PROBABILITY SIGNATURE: BRIDGING DATA SEMANTICS AND EMBEDDING STRUCTURE IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The embedding space of language models is widely believed to capture the semantic relationships; for instance, embeddings of digits often exhibit an ordered structure that corresponds to their natural sequence. However, the mechanisms driving the formation of such structures remain poorly understood. In this work, we interpret the embedding structures via the token relationships. We propose a set of probability signatures that reflect the semantic relationships among tokens. Through experiments on the composite addition tasks using the linear model and feedforward network, combined with theoretical analysis of gradient flow dynamics, we reveal that these probability signatures significantly influence the embedding structures. We further generalize our analysis to large language models (LLMs). Our results show that the probability signatures are faithfully aligned with the embedding structures, particularly in capturing strong pairwise similarities among embeddings. Our work offers a universal analytical framework that investigates how token relationships direct embedding geometries, empowering researchers to trace how gradient flow propagates token relationships onto embedding structures of their models.

1 INTRODUCTION

In recent years, deep neural network-based large language models (LLMs) have demonstrated remarkable performance (Comanici et al., 2025; OpenAI et al., 2024; DeepSeek-AI et al., 2025). The development of these models has largely followed what Richard Sutton termed "the bitter lesson"– that the most effective approach to improving AI performance has historically been to leverage greater computational resources, larger models, and more data, rather than incorporating human knowledge or specialized architectures (Sutton, 2019). This trend has been formalized through scaling laws (Kaplan et al., 2020). While these scaling laws provide valuable quantitative predictions for model performance, they also reveal a concerning limitation: achieving further significant improvements may require prohibitively large increases in model and data size, making continued scaling increasingly impractical and resource-intensive.

A more sustainable path forward lies in developing a mechanistic understanding of deep learning's success. Recent research has uncovered key properties such as the edge-of-stability phenomenon (Wu et al., 2018; Cohen et al., 2021), frequency principle (Xu et al., 2020; 2025a), attention patterns (Elhage et al., 2021; Olsson et al., 2022; Bhojanapalli et al., 2020), and parameter distribution characteristics (Kovaleva et al., 2021; Dar et al., 2023). Among these, the structure of the embedding space is fundamental: it serves as the gateway through which tokens are encoded, forming the basis of all subsequent learning. Indeed, embeddings often capture intuitive semantics—for instance, embeddings of digits 1,2,. . . ,9 form an ordered structure reflecting their numerical sequence (Mikolov et al., 2013b; Ethayarajh et al., 2019; Zhang et al., 2024; Yao et al., 2025). Yet, what drives this alignment between embedding geometry and semantic structure remains an open question: the precise mechanisms linking data distribution to embedding organization are still poorly characterized.

In this work, we establish a mechanistic link between embedding geometry and token relationship through the lens of gradient flow dynamics. For each token, we propose a set of probability signa-

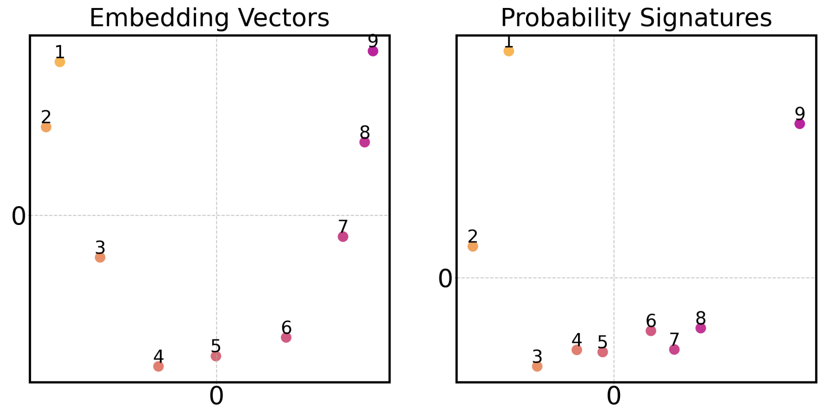


Figure 1: Left: The PCA projection of embedding vectors of the digits $1, 2, 3, \cdots, 9$ in Qwen2.5 3B-base. Right: The PCA projection of the probability signatures of the digits $1, 2, 3, \cdots, 9$ estimated by subsets of Pile(detailed formulation see (4)).

tures based on its statistical relationships with the other tokens (e.g., label distribution, co-occurrence patterns). Such probability signatures systematically capture inherent token-level relationships and reflect semantic structures. Our gradient flow analysis reveals that these signatures actively govern the evolution of embedding vectors, forging a deterministic connection between probability signature and embedding structure. This is illustrated in Figure 1: both the embeddings of digits 1,2,. . . ,9 in Qwen2.5 3B-base (Team, 2024) and their probability signatures estimated from the Pile corpus (Gao et al., 2020; Biderman et al., 2022) exhibit an ordered arrangement aligned with their natural sequence, suggesting that probability signatures are the prime driver of embedding organization. We instantiate this framework by deriving the exact signature sets for linear models and feedforward networks, showing how architecture determines which token relationships are encoded. Through carefully controlled synthetic tasks, we verify that manipulating probability signatures predictably reshapes the embedding space. Finally, we extrapolate our framework to LLMs, demonstrating that even in realistic training regimes, next-token and previous-token distributions dominate the dynamics of embedding and unembedding vectors in Qwen2.5 and Llama-2 architectures.

The primary contribution of this work is a universal analytical framework that investigates how token relationships direct embedding geometries. Through exact gradient flow analysis, we demonstrate that any embedding-based architecture encodes a specific, predictable subset of data distribution statistics into its token representations. This framework not only explains observed embedding structures as a deterministic consequence of probability signatures, but also predicts which probability signatures dominate in a given model, transforming representation learning from a black-box phenomenon into a transparent, distribution-driven process.

2 Related Work

Parameter analysis in LLMs Investigating the underlying parameter properties in LLMs is crucial for understanding the foundation of models. Some works focus on the specific modules in models. Elhage et al. (2021); Olsson et al. (2022) uncover mechanisms such as induction heads from the attention module. Bhojanapalli et al. (2020) reveals the rank-collapse phenomenon of the attention matrix. Geva et al. (2021; 2022); Dai et al. (2022) investigates the characteristics and functions of the FFN in LLMs. Additionally, analysis of a single neuron has been widely employed in mechanism interpretation, particularly in circuits analysis Hanna et al. (2023); Wang et al. (2023); Hanna et al. (2024); Wang et al. (2025), sparse autoencoders (SAE) Huben et al. (2024); Bricken et al. (2023), transcoders Dunefsky et al. (2024), and cross-layer transcoders (CLT) Ameisen et al. (2025). There are also some studies investigating the global properties of all parameters. Dar et al. (2023); Katz et al. (2024) introduce a framework for interpreting all parameters of Transformer models by projecting them into the embedding space. Kovaleva et al. (2021); Yu et al. (2025) provide an analysis of the parameter distribution, demonstrating the significance of these outliers. In this work, we will focus on the embedding space, explaining the formation of its structure from both experimental and theoretical perspectives.

Embedding structure and representation learning Since the introduction of static word embeddings by Mikolov et al. (2013a); Pennington et al. (2014) and the adoption of contextualized embeddings (Devlin et al., 2019; Peters et al., 2018), significant attention has been devoted to analyzing embedding properties. Gao et al. (2019); Ethayarajh (2019); Timkey & van Schijndel (2021) explore the anisotropy of embedding space, while Cai et al. (2021) show that embeddings exhibit isotropy within clusters. Liu et al. (2022) offers insights into grokking by emphasizing the role of well-organized embedding structures. Zhang et al. (2024) establishes a connection between embedding structure and model generalization, and Yao et al. (2025) provides an analysis of this relationship. Crucially, these studies characterize embedding geometry post hoc, treating it as an empirical phenomenon to be observed rather than a deterministic outcome to be explained. In contrast, we mechanistically interpret how embedding structures arise from token relationships. Our gradient-flow-driven framework reveals that token-wise probability signatures dictate the evolution of embedding vectors, offering not merely a new perspective, but a predictive, architecture-agnostic protocol for understanding representation formation.

3 Preliminary

3.1 Embedding-based Model

We denote the models functioning on the trainable embedding of the input sequence as embedding-based models. We provide the following formulation:

Definition 1. *Given a vocabulary $\mathcal{V} \subset \mathbb{N}^+$ with size d_{vob}, we denote $\boldsymbol{e}_x \in \mathbb{R}^{d_{\text{vob}}}$ as the one-hot vector of x for any $x \in \mathcal{V}$. The trainable embedding matrix and unembedding matrix are $\boldsymbol{W}^E \in \mathbb{R}^{d\times d_{\text{vob}}}$ and $\boldsymbol{W}^U \in \mathbb{R}^{d_{\text{vob}}\times d}$, respectively. For a sequence $\boldsymbol{X} := [x_1, x_2, \cdots, x_L] \in \mathcal{V}^L$ with length L. The trainable embedding of $\boldsymbol{X}$ and an embedding-based model F taking $\boldsymbol{X}$ as input could be formulated as*

$$\boldsymbol{W}^E_{\boldsymbol{X}} = \boldsymbol{W}^E \boldsymbol{e}_{\boldsymbol{X}} := \left[\boldsymbol{W}^E_{x_1}, \boldsymbol{W}^E_{x_2}, \cdots, \boldsymbol{W}^E_{x_L}\right],$$
$$F\left(\boldsymbol{X}\right) = \boldsymbol{W}^U G\left(\boldsymbol{W}^E_{\boldsymbol{X}}\right),$$

where G means the mapping in the hidden space, $\boldsymbol{W}^E_{x_i} = \boldsymbol{W}^E \mathbf{e}_{x_i}$ represents the embedding vector of elements $x_i \in \boldsymbol{X}$.

Embedding-based models have been widely applied in various domains, particularly in NLP. In this work, our objective is to investigate how the token relationships impact the characteristics of the embedding space. We will begin with the following simplified models, facilitating our analysis.

- Linear model. $F_{\text{lin}}\left(\boldsymbol{X}\right) = \boldsymbol{W}^U \sum_{x\in\boldsymbol{X}} \boldsymbol{W}^E_x$.
- Feedforward network. $F_{\text{ffn}}\left(\boldsymbol{X}\right) = \boldsymbol{W}^U \sigma\left(\sum_{x\in\boldsymbol{X}} \boldsymbol{W}^E_x\right)$, where σ denotes the element-wise nonlinear activation.

Furthermore, we will provide an elementary analysis of the Transformer architecture in language tasks and verify our results by the Qwen2.5 architecture and the Llama 2 architecture (Touvron et al., 2023).

3.2 Token relationships & Probability Signatures

In natural language, a token's meaning is fully constituted by its statistical context: how it predicts downstream labels, what tokens it co-occurs with, and how these relationships jointly evolve. Formally, these semantic regularities manifest as conditional probability distributions over the vocabulary. Denote the label of a sequence $\boldsymbol{X}$ by y and assume $(\boldsymbol{X}, y) \sim \pi$. For a token x in input $\boldsymbol{X}$, we consider four representative families of such distributions:

- **Label relationship**: $\mathbb{P}_\pi(y = \nu \mid x \in \boldsymbol{X})$ encodes what x signals about the output—e.g., "excellent" in a review robustly predicts a positive label ν, while "frustrated" skews toward negative.
- **Co-occurrence relationship**: $\mathbb{P}_\pi(x' \in \boldsymbol{X} \mid x \in \boldsymbol{X})$ captures syntactic-semantic neighborhoods—"stock" frequently co-occurs with "market" but rarely with "apple" (in the financial sense). Higher-order terms like $\mathbb{P}_\pi(x', x'' \in \boldsymbol{X} \mid x \in \boldsymbol{X})$ encode compositional contexts.

- **Joint relationship**: The joint $\mathbb{P}_\pi(x' \in \boldsymbol{X}, y = \nu \mid x \in \boldsymbol{X})$ reveals context-dependent labeling—"apple" co-occurring with "pie" predicts a food label, while with "store" predicts a tech label.
- **Inverse relationship**: $\mathbb{P}_\pi(x_i \in \boldsymbol{X} \mid y = x)$ describes what precedes a token as its cause—the tokens that predict x itself (e.g., what contexts make "surprised" likely to appear).

These token-wise relationships are semantic primitives: they are computable from data, independent of any model, but depend on the contexts and tokenizers, yet fully determine the token's functional role in the corpus. Critically, a sequence of length L yields exponentially many such relationships—our four families merely scratch the surface. **Rather than exhaustively enumerating them, we propose a systematic principle: the gradient flow dynamics of any embedding-based model will automatically select a specific subset of these relationships to encode.** To showcase this principle, we distill each family into a compact **probability signature**—a vector/matrix that aggregates the relevant conditional probabilities (Definition 2). This choice is deliberate: we aim not to prescribe a fixed signature set, but to demonstrate that any such set derived from gradient flow analysis will faithfully sculpt the embedding space.

Definition 2 (Probability Signatures)**.** *For token $x \in \mathcal{V}$, we define four probability signatures that capture distinct token relationships:*

$$\boldsymbol{\phi}_x^y = \sum_{\nu \in \mathcal{V}} \mathbb{P}_\pi(y = \nu \mid x \in \boldsymbol{X}) \boldsymbol{e}_\nu, \qquad \boldsymbol{\phi}_x^{\boldsymbol{X}} = \sum_{x' \in \mathcal{V}} \mathbb{P}_\pi(x' \in \boldsymbol{X} \mid x \in \boldsymbol{X}) \boldsymbol{e}_{x'},$$

$$\boldsymbol{\phi}_x^{\boldsymbol{X}|y} = \sum_{\nu, x'} \mathbb{P}_\pi(x' \in \boldsymbol{X}, y = \nu \mid x \in \boldsymbol{X})\, \boldsymbol{e}_\nu \times \boldsymbol{e}_{x'}^\top, \qquad \boldsymbol{\varphi}_x^{\boldsymbol{X}} = \sum_{x' \in \mathcal{V}} \mathbb{P}_\pi(x' \in \boldsymbol{X} \mid y = x) \boldsymbol{e}_{x'}.$$

We have $\boldsymbol{\phi}_x^y, \boldsymbol{\phi}_x^{\boldsymbol{X}}, \boldsymbol{\varphi}_x^{\boldsymbol{X}} \in \mathbb{R}^{d_{\text{vob}}}$, $\boldsymbol{\phi}_x^{\boldsymbol{X}|y} \in \mathbb{R}^{d_{\text{vob}} \times d_{\text{vob}}}$.

Each probability signature is a data-derived feature vector/matrix for x. For example, the ν-th element of $\boldsymbol{\phi}_x^y$ is $\mathbb{P}_\pi(y = \nu \mid x \in \boldsymbol{X})$. The signatures above are exemplars; our framework empowers researchers to derive more probability signatures for their models of interest by tracing how gradient flow propagates token relationships onto embedding structures.

4 Gradient Flow of Embedding Vector

To understand why embeddings organize as they do, we examine the continuous dynamics of training via gradient flow, the limit of gradient descent as the learning rate vanishes. This tool acts as a microscope, revealing the "force field" that sculpts each embedding vector. Formally, Given a dataset $\left\{\left(\boldsymbol{X}^i, y^i\right)\right\}_{i=1}^N$ with loss function $\ell^i = \ell\left(F\left(\boldsymbol{X}^i; \theta\right), y^i\right)$, the gradient descent implies that $\theta^{k+1} - \theta^k = -\eta \frac{1}{N} \sum_{i=1}^N \frac{\partial \ell^i}{\partial \theta} \big|_{\theta = \theta^k}$. Then the gradient flow of θ is defined as:

$$\frac{d\theta}{dt} := \lim_{\eta \to 0} \frac{\theta^{k+1} - \theta^k}{\eta} = -\frac{1}{N} \sum_{i=1}^N \frac{\partial \ell^i}{\partial \theta}.$$

Our goal is to trace how this dynamics acts on the embedding vector $\boldsymbol{W}_x^E$ for any token $x \in \mathcal{V}$. Using the standard cross-entropy loss:

$$\ell^i = -\log \operatorname{Softmax}\left(F\left(\boldsymbol{X}^i\right)\right)_{y^i} = -\log \frac{\exp F\left(\boldsymbol{X}^i\right)_{y^i}}{\sum_{j=1}^{d_{\text{vob}}} \exp F\left(\boldsymbol{X}^i\right)_j},$$

we derive the exact evolution equation:

Proposition 1. *Let $\odot$ represent the Hadamard product and T mean the matrix transpose. Given an embedding-based model F with an embedding matrix $\boldsymbol{W}^E$. For any token $x \in \mathcal{V}$, the gradient flow of $\boldsymbol{W}_x^E$ (the embedding vector of x) can be formulated as follow when $N \to \infty$:*

$$\begin{aligned}
\frac{d\boldsymbol{W}_x^E}{dt} = r_x^{\text{in}} \Bigg(& \sum_{\nu \in \mathcal{V}} \mathbb{P}_\pi\left(y = \nu \mid x \in \boldsymbol{X}\right) \left(\boldsymbol{W}^{U,T} \boldsymbol{e}_\nu\right) \odot \mathbb{E}_\pi \left[G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right) \mid x \in \boldsymbol{X}, y = \nu\right] \\
& - \mathbb{E}_\pi \left[\left(\boldsymbol{W}^{U,T} \boldsymbol{p}\right) \odot G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right) \mid x \in \boldsymbol{X}\right] \Bigg) \\
:= r_x^{\text{in}} & \left(\boldsymbol{U} \boldsymbol{\phi}_x^y - \mathbb{E}_\pi \left[\left(\boldsymbol{W}^{U,T} \boldsymbol{p}\right) \odot G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right) \mid x \in \boldsymbol{X}\right] \right),
\end{aligned}$$

where $\boldsymbol{U} \in \mathbb{R}^{d \times d_{\text{vob}}}$ *and the* ν*-th column of* $\boldsymbol{U}$ *equals* $\left(\boldsymbol{W}^{U,T} \boldsymbol{e}_\nu\right) \odot \mathbb{E}_\pi \left[G^{(1)}\left(\boldsymbol{W}^E_{\boldsymbol{X}}\right) \mid x \in \boldsymbol{X}, y = \nu\right]$. r^{in}_x *denotes the ratio of input sequences containing* x *in the training set,* $G^{(1)}$ *represents the derivative of* G *with respect to* $\boldsymbol{W}^E_x$ *and* $\boldsymbol{p} = \text{softmax}\left(F\left(\boldsymbol{X}\right)\right)$.

This equation reveals that $\boldsymbol{\phi}^y_x$ drives $\boldsymbol{W}^E_x$ toward a direction determined by the token-label semantics. This means: if two tokens share similar label distributions, their embeddings will be forced to evolve in similar directions from the very start of training. The emergence of other probability signatures ($\boldsymbol{\phi}^{\boldsymbol{X}}_x$, $\boldsymbol{\phi}^{\boldsymbol{X}|y}_x$) is dependent on the formulation of G, as we will show next.

To make this analysis concrete, we dissect linear model and feedforward networks, deriving their exact probability signature sets from Proposition 1. This demonstrates how our framework systematically extracts the relevant probability signatures for any given G.

4.1 Linear Model

For linear models F_{lin}, the hidden mapping G is simply the sum of embeddings. Substituting this into Proposition 1 yields a simplified dynamics where the gradient flow depends on only two probability signatures:

Corollary 1 (Embedding of Linear Model)**.** *Let* $N \to \infty$, π *denotes the data distribution over the training set. The gradient flow of* $\boldsymbol{W}^E_x$ *in* F_{lin} *can be approximated by*

$$\frac{d\boldsymbol{W}^E_x}{dt} = r^{\text{in}}_x \boldsymbol{W}^{U,T}\left(\boldsymbol{\phi}^y_x - \frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\boldsymbol{W}^E\boldsymbol{\phi}^{\boldsymbol{X}}_x + \boldsymbol{\eta}\right), \tag{1}$$

where $\boldsymbol{\eta}$ *denotes the data-independent and higher-order terms.*

The Corollary 1 indicates that the term $\boldsymbol{\phi}^y_x$ acts as the primary steering force. Early in training, when $\|\boldsymbol{W}^U\boldsymbol{W}^E\|$ is small, $\boldsymbol{\phi}^y_x$ alone dictates the update direction. The term $\boldsymbol{\phi}^{\boldsymbol{X}}_x$ modulates the embedding update based on contextual co-occurrence statistics, but its influence is scaled by $\frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\boldsymbol{W}^E$ and thus emerges later in training.

Experimental Validation: Controllable Addition Tasks If two tokens α, α' satisfy $\boldsymbol{\phi}^y_\alpha \approx \boldsymbol{\phi}^y_{\alpha'}$ and $\boldsymbol{\phi}^{\boldsymbol{X}}_\alpha \approx \boldsymbol{\phi}^{\boldsymbol{X}}_{\alpha'}$, Corollary 1 forces their embeddings to align: $\cos(\boldsymbol{W}^E_\alpha, \boldsymbol{W}^E_{\alpha'}) = \frac{\boldsymbol{W}^{E,T}_\alpha \boldsymbol{W}^E_{\alpha'}}{||\boldsymbol{W}^E_\alpha||_2 ||\boldsymbol{W}^E_{\alpha'}||_2} \to 1$. We design three **variable-controlled addition tasks** to isolate and verify each probability signature's influence. In each task, $\boldsymbol{\phi}^y_\alpha$ or $\boldsymbol{\phi}^{\boldsymbol{X}}_\alpha$ or both of them will be identical across α. Assuming all tokens belong to positive integers, and we denote an anchor set by $\mathcal{A}$, whose elements represent different addition operations, i.e., anchor α_1 means addition with α_1. Given a input sequence $\boldsymbol{X} = [z, \alpha_1, \alpha_2]$, we define the following tasks:

- **Addition task** (Varying $\boldsymbol{\phi}^y_\alpha$). $y = f_{\text{add}}\left(\boldsymbol{X}\right) = z + \alpha_1 + \alpha_2, \quad \alpha_1, \alpha_2 \in \mathcal{A}$. For each anchor pair (α_1, α_2), z is sampled from the same set $\mathcal{Z}$ with $\mathcal{Z} \cap \mathcal{A} = \emptyset$. In this task, $\boldsymbol{\phi}^{\boldsymbol{X}}_\alpha$ are identical across anchors while $\boldsymbol{\phi}^y_\alpha$ are distinct with varying anchors α.
- **Addition task with the same value domain** (Varying $\boldsymbol{\phi}^{\boldsymbol{X}}_\alpha$). $y = \tilde{f}_{\text{add}}\left(\boldsymbol{X}\right) = z + \alpha_1 + \alpha_2, \quad \alpha_1, \alpha_2 \in \mathcal{A}$. For anchor pair (α_1, α_2), $z \in \mathcal{Z}_{(\alpha_1, \alpha_2)} = \mathcal{Y} - \alpha_1 - \alpha_2$ where $\mathcal{Y}$ denotes the label set, which is identical for all anchor pairs. In $\tilde{f}_{\text{add}}$, $\boldsymbol{\phi}^{\boldsymbol{X}}_\alpha$ are distinct across anchors α while $\boldsymbol{\phi}^y_\alpha$ are identical for all $\alpha \in \mathcal{A}$.
- **Module addition** (Both signatures identical). $y = f_{\text{mod}}\left(\boldsymbol{X}\right) = \min \mathcal{Z} + \left(z + \alpha_1 + \alpha_2 \mod |\mathcal{Z}|\right), \quad \alpha_1, \alpha_2 \in \mathcal{A}$ and $z \in \mathcal{Z}$. Both $\boldsymbol{\phi}^{\boldsymbol{X}}_\alpha$ and $\boldsymbol{\phi}^y_\alpha$ are identical with different anchors.

In this work, we set $\mathcal{A} = \{11, 12, \cdots, 20\}$ and $\mathcal{Y} = \mathcal{Z} = \{101, 102, \cdots, 140\}$. Figure 2A visualizes the probability signature similarities for each task, confirming our manipulations. The detailed mathematical formulations of these signatures in each task are provided in Appendix B.1.

Results: Theory Predicts Embedding Structure We train F_{lin} for each task with $d = 200$. Tasks f_{add} and $\tilde{f}_{\text{add}}$ are well learned, while f_{mod} fails to be fitted. The details are provided in Appendix A. Figure 2 B represents the value of $\cos\left(\boldsymbol{W}^E_\alpha, \boldsymbol{W}^E_{\alpha'}\right)$ in the three tasks.

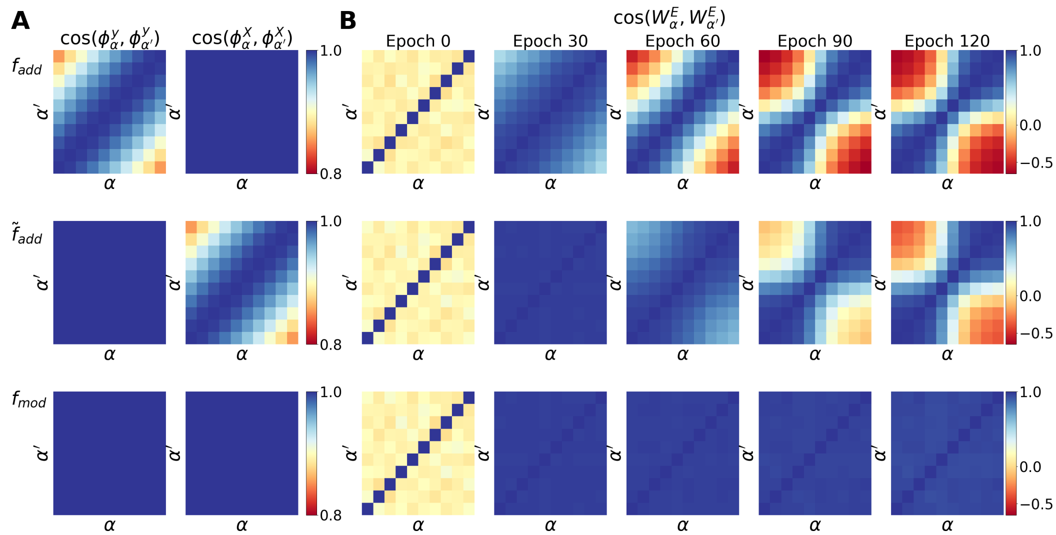


Figure 2: A: The heatmap of $\cos\left(\phi^y_\alpha, \phi^y_{\alpha'}\right)$ and $\cos\left(\phi^{\boldsymbol{X}}_\alpha, \phi^{\boldsymbol{X}}_{\alpha'}\right)$ in three addition tasks. B: The heatmap of $\cos\left(\boldsymbol{W}^E_\alpha, \boldsymbol{W}^E_{\alpha'}\right)$ in F_{lin} across different tasks.

- **Task f_{add}:** different anchor embeddings quickly form an ordered structure, where the cosine similarity gets smaller as the anchor distance gets larger. The distribution of $\cos\left(\boldsymbol{W}^E_\alpha, \boldsymbol{W}^E_{\alpha'}\right)$ is consistent with the $\cos\left(\phi^y_\alpha, \phi^y_{\alpha'}\right)$ (Figure 2 A), implying the impact of ϕ^y_α in directing $\boldsymbol{W}^E_\alpha$.
- **Task $\tilde{f}_{\text{add}}$:** The anchor embeddings also develop a similar hierarchical structure, aligned with the structure of $\phi^{\boldsymbol{X}}_\alpha$ in $\tilde{f}_{\text{add}}$. But its convergence is slower, validating that ϕ^y_α dominates early dynamics.
- **Task f_{mod}:** Although the task is unsolvable by a linear model, all anchor embeddings collapse to the same direction, exactly as Corollary 1 predicts when both signatures are identical.

4.2 FFN Unlocks Joint relationships: Solving the Modular Addition Puzzle

Recall that in Section 4.1, the linear model failed to learn f_{mod}, whose embeddings collapsed to a single direction. It's not because the task lacked structure, but the linear model cannot encode the probability signature $\phi^{\boldsymbol{X}|y}_x$. We find that the nonlinear activation could resolve this problem and provide the following results.

Corollary 2 (Embedding of FFN). *Let $N \to \infty$, π denotes the data distribution over the training set. The gradient flow of $\boldsymbol{W}^E_x$ in F_{ffn} could be approximated by*

$$\frac{d\boldsymbol{W}^E_x}{dt} = r^{\text{in}}_x \left(\boldsymbol{W}^{U,T} \left(\phi^y_x - \frac{1}{d_{\text{vob}}} \boldsymbol{W}^U \boldsymbol{W}^E \phi^{\boldsymbol{X}}_x \right) + \mathbb{T} \cdot \phi^{\boldsymbol{X}|y}_x + \boldsymbol{\epsilon} \right), \tag{2}$$

where $\mathbb{T} \in \mathbb{R}^{d \times d_{\text{vob}} \times d_{\text{vob}}}$, $\mathbb{T}_{:,x',\nu} = \boldsymbol{W}^U_\nu \odot \boldsymbol{W}^E_{x'}$ for $\nu, x' \in \mathcal{V}$ and 0 otherwise. $\boldsymbol{\epsilon}$ represents the higher-order term.

This is a qualitative leap beyond F_{lin}: The new term $\mathbb{T} \cdot \phi^{\boldsymbol{X}|y}_x$ directly encodes how the presence of x influences the co-occurrence distribution conditioned on future labels. For f_{mod}, $\phi^{\boldsymbol{X}|y}_x$ varies systematically with α (shown in Figure 3 A), thereby providing the necessary signal that the linear model could not access. We train the f_{mod} with F_{ffn} to test whether $\phi^{\boldsymbol{X}|y}_x$ enables structure formation. Figure 3 B depicts the cosine similarity among anchor embeddings, demonstrating that the embedding structure in f_{mod} is ordered, which validates our analysis. This contract validates that the specific probability signatures encoded are architecture-dependent, but the governing principle—gradient flow transforms signatures into structure—is universal.

Geometric Proof: PCA Visualization of Signature-Embedding Alignment Proposition 1 and Corollaries 1-2 make algebraic predictions; we now render them as visible geometry. Figure 4

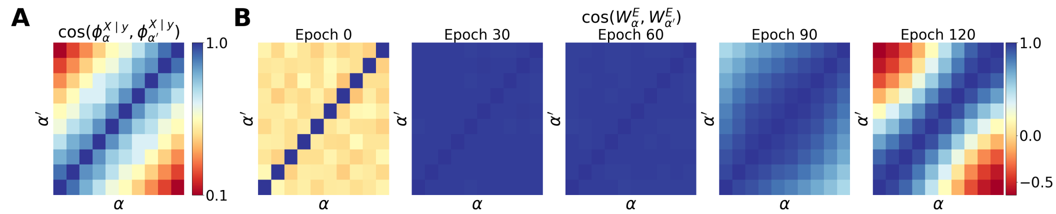


Figure 3: A: The heatmap of $\cos\left(\phi_\alpha^{\boldsymbol{X}|y}, \phi_{\alpha'}^{\boldsymbol{X}|y}\right)$ in f_{mod}. B: $\cos\left(\boldsymbol{W}_\alpha^E, \boldsymbol{W}_{\alpha'}^E\right)$ in F_{ffn} learning f_{mod}.

projects all probability signatures (left 3 columns) and learned embeddings (right 2 columns) into 2D space via PCA. This result reveals that in F_{lin}, the embedding structure is primarily influenced by ϕ_α^y and $\phi_\alpha^{\boldsymbol{X}}$. Specifically, when both ϕ_α^y and $\phi_\alpha^{\boldsymbol{X}}$ are controlled in f_{mod}, the embedding structure is chaotic. Besides, the embedding space in F_{ffn} is impacted by another probability signature $\phi_\alpha^{\boldsymbol{X}|y}$. These phenomena are consistent with our theoretical analysis, illustrating that analyzing the embedding space via the gradient flow and linking to the token relationships is viable.

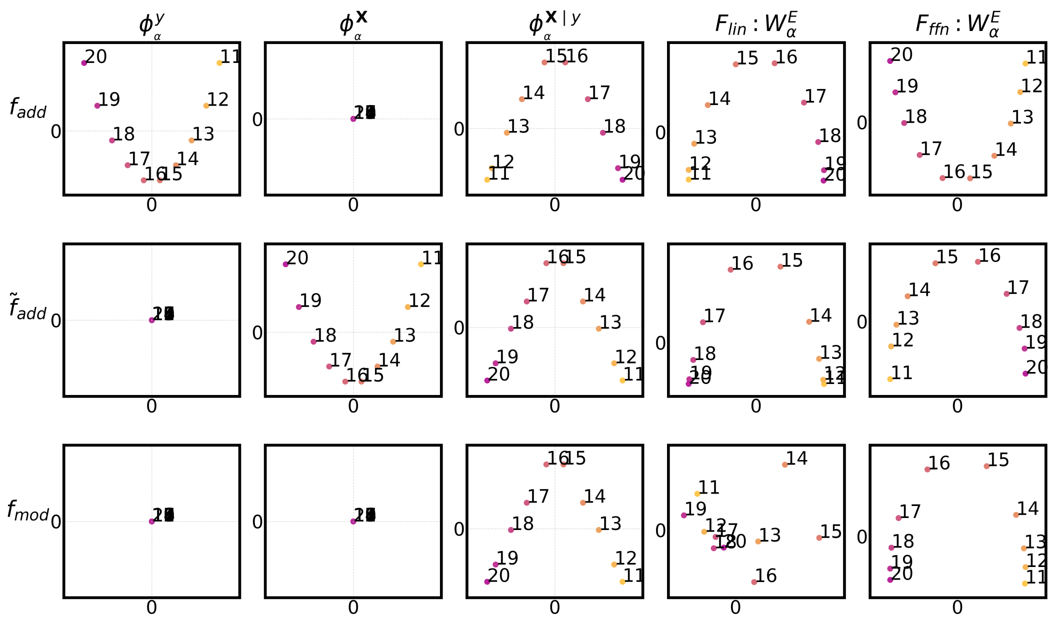


Figure 4: PCA projection of the three types of probability signatures and the embedding vectors in F_{lin} and F_{ffn} (epoch 120).

5 GRADIENT FLOW OF UNEMBEDDING VECTOR

Our analysis thus far has focused on the encoding side—how tokens are embedded into hidden space. A complete theory must also explain the decoding side: how the unembedding matrix $\boldsymbol{W}^U$ learns to map hidden representations back to token probabilities. Remarkably, gradient flow reveals a perfect symmetry: just as embeddings evolve under token-level probability signatures, unembeddings evolve under inverse signatures that capture how tokens are predicted from contexts.

Proposition 2. *Given an embedding-based model F with an unembedding matrix $\boldsymbol{W}^U$. For any token $\nu \in \mathcal{V}$, the gradient flow of $\boldsymbol{W}_\nu^U$ (the ν-th row of $\boldsymbol{W}^U$) can be written as*

$$\frac{d\boldsymbol{W}_\nu^U}{dt} = r_\nu^{\text{out}} \mathbb{E}_\pi \left[G\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right)^T \mid y = \nu \right] - \mathbb{E}_\pi \left[\boldsymbol{p}_\nu G\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right)^T \right],$$

where r_ν^{out} denotes the ratio of sequences whose label is ν and $\boldsymbol{p}_\nu$ means the ν-th element of $\boldsymbol{p}$.

Specifically, we have the following formulation for the linear model:

Corollary 3 (Unembedding of Linear Model)**.** *Let $N \to \infty$, π denotes the data distribution over the training set. The gradient flow of $\boldsymbol{W}^U_\nu$ in F_{lin} could be approximated by*

$$\frac{d\boldsymbol{W}^U_\nu}{dt} = Lr^{\text{out}}_\nu \left(\boldsymbol{W}^E \varphi^{\boldsymbol{X}}_\nu\right)^T + \boldsymbol{\eta}, \tag{3}$$

where $\boldsymbol{\eta}$ denotes the output term.

Corollary 3 demonstrates that $\phi^{\boldsymbol{X}}_\nu$ directs the dynamics of the unembedding vector. We extract the unembedding matrix from the addition tasks and compare its geometry to $\varphi^{\boldsymbol{X}}_\nu$. Figure 5 reveals the same striking alignment observed for embeddings. Figure 5 B depicts the distribution of $\cos\left(\varphi^{\boldsymbol{X}}_\nu, \varphi^{\boldsymbol{X}}_{\nu'}\right)$, which is aligned with the distribution of the $\cos\left(\boldsymbol{W}^U_\nu, \boldsymbol{W}^U_{\nu'}\right)$. Furthermore, Figure 5 C compares the PCA projection of $\varphi^{\boldsymbol{X}}_\nu$ and $\boldsymbol{W}^U_\nu$ in all tasks, revealing a high consistency and validating our analysis. This symmetric validation completes our framework: Gradient flow does not arbitrarily shape parameters—it encodes data statistics into model weights with mathematical precision, whether on the input or output side.

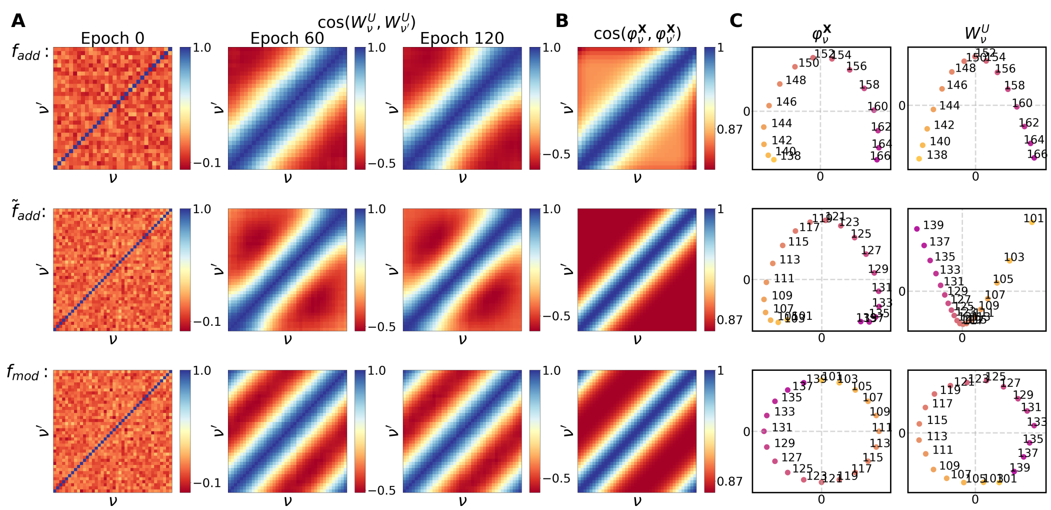


Figure 5: A: The heatmap of the $\cos\left(\boldsymbol{W}^U_\nu, \boldsymbol{W}^U_{\nu'}\right)$ in F_{lin} during the training process. B: The heatmap of $\cos\left(\varphi^{\boldsymbol{X}}_\nu, \varphi^{\boldsymbol{X}}_{\nu'}\right)$ across different tasks. C: PCA projection of $\varphi^{\boldsymbol{X}}_\nu$ and $\boldsymbol{W}^U_\nu$ (epoch 120).

6 Language Model

Our analysis of synthetic tasks demonstrates that gradient flow dynamics encode probability signatures into embedding structures. We now ask: Does this principle scale to language models trained on real-world corpora? A full analysis of all terms in Proposition 1 for Transformers would be intractable and, more importantly, unnecessary for validating our core contribution. We therefore adopt a minimalist validation strategy: analyze the dominant probability signature predicted by gradient flow and test whether it alone can predict embedding structure. If this simplified analysis succeeds, it proves that our framework captures the essential mechanism and researchers can then extend it to additional modules as needed.

For decoder-only Transformers with next-token prediction, the gradient flow of embeddings is dominated by the next-token distribution since the model could be formulated as follows.

$$F_{\text{lan}}\left(\boldsymbol{X}\right) = \boldsymbol{W}^U \left(\boldsymbol{W}^E_{\boldsymbol{X}} + \tilde{F}\left(\boldsymbol{X}\right)\right).$$

Formally, given the training corpus $\left\{\boldsymbol{X}^i\right\}_{i=1}^N$, we define the following probability signatures for any $s \in \mathcal{V}$:

$$\begin{aligned} \phi^{\text{next}}_s &= \sum_{s' \in \mathcal{V}} \mathbb{P}_\pi \left(\cup_{t=1}^{L-1} \left\{X_{t+1} = s' \mid X_t = s\right\}\right) \boldsymbol{e}_{s'}, \\ \varphi^{\text{pre}}_s &= \sum_{s' \in \mathcal{V}} \mathbb{P}_\pi \left(\cup_{t=1}^{L-1} \left\{X_t = s' \mid X_{t+1} = s\right\}\right) \boldsymbol{e}_{s'}, \end{aligned} \tag{4}$$

We derive the following result:

Corollary 4. *Let $N \to \infty$, π denotes the token distribution in the training dataset. The gradient flow of the embedding vector $\boldsymbol{W}_s^E$ of token s could be fomulated as*

$$\frac{d\boldsymbol{W}_s^E}{dt} = r_s^{\text{in}} \boldsymbol{W}^{U,T} \boldsymbol{\phi}_s^{\text{next}} + \boldsymbol{\eta}^E.$$

Furthermore, the gradient flow of the unembedding vector $\boldsymbol{W}_s^U$ could be approximated as

$$\frac{d\boldsymbol{W}_s^U}{dt} = r_s^{\text{out}} \left(\boldsymbol{W}^E \boldsymbol{\varphi}_s^{\text{pre}}\right)^T + \boldsymbol{\eta}^U.$$

The $\boldsymbol{\eta}^E$ and $\boldsymbol{\eta}^U$ denote the output probability and the higher-order term.

Probability signatures impact the embedding space in language models Corollary 4 suggests that given any token s, the distributions of its next token and previous token significantly impact its embedding. To verify this result, we trained a group of Qwen2.5 models on different subsets of the Pile. Figure 6 A shows these similarity matrices for the dataset Pile-dm-mathematics, where the tokens displayed are those that occur most frequently in the corpus. We define the following correlation coefficient $R_{\cos}\left(\boldsymbol{W}^E, \boldsymbol{\phi}^{\text{next}}\right) := \text{Corr}\left(\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}_{s'}^E\right), \cos\left(\boldsymbol{\phi}_s^{\text{next}}, \boldsymbol{\phi}_{s'}^{\text{next}}\right)\right)$, and similarly $R_{\cos}\left(\boldsymbol{W}^U, \boldsymbol{\varphi}^{\text{pre}}\right)$. Figure 6 B tracks the $R_{\cos}\left(\boldsymbol{W}^E, \boldsymbol{\phi}^{\text{next}}\right)$ and $R_{\cos}\left(\boldsymbol{W}^U, \boldsymbol{\varphi}^{\text{pre}}\right)$ across all subsets during training (20 epochs). Correlations increase during the first epoch, indicating that gradient flow rapidly encodes next-token and previous-token statistics into embeddings and unembeddings. After reaching peak alignment, correlations plateau and dip slightly, showing that the embedding structure is still largely impacted by $\boldsymbol{\phi}_s^{\text{next}}$ and $\boldsymbol{\varphi}_s^{\text{pre}}$. The fact that a single simplified probability signature maintains predictive power throughout training, proves that our gradient flow analysis captures the essential mechanism of embedding structure. Researchers can now systematically uncover additional probability signatures (e.g., from attention patterns or higher-order terms) to account for residual variance. Furthermore, we find that the probability signatures reflect the strong connections of embeddings more faithfully, and we provide a detailed analysis in the Appendix C.3. Additionally, we provide another set of experiments using the Llama2 architecture in Appendix C.4.

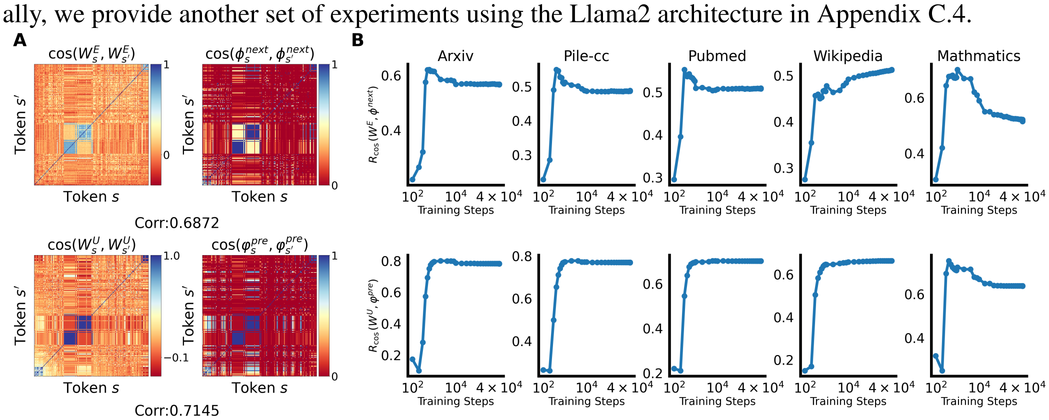


Figure 6: A: Heatmap of $\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}_{s'}^E\right)$ (left up), $\cos\left(\boldsymbol{\phi}_s^{\text{next}}, \boldsymbol{\phi}_{s'}^{\text{next}}\right)$ (right up), $\cos\left(\boldsymbol{W}_s^U, \boldsymbol{W}_{s'}^U\right)$ (left down) and $\cos\left(\boldsymbol{\varphi}_s^{\text{pre}}, \boldsymbol{\varphi}_{s'}^{\text{pre}}\right)$ (right up) in the experiment on dataset Pile-dm-mathematics (1 epoch). B: The dynamics of $R_{\cos}\left(\boldsymbol{W}^E, \boldsymbol{\phi}^{\text{next}}\right)$ (top) and $R_{\cos}\left(\boldsymbol{W}^U, \boldsymbol{\varphi}^{\text{pre}}\right)$ (bottom) during training (20 epochs) across different datasets.

Validating with the open-source model Since general-purpose pretrained base models are trained on broad corpora, we attempt to directly estimate their embedding structure by the probability signature. We employ Qwen2.5-3B-base for comparison and define $\tilde{\boldsymbol{\phi}}_s = \boldsymbol{\phi}_s^{\text{next}} + \boldsymbol{\varphi}_s^{\text{pre}}$, since $\boldsymbol{W}^E = \boldsymbol{W}^{U,T}$ in Qwen2.5-3B-base (the detail is provided in Appendix C.2). We compute $\tilde{\boldsymbol{\phi}}_s$ from the subsets of Pile. As shown in Figure 7 A, the structure of $\tilde{\boldsymbol{\phi}}_s$ could capture the main properties of the embedding structure, particularly the presence of sub-blocks with high similarity. Furthermore, we examine the instance for the digits ranging from 1 to 9. Figure 1 exhibits the PCA projections of $\boldsymbol{W}_s^E$ and $\tilde{\boldsymbol{\phi}}_s$, while Figure 7B illustrates their respective cosine similarities $\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}_{s'}^E\right)$ and

$\cos\left(\tilde{\phi}_s, \tilde{\phi}_{s'}\right)$, with both figures revealing an ordered organization aligned with the numerical sequence. However, this estimation does not always hold. On the one hand, Zhang et al. (2024) finds that initialization scale significantly affects the emergence of such embedding structures, demonstrating that in the NTK regime, the embedding structure may fail to capture token relationships. On the other hand, since probability signatures are computed from the training dataset, obtaining the correct data distribution becomes difficult when the corpus is carefully curated.

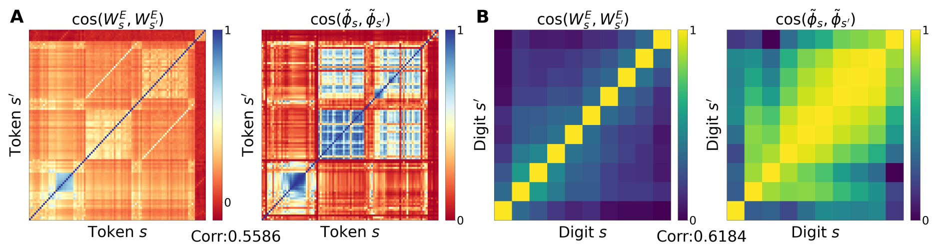


Figure 7: $\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}_{s'}^E\right)$ of the Qwen2.5-3B-base and $\cos\left(\tilde{\boldsymbol{\phi}}_s, \tilde{\boldsymbol{\phi}}_{s'}\right)$, respectively, with the frequently-appearing tokens (A) and the digits from 0 to 9 (B).

7 Discussion & Conclusion

We have shown that the geometry of embedding spaces is not a mysterious emergent phenomenon, but a deterministic encoding of probability signatures sculpted by gradient flow dynamics. More importantly, we have demonstrated that this encoding can be reverse-engineered: given any embedding-based architecture, our framework systematically extracts the exact set of statistical relationships that drive embedding evolution. This transforms representation learning from a black box into a transparent, distribution-driven process.

Guidance for Model Architectures and Training Methods We illustrate that each architecture implicitly selects which probability signatures it can encode. Our gradient-flow analysis makes this selection explicit and quantifiable: Corollary 1 proves that linear models cannot encode joint token-label relationships ($\boldsymbol{\phi}_x^{\boldsymbol{X}|y}$). Any task requiring this relationship will fail, regardless of scale. Adding a nonlinear activation unlocks $\boldsymbol{\phi}_x^{\boldsymbol{X}|y}$ (Corollary 2), enabling models to learn such semantics. This suggests a principled architecture search: introduce modules whose Jacobians $G^{(1)}$ encode desired probability signatures. On the other hand, our results have shown that the loss function is not merely a performance metric but also a gradient flow sculptor that determines which probability signatures dominate. Corollary 4 shows that next-token prediction makes $\boldsymbol{\phi}_s^{\text{next}}$ the dominant signature, embedding tokens based on immediate neighbors. This explains why standard autoregressive models excel at local coherence but struggle with long-range dependencies. If the loss predicts k future tokens, gradient flow will encode the k-gram relationship distribution. This provides a theoretical explanation for why multi-token prediction could easily capture the global relationships (Gloeckle et al., 2024).

Future Work We deliberately analyzed only four signature families and a simplified LLM gradient flow. This was not due to theoretical incompleteness, but to demonstrate the framework's modular extensibility. Just as we derived $\boldsymbol{\phi}_x^{\boldsymbol{X}|y}$ for feedforward networks and $\boldsymbol{\phi}_s^{\text{next}}$ for Transformers, researchers can now systematically mine custom signatures for their architectures of interest. The framework is designed to be extended. As a future direction, we will focus on analyzing the probability signatures in the self-attention module and the completed Transformer layer. This is not a correction to our theory, but its natural evolution.

References

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 864–873. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bhojanapalli20a.html.

Stella Biderman, Kieran Bicheno, and Leo Gao. Datasheet for the pile. *arXiv preprint arXiv:2201.07311*, 2022.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xYGNO86OWDH.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, and Inderjit Dhillon et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, and Runxin Xu et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 24375–24410. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2b8f4db0464cc5b6e9d5e6bea4b9f308-Paper-Conference.pdf.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep

Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *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. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkEYojRqtm.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15706–15734. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/gloeckle24a.html.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=p4PckNQR8k.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

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.

Shahar Katz, Yonatan Belinkov, Mor Geva, and Lior Wolf. Backward lens: Projecting language model gradients into the vocabulary space. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2390–2422, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.142. URL https://aclanthology.org/2024.emnlp-main.142/.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. BERT busters: Outlier dimensions that disrupt transformers. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3392–3405, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.300. URL https://aclanthology.org/2021.findings-acl.300/.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer relu neural networks at infinite-width limit. *Journal of Machine Learning Research*, 22(71):1–47, 2021.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013a. URL https://arxiv.org/abs/1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013b.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, and Florencia Leoni Aleman et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202/.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.372. URL https://aclanthology.org/2021.emnlp-main.372/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=45EIiFd6Oa.

Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020.

Zhi-Qin John Xu, Yaoyu Zhang, and Tao Luo. Overview frequency principle/spectral bias in deep learning. *Communications on Applied Mathematics and Computation*, 7(3):827–864, 2025a.

Zhi-Qin John Xu, Yaoyu Zhang, and Zhangchen Zhou. An overview of condensation phenomenon in deep learning. *arXiv preprint arXiv:2504.09484*, 2025b.

Junjie Yao, Zhongwang Zhang, and Zhi-Qin John Xu. An analysis for reasoning bias of language models with small initialization. In *Forty-second International Conference on Machine Learning*, 2025.

Mengxia Yu, De Wang, Qi Shan, Colorado Reed, and Alvin Wan. The super weight in large language models, 2025. URL https://arxiv.org/abs/2411.07191.

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, and Zhi-Qin John Xu. Initialization is critical to whether transformers fit composite functions by reasoning or memorizing. In *Advances in Neural Information Processing Systems*, 2024.

LLMs USAGE

In this work, the LLMs are employed to correct grammatical errors and inappropriate words.

A EXPERIMENTAL SETUPS

Addition tasks For each type of addition task, we trained a linear model F_{lin} and a Feedforward network F_{ffn}. The hidden size $d = 200$, and we employed the ReLU as the activation function. Each dataset contains 50000 data pairs. The training is conducted for 1000 epochs with a batch size of 100. The AdamW optimizer is employed with an initial learning rate of 10^{-5}. Inspired by the work of Luo et al. (2021); Xu et al. (2025b), we initialize the model parameters by $\boldsymbol{W}_{i,j} \sim \mathcal{N}\left(0, d^{-0.8}\right)$, indicating a small initialization scale.

Language models In the analysis of the LLMs, we employ the Qwen2.5 architecture with 12 layers and 12 attention heads in each layer. We set up that the hidden size is 512, and the intermediate size in FFN is 1024. The dimension of the key vectors and value vectors in each head is 64. Similarly, we initialize the parameter by $\boldsymbol{W}_{i,j} \sim \mathcal{N}\left(0, d_{\text{in}}^{-1}\right)$ where d_{in} means the input dimension of $\boldsymbol{W}$. We select five subsets of Pile, including Pile-arxiv, Pile-dm-mathematics, Pile-cc, Pile-pubmed-central, and Pile-wikipedia-en. The length of each sequence is 2048. The training is conducted for 1 epoch in each experiment, with the AdamW optimizer and a cosine learning rate schedule utilized. The initial learning rate is 10^{-4}.

B ADDITION TASK

B.1 PROBABILITY SIGNATURES IN ADDITION TASKS

We provide a formulation of the following probability in the three addition tasks. We denote $U\left(\mathcal{A}\right)$ and $U\left(\mathcal{Z}\right)$ as the discrete uniform distribution over $\mathcal{A}$ and $\mathcal{Z}$, respectively. A and Z are the random variables following $U\left(\mathcal{A}\right)$ and $U\left(\mathcal{Z}\right)$. For the task f_{add}, we have that

$$
\begin{aligned}
&\mathbb{P}_\pi\left(y=\nu \mid \alpha \in \boldsymbol{X}\right)=\mathbb{P}_\pi\left(A+Z=\nu-\alpha\right), \quad \mathbb{P}_\pi\left(z \in \mathcal{X} \mid \alpha \in \boldsymbol{X}\right)=\frac{1}{|\mathcal{Z}|}, \\
&\mathbb{P}_\pi\left(z \in \boldsymbol{X} \mid \alpha \in \boldsymbol{X}, y=\nu\right)=\mathbb{P}_\pi\left(A=\nu-\alpha-z\right)=\frac{1}{|\mathcal{A}|}\delta_{\nu-\alpha-z\in\mathcal{A}}, \\
&\mathbb{P}_\pi\left(\alpha' \in \boldsymbol{X} \mid \alpha \in \boldsymbol{X}, y=\nu\right)=\mathbb{P}_\pi\left(Z=\nu-\alpha-\alpha'\right)=\frac{1}{|\mathcal{Z}|}\delta_{\nu-\alpha-\alpha'\in\mathcal{Z}}, \\
&\mathbb{P}_\pi\left(z \in \boldsymbol{X} \mid y=\nu\right)=\mathbb{P}_\pi\left(A+A=\nu-z\right), \quad \mathbb{P}_\pi\left(\alpha \in \boldsymbol{X} \mid y=\nu\right)=\mathbb{P}_\pi\left(A+Z=\nu-\alpha\right),
\end{aligned}
$$

where $\alpha, \alpha' \in \mathcal{A}, z \in \mathcal{Z}$.It's noted that besides the co-occurrence probability $\mathbb{P}_\pi\left(z \in \mathcal{X} \mid \alpha \in \boldsymbol{X}\right)$, the value of other ones is dependent on α or ν. Figure 8 (left) displays the distribution of these probabilities, which intuitively reveals the cause of the hierarchy structure in the similarity matrix. Similarly, for $\tilde{f}_{\text{add}}$, denote $Y \sim U\left(\mathcal{Y}\right)$ and we have

$$
\begin{aligned}
&\mathbb{P}_\pi\left(y=\nu \mid \alpha \in \boldsymbol{X}\right)=\frac{1}{|\mathcal{Y}|}, \quad \mathbb{P}_\pi\left(z \in \mathcal{X} \mid \alpha \in \boldsymbol{X}\right)=\mathbb{P}_\pi\left(Y-A=z+\alpha\right), \\
&\mathbb{P}_\pi\left(z \in \boldsymbol{X} \mid \alpha \in \boldsymbol{X}, y=\nu\right)=\mathbb{P}_\pi\left(A=\nu-\alpha-z\right)=\frac{1}{|\mathcal{A}|}\delta_{\nu-\alpha-z\in\mathcal{A}}, \\
&\mathbb{P}_\pi\left(\alpha' \in \boldsymbol{X} \mid \alpha \in \boldsymbol{X}, y=\nu\right)=\frac{1}{|\mathcal{Z}|}, \\
&\mathbb{P}_\pi\left(z \in \boldsymbol{X} \mid y=\nu\right)=\mathbb{P}_\pi\left(A+A=\nu-z\right), \quad \mathbb{P}_\pi\left(\alpha \in \boldsymbol{X} \mid y=\nu\right)=\mathbb{P}_\pi\left(A+Z=\nu-\alpha\right).
\end{aligned}
$$

For f_{mod}, we have

$$
\begin{aligned}
&\mathbb{P}_\pi\left(y=\nu \mid \alpha \in \boldsymbol{X}\right)=\frac{1}{|\mathcal{Z}|}, \quad \mathbb{P}_\pi\left(z \in \mathcal{X} \mid \alpha \in \boldsymbol{X}\right)=\frac{1}{|\mathcal{Z}|}, \\
&\mathbb{P}_\pi\left(z \in \boldsymbol{X} \mid \alpha \in \boldsymbol{X}, y=\nu\right)=\frac{1}{|\mathcal{A}|} \delta_{\nu-\min \mathcal{Z}-(\alpha-z \bmod |\mathcal{Z}|) \in(A \bmod |\mathcal{Z}|)}, \\
&\mathbb{P}_\pi\left(\alpha' \in \boldsymbol{X} \mid \alpha \in \boldsymbol{X}, y=\nu\right)=\frac{1}{|\mathcal{Z}|}, \\
&\mathbb{P}_\pi\left(z \in \boldsymbol{X} \mid y=\nu\right)=\mathbb{P}_\pi\left((A+A \bmod |\mathcal{Z}|)=\nu-\min \mathcal{Z}-(z \bmod |\mathcal{Z}|)\right), \\
&\mathbb{P}_\pi\left(\alpha \in \boldsymbol{X} \mid y=\nu\right)=\mathbb{P}_\pi\left((A+Z \bmod |\mathcal{Z}|)=\nu-\min \mathcal{Z}-(\alpha \bmod |\mathcal{Z}|)\right).
\end{aligned}
$$

Figure 8 depicts all these probability distributions.

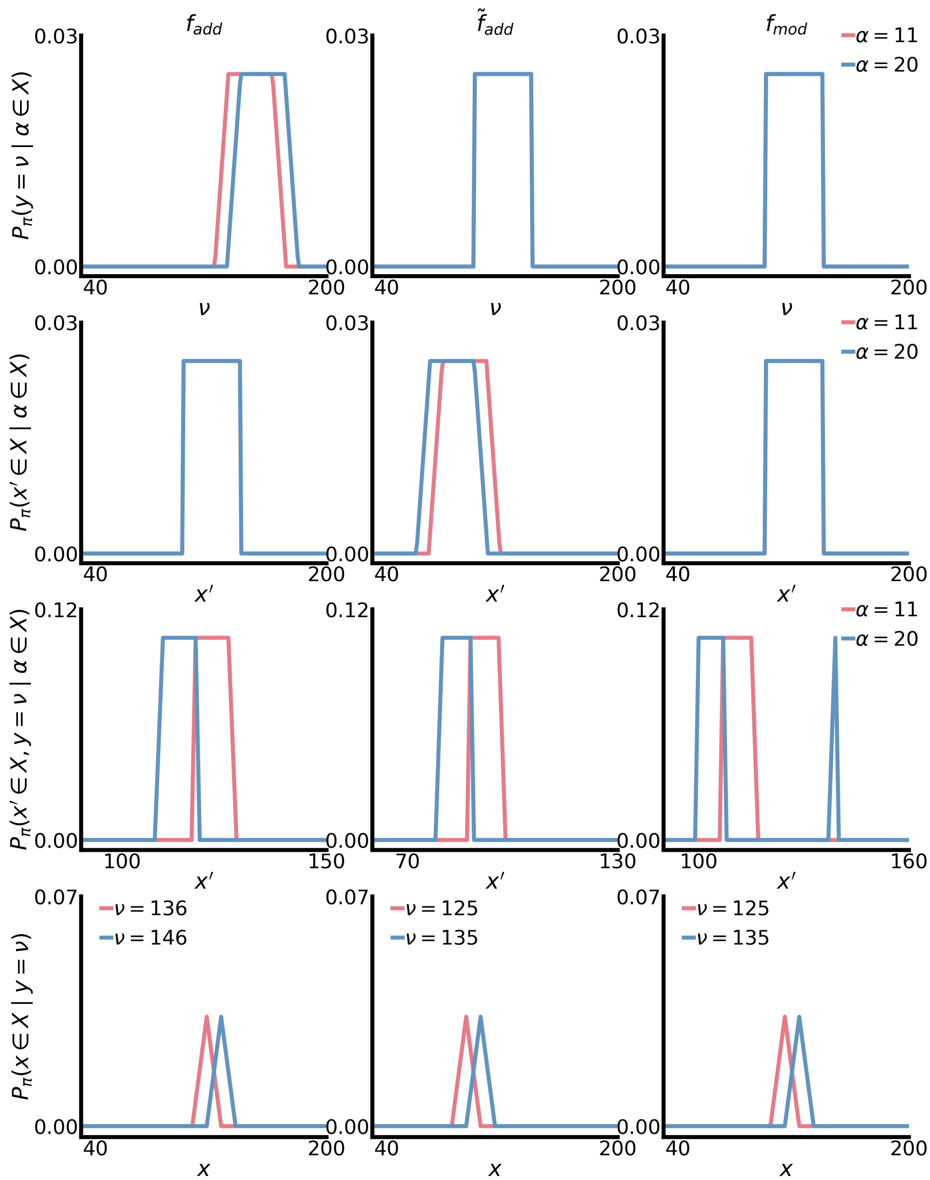


Figure 8: Probability signatures in each task under distinct α and ν. In the distribution of $\mathbb{P}_\pi\left(x' \in \boldsymbol{X}, y = \nu \mid \alpha \in \boldsymbol{X}\right)$, $\nu = 150$ is displayed in f_{add} and $\nu = 120$ in $\tilde{f}_{\text{add}}$ and f_{mod}, since 150 and 120 are the average label value in each task.

B.2 Training result

Figure 9 shows the training accuracy of F_{lin} and F_{ffn} on the three addition tasks. The results reveal that both f_{add} and $\tilde{f}_{\text{add}}$ are learned well by the linear model, whereas f_{mod} requires the nonlinear model to achieve an effective fit.

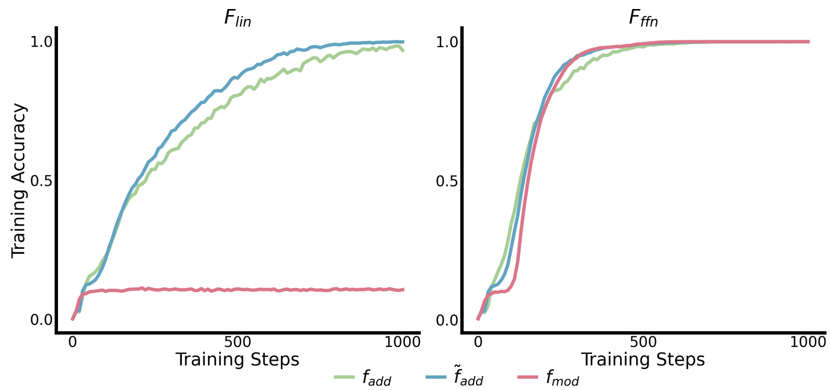


Figure 9: Training accuracy of the F_{lin} (left) and F_{ffn} (right) on the three addition tasks.

B.3 Quantify the hierarchy embedding structure

In the addition tasks, the anchors exhibit a strict ordering due to the numerical sequence. This provides an ideal setting for the embedding space to develop a corresponding ordered relationship. To formally quantify the formation of the ordered structure, we define the following metric:

$$R_{\text{order}}\left(\boldsymbol{W}_{\mathcal{A}}^{E}\right) = \text{Corr}\left(\cos\left(\boldsymbol{W}_{\alpha}^{E}, \boldsymbol{W}_{\alpha'}^{E}\right), \mid \alpha - \alpha' \mid\right).$$

$R_{\text{order}}\left(\boldsymbol{W}_{\mathcal{A}}^{E}\right)$ reflects the relationship between embedding similarity and anchor difference. A strong negative $R_{\text{order}}\left(\boldsymbol{W}_{\mathcal{A}}^{E}\right)$ (approximately -1) indicates that the similarity decreases systematically with increasing anchor difference, confirming the presence of a hierarchical organization in the anchor embeddings. Figure 10 depicts the corresponding evolution of $R_{\text{order}}\left(\boldsymbol{W}_{\mathcal{A}}^{E}\right)$ in F_{lin} and F_{ffn}, which is consistent with our analysis.

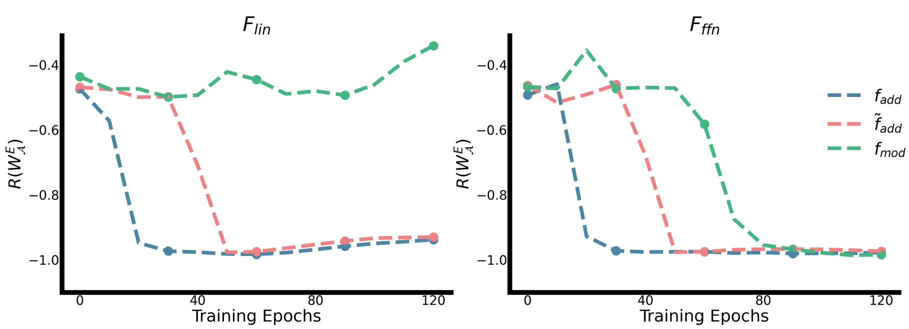


Figure 10: Dynamics of $R_{\text{order}}\left(\boldsymbol{W}_{\mathcal{A}}^{E}\right)$ in F_{lin} (left) and F_{ffn} (right). Line colors represent task types.

B.4 Umembedding matrix in Feedforward Network

Figure 11 displays the structure of the unembedding matrix in F_{ffn} with the three types of addition tasks. The distribution of $\cos\left(\boldsymbol{W}_{\nu}^{U}\right)$ (A) and the PCA projection (B) jointly reveal that the unembedding vectors of those label tokens establish a hierarchy structure, which is consistent with their natural sequence.

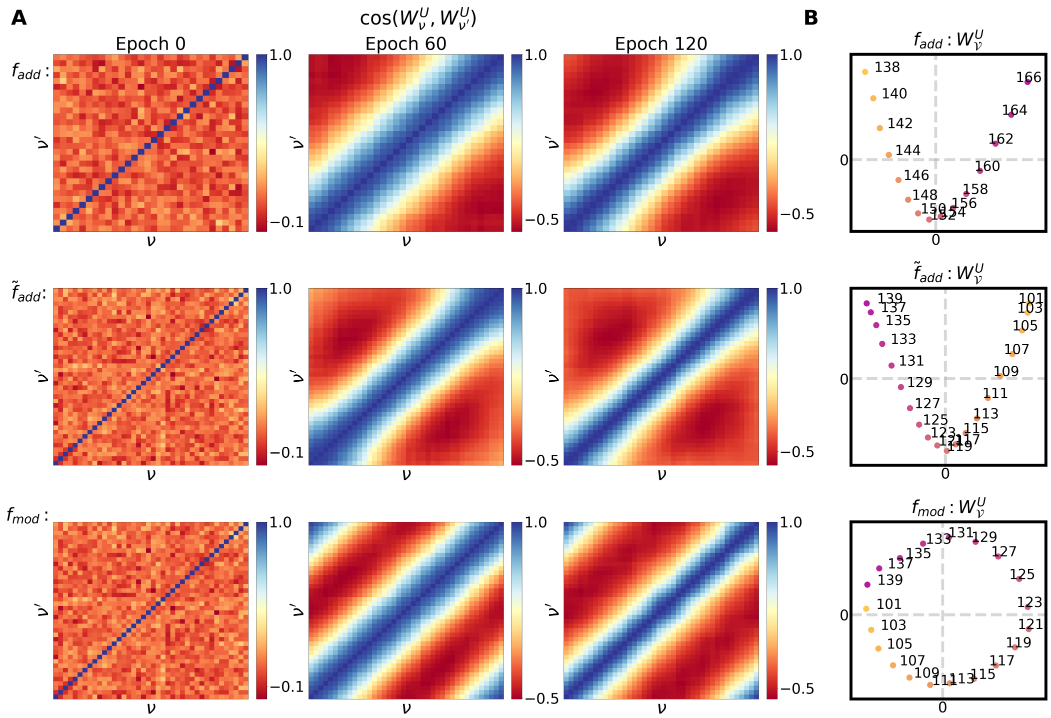


Figure 11: A: The heatmap of the $\cos\left(\boldsymbol{W}^U_{\mathcal{V}}\right)$ with label index in F_{ffn} during the training process. B: PCA projection of $\boldsymbol{W}^U_{\mathcal{V}}$ in F_{ffn} (epoch 120).

C Language Models

C.1 Complete results

Figure 12 represents the cosine similarity distribution of $\boldsymbol{W}^E, \boldsymbol{\phi}^{\text{next}}, \boldsymbol{W}^U$ and $\boldsymbol{\varphi}^{\text{pre}}$ at epoch 1 in the other 4 subsets of Pile we selected, exhibiting an analogous phenomenon with the observation in Figure 6. The distribution representations $\boldsymbol{\phi}^{\text{next}}$ and $\boldsymbol{\varphi}^{\text{pre}}$ could effectively capture the high similarity among embedding vectors and unembedding vectors, respectively. Figure 13 depicts the comparison at epoch 20.

C.2 Tied Embedding

In the Qwen2.5-3B-base model, $\boldsymbol{W}^E = \boldsymbol{W}^{U,T}$, which aims for computational source saving. Under this condition, we have that

$$\begin{aligned}\frac{d\boldsymbol{W}^E_s}{dt} &= r^{\text{in}}_s \boldsymbol{W}^{U,T} \boldsymbol{\phi}^{\text{next}}_s + r^{\text{out}}_s \boldsymbol{W}^E \boldsymbol{\varphi}^{\text{pre}}_s + \boldsymbol{\eta} \\ &= \boldsymbol{W}^E \left(r^{\text{in}}_s \boldsymbol{\phi}^{\text{next}}_s + r^{\text{out}}_s \boldsymbol{\varphi}^{\text{pre}}_s\right) + \boldsymbol{\eta}.\end{aligned}$$

Since the next-token-prediction, each token will be an input and an output, except the last token in a sequence, resulting in $r^{\text{in}}_s \approx r^{\text{out}}_s$. Denote $r_s = r^{\text{in}}_s$ and $\tilde{\boldsymbol{\phi}}_s = \boldsymbol{\phi}^{\text{next}}_s + \boldsymbol{\varphi}^{\text{pre}}_s$, then we have

$$\frac{d\boldsymbol{W}^E_s}{dt} = r_s \boldsymbol{W}^E \tilde{\boldsymbol{\phi}}_s + \boldsymbol{\eta}.$$

C.3 Probability Signature capture strong embedding similarities

We find that the probability signatures reflect the strong connections of embeddings more faithfully. As shown in Figure 14 A, the correlation between $\text{Corr}\left(\cos\left(\boldsymbol{W}^E_s, \boldsymbol{W}^E\right), \cos\left(\boldsymbol{\phi}^{\text{next}}_s, \boldsymbol{\phi}^{\text{next}}\right)\right)$ and

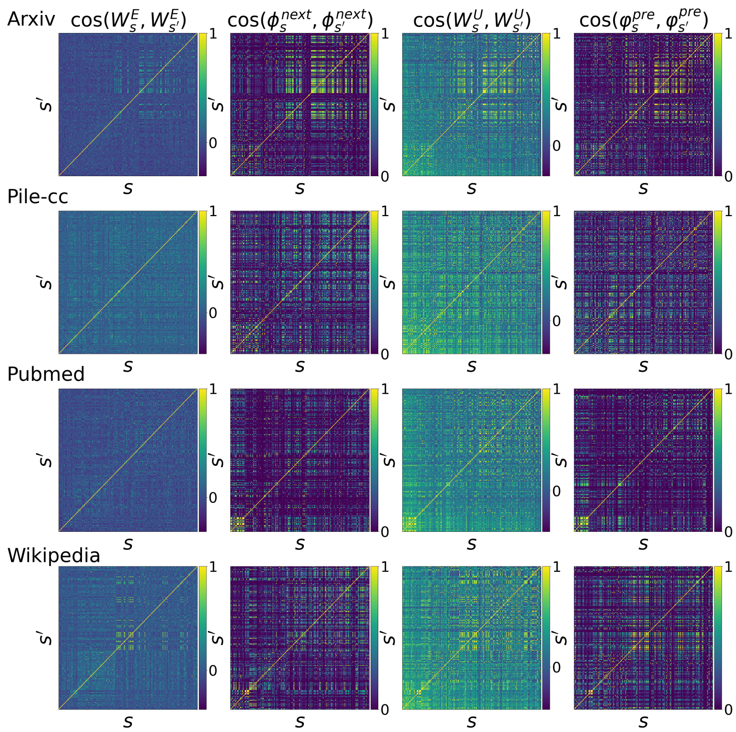


Figure 12: Heatmap of $\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}_{s'}^E\right)$ (left up), $\cos\left(\boldsymbol{\phi}_s^{\text{next}}, \boldsymbol{\phi}_{s'}^{\text{next}}\right)$ (right up), $\cos\left(\boldsymbol{W}_s^U, \boldsymbol{W}_{s'}^U\right)$ (left down) and $\cos\left(\boldsymbol{\varphi}_s^{\text{pre}}, \boldsymbol{\varphi}_{s'}^{\text{pre}}\right)$ (right up) (epoch 1) in each experiment with distinct dataset. The tokens displayed are those with the most appearances in the dataset.

$\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}^E\right)$ is plotted against for all tokens s, demonstrating stronger consistency in high-similarity regions. We define $p_{\cos(\boldsymbol{W}^E)}$ and $p_{\cos(\boldsymbol{\phi}^{\text{next}})}$ as the percentile matrix of each elements in $\cos\left(\boldsymbol{W}^E\right)$ and $\cos\left(\boldsymbol{\phi}^{\text{next}}\right)$, respectively. Figure 14 B displays the distribution of $p_{\cos(\boldsymbol{\phi}^{\text{next}})}$, conditioned on different intervals of the $p_{\cos(\boldsymbol{W}^E)}$, and Figure 14 C shows the average value of $p_{\cos(\boldsymbol{\phi}^{\text{next}})}$ within each interval of $p_{\cos(\boldsymbol{W}^E)}$. It can be observed that the alignment is significantly stronger in the regions with large embedding similarity.

Remark about Figure 14 A In each subset $D_i, i = 1, 2, \cdots M$, we define the set $\mathcal{S}_i = \left\{s_j^i\right\}_{j=1}^{C_i}$ as the set of the C_i tokens which appear most frequently in D_i. Based on the dataset D_i, and denote $\boldsymbol{W}^{E_i}$ as the embedding matrix of the model corresponding to dataset D_i, we compute that

$$\cos_{D_i}\left(\boldsymbol{W}_{s_j^i}^E, \boldsymbol{W}^E\right) = \left[\cos\left(\boldsymbol{W}_{s_j^i}^{E_i}, \boldsymbol{W}_{s'}^{E_i}\right)\right]_{s' \in \mathcal{S}_i} \in \mathbb{R}^{C_i},$$

and

$$\cos_{D_i}\left(\boldsymbol{\phi}_{s_j^i}^{\text{next}}, \boldsymbol{\phi}^{\text{next}}\right) = \left[\cos\left(\boldsymbol{\phi}_{s_j^i}^{\text{next}}, \boldsymbol{\phi}_{s'}^{\text{next}}\right)\right]_{s' \in \mathcal{S}_i} \in \mathbb{R}^{C_i}.$$

for any token $s_j^i \in \mathcal{S}_i$. Then we define the correlation coefficient

$$R_{D_i}\left(s_j^i\right) = \operatorname{Corr}\left(\cos_{D_i}\left(\boldsymbol{W}_{s_j^i}^E, \boldsymbol{W}^E\right), \cos_{D_i}\left(\boldsymbol{\phi}_{s_j^i}^{\text{next}}, \boldsymbol{\phi}^{\text{next}}\right)\right)$$

and the average embedding similarity as

$$\operatorname{Mean}_{\boldsymbol{W}^E, D_i}\left(s_j^i\right) = \frac{1}{C_i} \cos_{D_i}\left(\boldsymbol{W}_{s_j^i}^E, \boldsymbol{W}^E\right) \cdot \mathbf{1}.$$

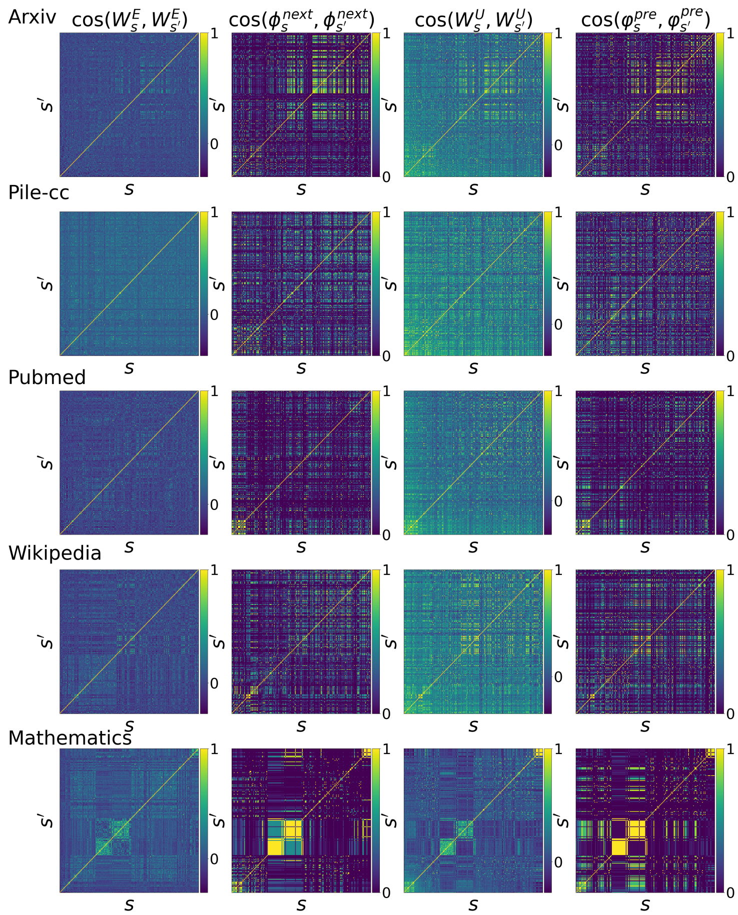


Figure 13: Heatmap of $\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}_{s'}^E\right)$ (left up), $\cos\left(\boldsymbol{\phi}_s^{\text{next}}, \boldsymbol{\phi}_{s'}^{\text{next}}\right)$ (right up), $\cos\left(\boldsymbol{W}_s^U, \boldsymbol{W}_{s'}^U\right)$ (left down) and $\cos\left(\boldsymbol{\varphi}_s^{\text{pre}}, \boldsymbol{\varphi}_{s'}^{\text{pre}}\right)$ (right up) (epoch 20) in each experiment with distinct dataset. The tokens displayed are those with the most appearances in the dataset.

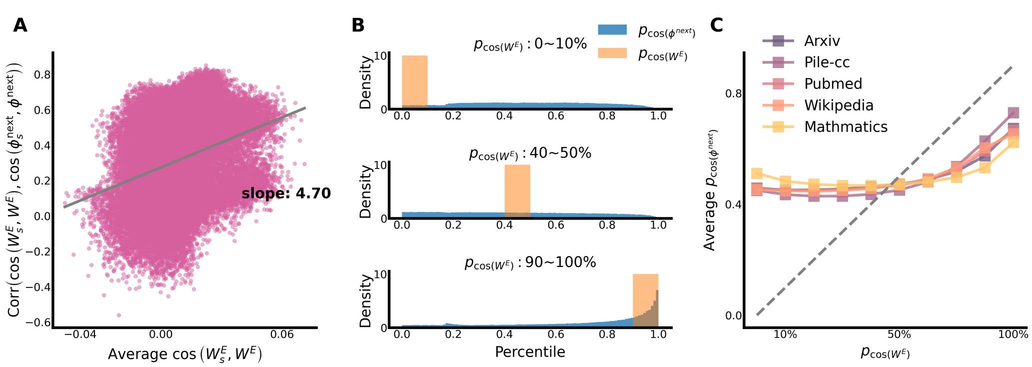


Figure 14: A: Relation between $\text{Corr}\left(\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}^E\right), \cos\left(\boldsymbol{\phi}_s^{\text{next}}, \boldsymbol{\phi}^{\text{next}}\right)\right)$ and the average value of $\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}^E\right)$. Each point denotes a token s. B: Distribution of $p_{\cos(\boldsymbol{\phi}^{\text{next}})}$, conditioned on intervals $0 \sim 10\%, 40 \sim 50\%$ and $90 \sim 100\%$ of the $p_{\cos(\boldsymbol{W}^E)}$. C: Average value of $p_{\cos(\boldsymbol{\phi}^{\text{next}})}$ within each interval of $p_{\cos(\boldsymbol{W}^E)}$.

Then we concatenate the metrics with all token $s_j^i \in \mathcal{S}_i, j = 1, 2, \cdots, C_i$ and all datasets $\mathcal{S}_i, i = 1, 2, \cdots, M$, i.e.

$$\text{Corr}\left(\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}^E\right), \cos\left(\boldsymbol{\phi}_s^{\text{next}}, \boldsymbol{\phi}^{\text{next}}\right)\right) = \left[R_{D_i}\left(s_j^i\right)\right]_{j=1,2,\cdots,C_i}^{i=1,2,\cdots,M} \in \mathbb{R}^{\sum_{i=1}^M C_i},$$
$$\text{Mean}\left(\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}^E\right)\right) = \left[\text{Mean}_{\boldsymbol{W}^E, D_i}\left(s_j^i\right)\right]_{j=1,2,\cdots,C_i}^{i=1,2,\cdots,M} \in \mathbb{R}^{\sum_{i=1}^M C_i}.$$

Figure 6 displays the relation between $\text{Corr}\left(\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}^E\right), \cos\left(\boldsymbol{\phi}_s^{\text{next}}, \boldsymbol{\phi}^{\text{next}}\right)\right)$ and $\text{Mean}\left(\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}^E\right)\right)$, revealing a positive correlation. In our work, $M = 5$, and we set up $C_i = 10000$ for each dataset.

Remark about Figure 14 B & C In each subset $D_i, i = 1, 2, \cdots M$, we define the set $\mathcal{S}_i = \left\{s_j^i\right\}_{j=1}^{C_i}$ as the set of the C_i tokens which appear most frequently in D_i. We compute that

$$\cos_{D_i}\left(\boldsymbol{W}^E\right) = \left[\cos\left(\boldsymbol{W}_s^{E_i}, \boldsymbol{W}_{s'}^{E_i}\right)\right]_{s,s' \in \mathcal{S}_i} \in \mathbb{R}^{C_i \times C_i}$$

and

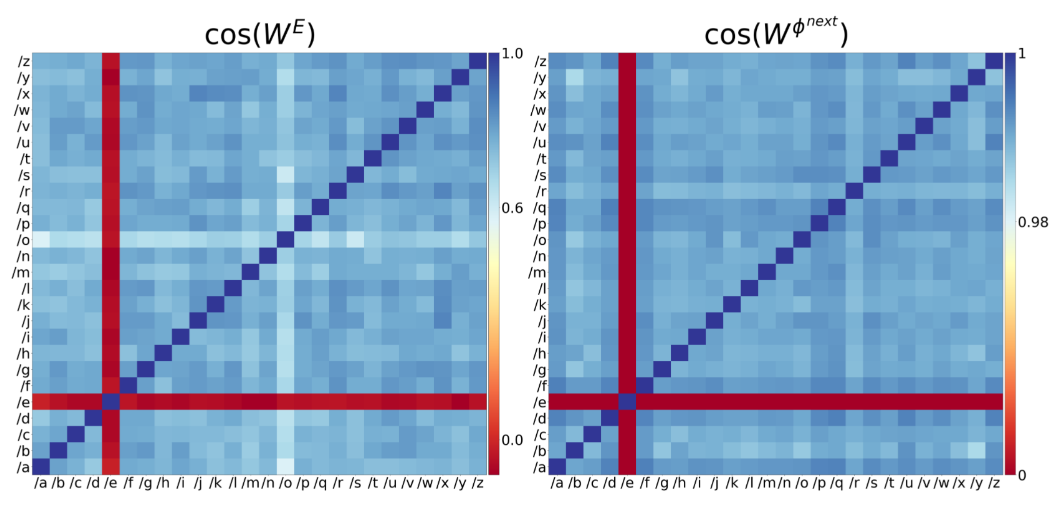
$$\cos_{D_i}\left(\boldsymbol{\phi}^{\text{next}}\right) = \left[\cos\left(\boldsymbol{\phi}_s^{\text{next}}, \boldsymbol{\phi}_{s'}^{\text{next}}\right)\right]_{s,s' \in \mathcal{S}_i} \in \mathbb{R}^{C_i \times C_i}.$$

Then translate the similarity matrix into a percentile formulation, i.e.

$$p_{\cos_{D_i}(\boldsymbol{W}^E)} = \text{Percentile}\left(\cos_{D_i}\left(\boldsymbol{W}^E\right)\right), \quad p_{\cos_{D_i}(\boldsymbol{\phi}^{\text{next}})} = \text{Percentile}\left(\cos_{D_i}\left(\boldsymbol{\phi}^{\text{next}}\right)\right)$$

and $p_{\cos(\boldsymbol{W}^E)} = \left[p_{\cos_{D_i}(\boldsymbol{W}^E)}\right]_{i=1,2,\cdots,M}$, $p_{\cos(\boldsymbol{\phi}^{\text{next}})} = \left[p_{\cos_{D_i}(\boldsymbol{\phi}^{\text{next}})}\right]_{i=1,2,\cdots,M}$. Figure 6 D and E reveal the distribution and average value of $p_{\cos(\boldsymbol{\phi}^{\text{next}})}$, where $k \times 10\% \leq p_{\cos(\boldsymbol{W}^E)} < (k+1) \times 10\%, k = 0, 1, 2, \cdots, 9$.

Case Analysis We provide a detailed case to explain the group of tokens exhibiting high embedding similarities. In experiments on the Pile-dm-mathematics dataset, tokens such as "$/a$", "$/b$", "$/c$", and "$/d$" often serve as denominators in mathematical expressions. Figure 15 shows the cosine similarities of both their embedding vectors and distribution representations, which are notably high for all tokens except "$/e$", which does not appear in the dataset. These tokens share highly similar semantics and also exhibit very similar next-token distributions, most frequently followed by "*" or ")". This similarity in next-token distribution leads to strong similarities in their embedding vectors. This example vividly illustrates how data distribution shapes semantic structure within the embedding space, particularly in the case of tokens with high semantic affinity.



Simplify ((k**(-9)*k*k/((k**0/k*k)/k)*k)/(k**(1/2))**(-26))/(((k*k**(2/3)*k)/k**0)/(k**(-1)*k)**14) assuming k is positive. k**(16/3)

Simplify (((a**31/a*a)/a)/a)/(a/(a/((a/(a/a**(6/5)))/a)))*(a*a**(-1/67))/((a*(a*a/(a*(a**14*a)/a))/a)/a) assuming a is positive. a**(14668/335)

Simplify ((j*(j**(-10/11)/j)/j)/j**(3/4)*(j/j**4)**7)**(-22/7) assuming j is positive.

Simplify (u*u**(-19))/u**(-1/3)*u/(u**(-5)*u)*u**0 assuming u is positive. u**(-38/3)

Simplify h*h/(h/(h**14/h)*h*h)*h/(h/(h/h**12))*h**(-11)/(((h**(-4)*h)/h)/h) assuming h is positive. h**(-5)

Simplify ((r/r**(-2/3))**4*r*r*r**(-2/9)*r*r/((r**0*r)/r))**33 assuming r is positive

Figure 15: A case analysis of the token group "$/a$", "$/b$", "$/c$", etc. The first row depicts the cosine similarity of their embeddings (left) and distribution representations (right). The second row exhibits the contexts containing these tokens, which are highlighted by different colors.

C.4 RESULTS OF LLAMA 2

To assess the generalizability of our analysis in Section 6 across different model architectures and tokenizers, we replicate the experiment using the Llama 2 architecture. We employ the same dataset from Pile, and the training configurations are the same as the experiments of Qwen2.5. As shown in Figure 16, the probability signatures effectively capture structural relationships in the embedding space, especially in regions exhibiting high embedding similarity. These results align closely with those in Figure 6, indicating that our analytical approach is robust to variations in model architecture.

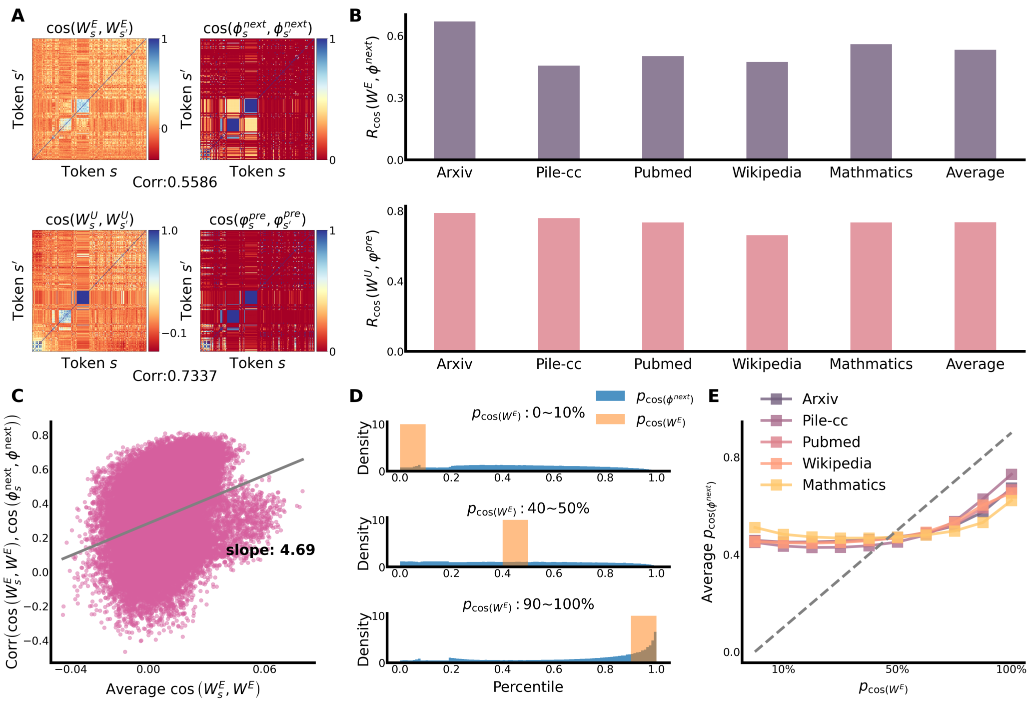


Figure 16: Results with Llama-2 architecture. A: Heatmap of the cosine similarity of $\boldsymbol{W}^E, \boldsymbol{W}^U, \boldsymbol{\phi}^{\text{next}}$ and $\boldsymbol{\varphi}^{\text{pre}}$. B: $R_{\cos}\left(\boldsymbol{W}^E, \boldsymbol{\phi}^{\text{next}}\right)$ (top) and $R_{\cos}\left(\boldsymbol{W}^U, \boldsymbol{\varphi}^{\text{pre}}\right)$ (bottom) with different datasets. C: Relation between $\text{Corr}\left(\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}^E\right), \cos\left(\boldsymbol{\phi}_s^{\text{next}}, \boldsymbol{\phi}^{\text{next}}\right)\right)$ and the average value of $\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}^E\right)$. Each point denotes a token s. D: Distribution of $p_{\cos(\boldsymbol{\phi}^{\text{next}})}$, conditioned on intervals $0 \sim 10\%, 40 \sim 50\%$ and $90 \sim 100\%$ of the $p_{\cos(\boldsymbol{W}^E)}$. E: Average value of $p_{\cos(\boldsymbol{\phi}^{\text{next}})}$ within each interval of $p_{\cos(\boldsymbol{W}^E)}$.

D Theoretical Details

D.1 Proof of Proposition 1

Lemma 1. *Given a model F and data pair $(\boldsymbol{X}, y) \in \mathbb{N}^{+,L} \times \mathbb{N}^{+}$, $\ell = -\log \mathit{Softmax}\left(F\left(\boldsymbol{X}\right)\right)_y$, we have that*

$$\frac{\partial \ell}{\partial F\left(\boldsymbol{X}\right)} = \boldsymbol{p} - \boldsymbol{e}_y, \tag{5}$$

where $\boldsymbol{p} = \mathit{softmax}\left(\boldsymbol{X}\right)$.

Proof. It's noted $\ell = -F\left(\boldsymbol{X}\right)_y + \log \sum_{j=1}^{d_{\text{vob}}} \exp F\left(\boldsymbol{X}\right)_j$, then we have

$$\frac{\partial \ell}{\partial F\left(X\right)_i} = -\delta_{i=y} + \frac{\exp F\left(\boldsymbol{X}\right)_i}{\sum_{j=1}^{d_{\text{vob}}} \exp F\left(\boldsymbol{X}\right)_j} = \boldsymbol{p}_i - \delta_{i=y},$$

where $\delta_{i=y} = 1$ if $i = y$ else 0. This indicates that $\frac{\partial \ell}{\partial F(\boldsymbol{X})} = \boldsymbol{p} - \boldsymbol{e}_y$. □

With Lemma 1, we could obtain the derivative of ℓ with respect to $\boldsymbol{W}_x^E$ for any $x \in \mathcal{V}$ as follows:

$$\begin{aligned}\frac{\partial \ell^i}{\partial \boldsymbol{W}_x^E} &= \frac{\partial F\left(\boldsymbol{X}^i\right)}{\partial \boldsymbol{W}_x^E} \frac{\partial \ell^i}{\partial F\left(\boldsymbol{X}^i\right)} \\ &= \left(\boldsymbol{W}^{U,T}\left(\boldsymbol{p}^i - \boldsymbol{e}_{y^i}\right)\right) \odot G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}^i}^E\right).\end{aligned}$$

Then the gradient flow of $\boldsymbol{W}_x^E$ could be obtained by

$$\frac{d\boldsymbol{W}_x^E}{dt} = -\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell^i}{\partial \boldsymbol{W}_x^E} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{W}^{U,T}\left(\boldsymbol{p}^i - \boldsymbol{e}_{y^i}\right)\right) \odot G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}^i}^E\right),$$

Since $\operatorname{diag}\left(G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}^i}^E\right)\right) = 0$ if $x \notin \boldsymbol{X}^i$, we have that

$$\begin{aligned}\frac{d\boldsymbol{W}_x^E}{dt} &= \frac{1}{N}\sum_{i=1}^{N_x^{\text{in}}}\left(\boldsymbol{W}^{U,T}\left(\boldsymbol{e}_{y_x^i} - \boldsymbol{p}_x^i\right)\right) \odot G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}_x^i}^E\right) \\ &= \frac{r_x^{\text{in}}}{N_x^{\text{in}}}\sum_{i=1}^{N_x^{\text{in}}}\left(\boldsymbol{W}^{U,T}\left(\boldsymbol{e}_{y_x^i} - \boldsymbol{p}_x^i\right)\right) \odot G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}_x^i}^E\right).\end{aligned}$$

Since that y_x^i takes value $\nu \in \mathcal{V}$, we can rewrite this formation as

$$\begin{aligned}\frac{d\boldsymbol{W}_x^E}{dt} &= \frac{r_x^{\text{in}}}{N_x^{\text{in}}}\left[\sum_{\nu\in\mathcal{V}}\sum_{i=1}^{N_{x,\nu}}\left(\boldsymbol{W}^{U,T}\boldsymbol{e}_\nu\right) \odot G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}_{x,\nu}^i}^E\right) - \sum_{i=1}^{N_x^{\text{in}}}\left(\boldsymbol{W}^{U,T}\boldsymbol{p}_x^i\right) \odot G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}_x^i}^E\right)\right] \\ &= r_x^{\text{in}}\left[\sum_{\nu\in\mathcal{V}}\left(\boldsymbol{W}^{U,T}\boldsymbol{e}_\nu\right) \odot \frac{N_{x,\nu}}{N_x^{\text{in}}}\frac{1}{N_{x,\nu}}\sum_{i=1}^{N_{x,\nu}} G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}_{x,\nu}^i}^E\right) - \frac{1}{N_x^{\text{in}}}\sum_{i=1}^{N_x^{\text{in}}}\left(\boldsymbol{W}^{U,T}\boldsymbol{p}_x^i\right) \odot G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}_x^i}^E\right)\right],\end{aligned}$$

where $N_x^{\text{in}}, N_{x,\nu}$ denotes the count of sequences containing x and the count of sequences containing x with label ν, $r_x^{\text{in}} = \frac{N_x^{\text{in}}}{N}, r_{x,\nu} = \frac{N_{s,\nu}}{N}$. Then let $N \to \infty$, by the law of large number we have

$$\begin{aligned}\frac{d\boldsymbol{W}_x^E}{dt} = r_x^{\text{in}}\Bigg(&\sum_{\nu\in\mathcal{V}}\mathbb{P}_\pi\left(y = \nu \mid x \in \boldsymbol{X}\right)\left(\boldsymbol{W}^{U,T}\boldsymbol{e}_\nu\right) \odot \mathbb{E}_\pi\left[G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right) \mid x \in \boldsymbol{X}, y = \nu\right] \\ &-\mathbb{E}_\pi\left[\left(\boldsymbol{W}^{U,T}\boldsymbol{p}\right) \odot G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right) \mid x \in \boldsymbol{X}\right]\Bigg).\end{aligned}$$

D.2 Proof of Proposition 2

Similar with the analysis of $\boldsymbol{W}_x^E$, we derive the gradient flow of $\boldsymbol{W}_\nu^U$ as follows:

$$
\begin{aligned}
\frac{d\boldsymbol{W}_\nu^U}{dt} &= -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ell^i}{\partial \boldsymbol{W}_\nu^U} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{e}_{y^{i,\nu}} - \boldsymbol{p}^{i,\nu}\right)\left[G\left(\boldsymbol{W}_{\boldsymbol{X}^i}^E\right)\right]^T.
\end{aligned}
$$

Since $\boldsymbol{e}_{y^{i,\nu}} = 1$ if $y^i = \nu$ else 0, we have that

$$
\frac{d\boldsymbol{W}_\nu^U}{dt} = \frac{r_\nu^{\text{out}}}{N_\nu^{\text{out}}}\sum_{i=1}^{N_\nu^{\text{out}}}\left[G\left(\boldsymbol{W}_{\boldsymbol{X}^i_{(\cdot,\nu)}}^E\right)\right]^T - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{p}^{i,\nu}\left[G\left(\boldsymbol{W}_{\boldsymbol{X}^i}^E\right)\right]^T,
$$

where N_ν^{out} denotes the count of sequences with label ν and $r_{\nu_j}^{\text{out}} = \frac{N_\nu^{\text{out}}}{N}$. Then let $N \to \infty$, by the law of large number we have

$$
\frac{d\boldsymbol{W}_\nu^U}{dt} = r_\nu^{\text{out}}\mathbb{E}_\pi\left[G\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right)^T \mid y = \nu\right] - \mathbb{E}_\pi\left[\boldsymbol{p}_\nu G\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right)^T\right].
$$

D.3 Proof of Corollary 1

With proposition 1, we have that

$$
\begin{aligned}
\frac{d\boldsymbol{W}_x^E}{dt} = r_x^{\text{in}}\Bigg(&\sum_{\nu\in\mathcal{V}}\mathbb{P}_\pi\left(y=\nu \mid x \in \boldsymbol{X}\right)\left(\boldsymbol{W}^{U,T}\boldsymbol{e}_\nu\right)\odot\mathbb{E}_\pi\left[G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right) \mid x \in \boldsymbol{X}\right] \\
&-\mathbb{E}_\pi\left[\left(\boldsymbol{W}^{U,T}\boldsymbol{p}\right)\odot G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right) \mid x \in \boldsymbol{X}\right]\Bigg).
\end{aligned}
$$

For the linear model, we have that $G^{(1)}\left(\boldsymbol{W}_{\boldsymbol{X}}^E\right) = \boldsymbol{1}$ if $x \in \boldsymbol{X}$. Utilizing that $\mathrm{softmax}\left(\boldsymbol{f}\right) = \frac{1}{d_{\text{vob}}}\boldsymbol{1} + \frac{1}{d_{\text{vob}}}\boldsymbol{f} + \mathcal{O}\left(d_{\text{vob}}^{-2}\boldsymbol{f}\right)$, we obtain that

$$
\begin{aligned}
\frac{d\boldsymbol{W}_x^E}{dt} &= \boldsymbol{W}^{U,T}r_x^{\text{in}}\left(\sum_{\nu\in\mathcal{V}}\mathbb{P}_\pi\left(y=\nu \mid x\in\boldsymbol{X}\right)\boldsymbol{e}_\nu - \mathbb{E}_\pi\left[\boldsymbol{p} \mid x \in \boldsymbol{X}\right]\right) \\
&= \boldsymbol{W}^{U,T}r_x^{\text{in}}\left(\boldsymbol{\phi}_x^y - \mathbb{E}_\pi\left[\frac{1}{d_{\text{vob}}}\boldsymbol{1} + \frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\sum_{x_i\in\boldsymbol{X}}\boldsymbol{W}_{x_i}^E + \mathcal{O}\left(d_{\text{vob}}^{-2}\boldsymbol{W}^U\boldsymbol{W}_x^E\right) \mid x\in\boldsymbol{X}\right]\right) \\
&= \boldsymbol{W}^{U,T}r_x^{\text{in}}\left(\boldsymbol{\phi}_x^y - \frac{1}{d_{\text{vob}}}\boldsymbol{1} - \frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\mathbb{E}_\pi\left[\sum_{x_i\in\boldsymbol{X}}\boldsymbol{W}_{x_i}^E \mid x \in \boldsymbol{X}\right] + \mathcal{O}\left(d_{\text{vob}}^{-2}\boldsymbol{W}^U\boldsymbol{W}_x^E\right)\right) \\
&= \boldsymbol{W}^{U,T}r_x^{\text{in}}\left(\boldsymbol{\phi}_x^y - \frac{1}{d_{\text{vob}}}\boldsymbol{1} - \frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\sum_{x'\in\mathcal{V}}\mathbb{P}_\pi\left(x'\in\boldsymbol{X} \mid x\in\boldsymbol{X}\right)\boldsymbol{W}_{x'}^E + \mathcal{O}\left(d_{\text{vob}}^{-2}\boldsymbol{W}^U\boldsymbol{W}_x^E\right)\right) \\
&= \boldsymbol{W}^{U,T}r_x^{\text{in}}\left(\boldsymbol{\phi}_x^y - \frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\boldsymbol{W}^E\boldsymbol{\phi}_x^{\boldsymbol{X}} - \frac{1}{d_{\text{vob}}}\boldsymbol{1} + \mathcal{O}\left(d_{\text{vob}}^{-2}\boldsymbol{W}^U\boldsymbol{W}_x^E\right)\right) \\
&:= \boldsymbol{W}^{U,T}r_x^{\text{in}}\left(\boldsymbol{\phi}_x^y - \frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\boldsymbol{W}^E\boldsymbol{\phi}_x^{\boldsymbol{X}} + \boldsymbol{\eta}\right),
\end{aligned}
$$

where $\boldsymbol{\eta} = -\frac{1}{d_{\text{vob}}}\boldsymbol{1} + \mathcal{O}\left(d_{\text{vob}}^{-2}\boldsymbol{W}^U\boldsymbol{W}_\alpha^E\right)$ contains the higher-order term and the data independent term.

D.4 PROOF OF COROLLARY 2

Proof. Since the small initialization, we assume that the activation function can be approximated by the following form with the Weierstrass approximation theorem.

$$\sigma\left(\sum_{x\in\boldsymbol{X}}\boldsymbol{W}_x^E\right)=C_0+C_1\left(\sum_{x\in\boldsymbol{X}}\boldsymbol{W}_x^E\right)+C_2\left(\sum_{x\in\boldsymbol{X}}\boldsymbol{W}_x^E\right)^{\odot 2}+\boldsymbol{\epsilon}.$$

With the loss of the generalization, we assume that $C_0=0, C_1=1, C_2=\frac{1}{2}$. Then we have

$$\begin{aligned}\frac{d\boldsymbol{W}_x^E}{dt}=r_x^{\text{in}}&\underbrace{\sum_{\nu\in\mathcal{V}}\mathbb{P}_\pi\left(y=\nu\mid x\in\boldsymbol{X}\right)\left(\boldsymbol{W}^{U,T}\boldsymbol{e}_\nu\right)\odot\mathbb{E}_\pi\left[\boldsymbol{1}+\sum_{x'\in\boldsymbol{X}}\boldsymbol{W}_{x'}^E\mid x\in\boldsymbol{X},y=\nu\right]}_{\boldsymbol{J}^y}\\&-r_x^{\text{in}}\underbrace{\mathbb{E}_\pi\left[\left(\boldsymbol{W}^{U,T}\boldsymbol{p}\right)\odot\left(\boldsymbol{1}+\sum_{x'\in\boldsymbol{X}}\boldsymbol{W}_{x'}^E\right)\mid x\in\boldsymbol{X}\right]}_{\boldsymbol{J}^p}.\end{aligned}$$

For the term $\boldsymbol{J}^y$ we have

$$\begin{aligned}\boldsymbol{J}^y&=\boldsymbol{W}^{U,T}\sum_{\nu\in\mathcal{V}}\mathbb{P}_\pi\left(y=\nu\mid x\in\boldsymbol{X}\right)\boldsymbol{e}_\nu+\sum_{\nu\in\mathcal{V}}\mathbb{P}_\pi\left(y=\nu\mid x\in\boldsymbol{X}\right)\left(\boldsymbol{W}^{U,T}\boldsymbol{e}_\nu\right)\odot\mathbb{E}_\pi\left[\sum_{x'\in\boldsymbol{X}}\boldsymbol{W}_{x'}^E\mid x\in\boldsymbol{X},y=\nu\right]\\&=\boldsymbol{W}^{U,T}\boldsymbol{\phi}_x^y+\sum_{\nu\in\mathcal{V}}\operatorname{diag}\left(\boldsymbol{W}_\nu^U\right)\sum_{x'\in\mathcal{V}}\mathbb{P}_\pi\left(y=\nu\mid x\in\boldsymbol{X}\right)\mathbb{P}_\pi\left(x'\in\boldsymbol{X}\mid x\in\boldsymbol{X},y=\nu\right)\boldsymbol{W}_{x'}^E.\end{aligned}$$

Since that $\mathbb{P}_\pi\left(y=\nu\mid x\in\boldsymbol{X}\right)\mathbb{P}_\pi\left(x'\in\boldsymbol{X}\mid x\in\boldsymbol{X},y=\nu\right)=\mathbb{P}_\pi\left(x'\in\boldsymbol{X},y=\nu\mid x\in\boldsymbol{X}\right)$, we have that

$$\begin{aligned}\boldsymbol{J}^y&=\boldsymbol{W}^{U,T}\boldsymbol{\phi}_x^y+\sum_{\nu,x'\in\mathcal{V}}\mathbb{P}_\pi\left(x'\in\boldsymbol{X},y=\nu\mid x\in\boldsymbol{X}\right)\boldsymbol{W}_\nu^U\odot\boldsymbol{W}_{x'}^E\\&=\boldsymbol{W}^{U,T}\boldsymbol{\phi}_x^y+\mathbb{T}\odot\boldsymbol{\phi}_x^{\boldsymbol{X}|y},\end{aligned}$$

where $\mathbb{T}\in\mathbb{R}^{d\times d_{\text{vob}}\times d_{\text{vob}}}$, $\mathbb{T}_{:,x',\nu}=\boldsymbol{W}_\nu^U\odot\boldsymbol{W}_{x'}^E$ for $\nu,x'\in\mathcal{V}$ and 0 otherwise.

Similarly, for the term $\boldsymbol{J}^p$, we have that

$$\begin{aligned}\boldsymbol{J}^p&=\mathbb{E}_\pi\left[\left(\boldsymbol{W}^{U,T}\left(\frac{1}{d_{\text{vob}}}\boldsymbol{1}+\frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\sum_{x'\in\boldsymbol{X}}\boldsymbol{W}_{x'}^E\right)\right)\odot\left(\boldsymbol{1}+\sum_{x'\in\boldsymbol{X}}\boldsymbol{W}_{x'}^E\right)\mid x\in\boldsymbol{X}\right]\\&=\frac{1}{d_{\text{vob}}}\boldsymbol{W}^{U,T}\boldsymbol{1}+\frac{1}{d_{\text{vob}}}\boldsymbol{W}^{U,T}\sum_{x'\in\mathcal{V}}\mathbb{P}_\pi\left(x'\in\boldsymbol{X}\mid x\in\boldsymbol{X}\right)\boldsymbol{W}_{x'}^E+\boldsymbol{\epsilon}\\&=\frac{1}{d_{\text{vob}}}\boldsymbol{W}^{U,T}\left(\boldsymbol{1}+\boldsymbol{W}^E\boldsymbol{\phi}_x^{\boldsymbol{X}}\right)+\boldsymbol{\epsilon},\end{aligned}$$

where $\boldsymbol{\epsilon}=\mathcal{O}\left(\frac{1}{d_{\text{vob}}^2}\boldsymbol{W}^U\boldsymbol{W}_\alpha^E\right)$. Then we have that

$$\frac{d\boldsymbol{W}_\alpha^E}{dt}=r_x^{\text{in}}\left(\boldsymbol{W}^{U,T}\boldsymbol{\phi}_x^y-\frac{1}{d_{\text{vob}}}\boldsymbol{W}^{U,T}\boldsymbol{W}^E\boldsymbol{\phi}_x^{\boldsymbol{X}}+\mathbb{T}\cdot\boldsymbol{\phi}_x^{\boldsymbol{X}|y}+\boldsymbol{\epsilon}\right),$$

where $\boldsymbol{\epsilon}=-\frac{1}{d_{\text{vob}}}\boldsymbol{W}^{U,T}\boldsymbol{1}+\mathcal{O}\left(\frac{1}{d_{\text{vob}}^2}\boldsymbol{W}^U\boldsymbol{W}_\alpha^E\right)$. □

D.5 Proof of Corollary 3

Proof. With Proposition 2, we have that

$$\begin{aligned}\frac{d\boldsymbol{W}_\nu^U}{dt} =&\frac{r_\nu^{\text{out}}}{N_\nu^{\text{out}}} r_\nu^{\text{out}} \mathbb{E}_\pi \left[\left(\sum_{x\in\boldsymbol{X}} \boldsymbol{W}_x^E\right)^T \mid y=\nu\right] - \mathbb{E}_\pi\left[\boldsymbol{p}_\nu \left(\sum_{x\in\boldsymbol{X}} \boldsymbol{W}_x^E\right)^T\right] \\ =& L r_\nu^{\text{out}} \sum_{x\in\mathcal{V}} \mathbb{P}_\pi\left(x\in\boldsymbol{X}\mid y=\nu\right)\boldsymbol{W}_x^{E,T} - \frac{1}{d_{\text{vob}}}\boldsymbol{W}^E\boldsymbol{1}+\boldsymbol{\epsilon} \\ =& L r_\nu^{\text{out}}\left(\boldsymbol{W}^E\boldsymbol{\varphi}_\nu^{\boldsymbol{X}}\right)^T - \boldsymbol{\eta},\end{aligned}$$

where $\boldsymbol{\eta} = -\frac{1}{d_{\text{vob}}}\boldsymbol{W}^E\boldsymbol{1} + \mathcal{O}\left(\frac{1}{d_{\text{vob}}}\boldsymbol{W}^E\boldsymbol{W}^E\boldsymbol{1}\right)$. □

D.6 Proof of Corollary 4

Proof. The next-token-prediction training loss could be formulated as

$$\ell^i = \frac{1}{L}\sum_{t=1}^{L-1}\text{CrossEntropy}\left(F_{\text{lan}}\left(\boldsymbol{X}_{:t}\right);\boldsymbol{e}_{\boldsymbol{X}_{t+1}}\right).$$

So we have that

$$\frac{\partial\ell^i}{\partial\boldsymbol{W}_s^E} = \frac{1}{L}\sum_{t=1}^{L-1}\boldsymbol{W}^{U,T}\left(\boldsymbol{p}_t^i - \boldsymbol{e}_{\boldsymbol{X}_{t+1}^i}\right)\odot\left(\delta_{\boldsymbol{X}_t^i=s}\boldsymbol{1}+\tilde{F}^{(1)}\left(\boldsymbol{X}_{:t}^i\right)\right).$$

Furthermore, we have that

$$\begin{aligned}\frac{d\boldsymbol{W}_s^E}{dt} =& \frac{1}{NL}\sum_{i=1}^N\sum_{t=1}^{L-1}\boldsymbol{W}^{U,T}\left(\boldsymbol{e}_{\boldsymbol{X}_{t+1}^i} - \boldsymbol{p}_t^i\right)\odot\left(\delta_{\boldsymbol{X}_t^i=s}\boldsymbol{1}+\tilde{F}^{(1)}\left(\boldsymbol{X}_{:t}^i\right)\right) \\ =& \frac{1}{NL}\boldsymbol{W}^{U,T}\sum_{i=1}^N\sum_{t=1}^{L-1}\delta_{\boldsymbol{X}_t^i=s}\boldsymbol{e}_{\boldsymbol{X}_{t+1}^i} + \frac{1}{NL}\boldsymbol{W}^{U,T}\sum_{i=1}^N\sum_{t=1}^{L-1}\boldsymbol{e}_{\boldsymbol{X}_{t+1}^i}\odot\tilde{F}^{(1)}\left(\boldsymbol{X}_{:t}^i\right) \\ &- \frac{1}{NL}\sum_{i=1}^N\sum_{t=1}^{L-1}\boldsymbol{W}^{U,T}\boldsymbol{p}_t^i\odot\left(\delta_{\boldsymbol{X}_t^i=s}\boldsymbol{1}+\tilde{F}^{(1)}\left(\boldsymbol{X}_{:t}^i\right)\right).\end{aligned}$$

Since the small initialization, assuming that $||\boldsymbol{W}||_\infty = \mathcal{O}\left(d^{-\gamma}\right)$ for any trainable parameter matrix $\boldsymbol{W}$, we have that $||\tilde{F}^{(1)}\left(\boldsymbol{X}_{:t}^i\right)||_\infty = \mathcal{O}\left(d^{1-2\gamma}\right)$ in the initial stage. Let $N\to\infty$, we have that

$$\frac{d\boldsymbol{W}_s^E}{dt} = r_s^{\text{in}}\boldsymbol{W}^{U,T}\left(\boldsymbol{\phi}_s^{\text{next}} - \boldsymbol{\eta}^E\right),$$

where $\boldsymbol{\eta}^E = \sum_{t=1}^{L-1}\mathbb{E}_\pi\left[\boldsymbol{p}\mid\boldsymbol{X}_t=s\right] + \mathcal{O}\left(d^{1-2\gamma}\boldsymbol{\phi}_s^{\text{next}}\right)$. Similarly, we have that

$$\frac{d\boldsymbol{W}_s^U}{dt} = \frac{1}{NL}\sum_{i=1}^N\sum_{t=1}^{L-1}\left(\delta_{\boldsymbol{X}_{t+1}^i=s} - \boldsymbol{p}_{\boldsymbol{X}_{:t}^i}^{i,s}\right)\left(\boldsymbol{W}_{\boldsymbol{X}_t^i}^{E,T} + \tilde{F}\left(\boldsymbol{X}_{:t}^i\right)^T\right),$$

where $\boldsymbol{p}_{\boldsymbol{X}_{:t}^i}^{i,s}$ means the s-th element of the output probability with input sequence $\boldsymbol{X}_{:t}^i$. Let $N\to\infty$, we have

$$\frac{d\boldsymbol{W}_s^U}{dt} = r_s^{\text{out}}\left(\boldsymbol{W}^E\boldsymbol{\varphi}_s^{\text{pre}}\right)^T + \boldsymbol{\eta}^U,$$

where $\boldsymbol{\eta}^U = \sum_{t=1}^{L-1}\mathbb{E}_\pi\left[\boldsymbol{p}_{\boldsymbol{X}_{:t}}^s\boldsymbol{W}_{\boldsymbol{X}_t}^{E,T}\right] + \mathcal{O}\left(r_s^{\text{out}}d^{1-2\gamma}\left(\boldsymbol{W}^E\boldsymbol{\varphi}_s^{\text{pre}}\right)^T\right)$. □