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

Junjie Yao[1,2] and Zhi-Qin John Xu[1,2,*]

[1]Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University
[2]School of Mathematical Sciences, Shanghai Jiao Tong University
[*]Corresponding author

**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 data distribution. 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) by training the Qwen2.5 architecture on the subsets of the Pile corpus. Our results show that the probability signatures are faithfully aligned with the embedding structures, particularly in capturing strong pairwise similarities among embeddings. Our work uncovers the mechanism of how data distribution guides the formation of embedding structures, establishing a novel understanding of the relationship between embedding organization and semantic patterns.

## 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, which quantify the relationship between model performance and factors such as model size, dataset size, and computational budget through power law relationships (Kaplan et al., 2020). While these scaling laws provide valuable quantitative predictions for model performance, they also reveal a concerning limitation: the power law relationship suggests that achieving further significant improvements may require prohibitively large increases in model and data size, making continued scaling increasingly impractical and resource-intensive.

One promising approach to address these limitations is to develop a deeper understanding of the underlying mechanisms that drive transformer models' success in natural language processing (NLP). The No Free Lunch theorem establishes that no single algorithm can perform optimally across all problem domains, highlighting the fundamental importance of understanding both the characteristics of the data and the properties of the algorithms that process it (Wolpert & Macready, 1997). Recent research has made significant progress in uncovering key properties of deep learning models, including 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), parameter distribution properties (Kovaleva et al., 2021; Dar et al., 2023), condensation phenomenon (Luo et al., 2021; Xu et al., 2025b), and embedding structure (Cai et al., 2021). There has also been some investigation into data characteristics—such as the power-law decay of correlations between elements (like pixels) as a function of their distance (Ruderman, 1994). However, there is a significant gap in understanding how these two fundamental aspects—algorithmic properties and data characteristics—interact to produce the remarkable performance.

The embedding space, which acts as the encoder of the language tokens, therefore provides an ideal entry point for investigating how algorithmic properties and data characteristics interact. Ideally, the structure of embeddings should reflect the semantic relationships among tokens. A concrete example involves digits such as $0, 1, 2, \ldots$, which possess a natural ordering. Their embedding vectors accordingly display an ordered structure consistent with this numerical sequence, reflecting basic reasoning capabilities in mathematical tasks (Mikolov et al., 2013b; Ethayarajh et al., 2019; Zhang et al., 2024; Yao et al., 2025). However, the cause of the consistency between embedding structures and semantic structures is still an open question, and the driving factors of the embedding structure are still not well characterized.

In this work, we identify a set of probability signatures that encode the semantic information in the data into the structure of the embedding space of language models. Such probability signatures are constructed based on label distribution, input distribution, input/output co-occurrence distribution, etc, that systematically capture inherent token-level relationships and reflect semantic structures. This result is achieved via utilizing an embedding-based model with gradient flow analysis of embedding vectors and unembedding vectors for well-designed variable-controlled experiments. We further extend our findings to LLMs with realistic corpora, such as the Qwen2.5 architecture (Team, 2024) and subsets of the Pile dataset (Gao et al., 2020; Biderman et al., 2022). The analysis approach with the controlled experiments offers a promising methodology to uncover more and more probability signatures that can bridge the data semantics and embedding structure in language models.

## 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. In contrast to prior work, our study focuses on the connection between embedding structure and data properties, offering a novel insight for understanding how embeddings are organized.

## 3 Embedding-based Model

We first explain the basic notation of embedding space. Given a vocabulary $\mathcal{V} \subset \mathbb{N}^+$ with size $d_{\text{vob}}$. We denote a trainable matrix $\boldsymbol{W}^E \in \mathbb{R}^{d \times d_{\text{vob}}}$ as the embedding matrix for $\mathcal{V}$, where $d$ is the hidden dimension. For any $x \in \mathcal{V}$, we denote $\boldsymbol{e}_x \in \mathbb{R}^{d_{\text{vob}}}$ as its one-hot vector and $\boldsymbol{W}^E_x := \boldsymbol{W}^E \boldsymbol{e}_x$ as the embedding vector of $x$, which is intuitively the $x$-th column of $\boldsymbol{W}^E$. For a sequence $\boldsymbol{X} = [x_1, x_2, \cdots, x_L] \subset \mathcal{V}$, we define its one-hot representation as $\boldsymbol{e}_{\boldsymbol{X}} := [\boldsymbol{e}_{x_1}, \boldsymbol{e}_{x_2}, \cdots, \boldsymbol{e}_{x_L}] \in \mathbb{R}^{d_{\text{vob}} \times L}$ and $\boldsymbol{W}^E_{\boldsymbol{X}} := [\boldsymbol{W}^E_{x_1}, \boldsymbol{W}^E_{x_2}, \cdots, \boldsymbol{W}^E_{x_L}] \in \mathbb{R}^{d \times L}$ as the embedding sequence of $\boldsymbol{X}$. Similarly, $\boldsymbol{W}^U \in \mathbb{R}^{d_{\text{vob}} \times d}$ represents the unembedding matrix and $\boldsymbol{W}^U_\nu := \boldsymbol{W}^{U,T} \boldsymbol{e}_\nu$ (the $\nu$-th row of $\boldsymbol{W}^U$) means the unembedding vector for any $\nu \in \mathcal{V}$.

We denote the models functioning on the 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 $d_{\text{vob}}$ and a sequence $\boldsymbol{X} \in \mathcal{V}^L$ with length $L$. An embedding-based model $F$ taking $\boldsymbol{X}$ as input could be formulated as*

$$F(\boldsymbol{X}) = \boldsymbol{W}^U G\left(\boldsymbol{W}^E_{\boldsymbol{X}}\right),$$

*where $G$ means the mapping in the hidden space.*

Embedding-based models have been widely applied in various domains, particularly in NLP. Investigating the characteristics of the embedding space via the gradient flow dynamics is essential for understanding the embedding structures. Given a dataset $\left\{\left(\boldsymbol{X}^i, y^i\right)\right\}_{i=1}^N$, we utilize cross-entropy as the loss function:

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

Denote that $\boldsymbol{p}^i = \text{Softmax}\left(F\left(\boldsymbol{X}^i\right)\right)$. Let $\odot$ represent the Hadamard (element-wise) product and $T$ mean the transpose, we obtain the following results.

**Proposition 1.** *Given an embedding-based model $F$ with an embedding matrix $\boldsymbol{W}^E$. For any token $x \in \mathcal{V}$, the gradient flow of $\boldsymbol{W}^E_x$ (the $x$-th column of $\boldsymbol{W}^E$) can be formulated as:*

$$\frac{d\boldsymbol{W}^E_x}{dt} = \sum_{\nu \in \mathcal{V}} \frac{r_{x,\nu}}{N_{x,\nu}} \left(\boldsymbol{W}^{U,T} \boldsymbol{e}_\nu\right) \odot \sum_{i=1}^{N_{x,\nu}} G^{(1)}\left(\boldsymbol{W}^E_{\boldsymbol{X}^i_{(x,\nu)}}\right) - \frac{r^{\text{in}}_x}{N^{\text{in}}_x} \sum_{i=1}^{N^{\text{in}}_x} G^{(1)}\left(\boldsymbol{W}^E_{\boldsymbol{X}^i_x}\right) \odot \left(\boldsymbol{W}^{U,T} \boldsymbol{p}^i_x\right),$$

*where $N^{\text{in}}_x, N_{x,\nu}$ denotes the count of sequences containing $x$ and the count of sequences containing $x$ with label $\nu$, $r^{\text{in}}_x = \frac{N^{\text{in}}_x}{N}, r_{x,\nu} = \frac{N_{s,\nu}}{N}$, $G^{(1)}$ represents the derivative of $G$ with respect to $\boldsymbol{W}^E_x$. $\boldsymbol{X}^i_x$ is the $i$-th sequence containing $x$ and $\boldsymbol{X}^i_{(x,\nu)}$ denotes the $i$-th sequence containing $x$ with label $\nu$. $\boldsymbol{p}^i_x$ means the output probability of the model over $\boldsymbol{X}^i_x$.*

**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 $\boldsymbol{W}^U_\nu$ (the $\nu$-th row of $\boldsymbol{W}^U$) can be written as*

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

*where $N^{\text{out}}_\nu$ denotes the count of sequences with label $\nu$ and $r^{\text{out}}_{\nu_j} = \frac{N^{\text{out}}_\nu}{N}$. $\boldsymbol{X}^i_{(\cdot,\nu)}$ means the $i$-th sample which takes $\nu$ as the label. $\boldsymbol{p}^{i,\nu}$ means the $\nu$-th element of $\boldsymbol{p}^i$.*

In this work, we employ three embedding-based architectures:

- Linear model. $F_{\text{lin}}(\boldsymbol{X}) = \boldsymbol{W}^U \sum_{x \in \boldsymbol{X}} \boldsymbol{W}^E_x$.
- Feedforward network. $F_{\text{ffn}}(\boldsymbol{X}) = \boldsymbol{W}^U \sigma\left(\sum_{x \in \boldsymbol{X}} \boldsymbol{W}^E_x\right)$, where $\sigma$ denotes the element-wise nonlinear activation.
- Transformer-based architecture. We employ the Qwen2.5 architecture in Section 6 and the Llama 2 architecture (Touvron et al., 2023) in Appendix C.4.

## 4 Probability Signature

In the field of deep learning, the data characteristics play a critically important role in both the training dynamics and the final performance of the model. Given a training dataset $\left\{\left(\boldsymbol{X}^i, y^i\right)\right\}_{i=1}^N$ sampled from distribution $\pi$, we define the following *probability signatures*:

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

$$\boldsymbol{\phi}^{\boldsymbol{X}|y}_x = \sum_{\nu \in \mathcal{V}} \boldsymbol{e}_\nu \left(\sum_{x' \in \mathcal{V}} \mathbb{P}\left(x' \in \boldsymbol{X} \mid x \in \boldsymbol{X}, y = \nu\right) \boldsymbol{e}_{x'}\right)^T, \quad \boldsymbol{\varphi}^{\boldsymbol{X}}_\nu = \sum_{x \in \mathcal{V}} \mathbb{P}_\pi\left(x \in \boldsymbol{X} \mid y = \nu\right) \boldsymbol{e}_x.$$

$\phi_x^y$ denotes the distribution of the label given that the input sequence $\boldsymbol{X}$ contains element $x$. It captures the association between a specific input element $x$ and its label. $\phi_x^{\boldsymbol{X}}$ represents the probability that, given $\boldsymbol{X}$ contains $x$, it also contains another input element $x'$. It characterizes the co-occurrence relationship between different elements. $\phi_x^{\boldsymbol{X}|y}$ indicates the probability that, when $\boldsymbol{X}$ contains $x$ and the label is fixed, the sequence additionally includes $x'$. It further delineates the relationship among different elements in the sequence under a specified label condition. $\varphi_\nu^{\boldsymbol{X}}$ denotes the probability distribution of the element $x$ contained in $\boldsymbol{X}$ conditional with the label being $\nu$. It reflects the dependency between input elements and the label from a reverse perspective.

## 5 Addition Task

The addition task has become an important benchmark for studying the characteristics of language models. Many studies have found that the digits' embeddings exhibit an ordered structure consistent with the natural sequence of numbers (Zhang et al., 2024; Yao et al., 2025). To demonstrate the significant influence of probability signatures on the model's embedding space, we design three types of composite addition tasks to perform *variable-controlled experiments*. 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 $\alpha_1$ means addition with $\alpha_1$. Given a input sequence $\boldsymbol{X} = [z, \alpha_1, \alpha_2]$, which means the composite function $(\alpha_1, \alpha_2)$ on the key $z$, we define the following tasks:

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

In this work, we set $\mathcal{A} = \{11, 12, \cdots, 20\}$ and $\mathcal{Y} = \mathcal{Z} = \{101, 102, \cdots, 140\}$. Figure 1 A displays the probability signature, which is distinct across the $\alpha$ in each task, revealing that the difference among $\alpha$ lies in the global horizontal shift. The detailed formulations are provided in Appendix B.1.

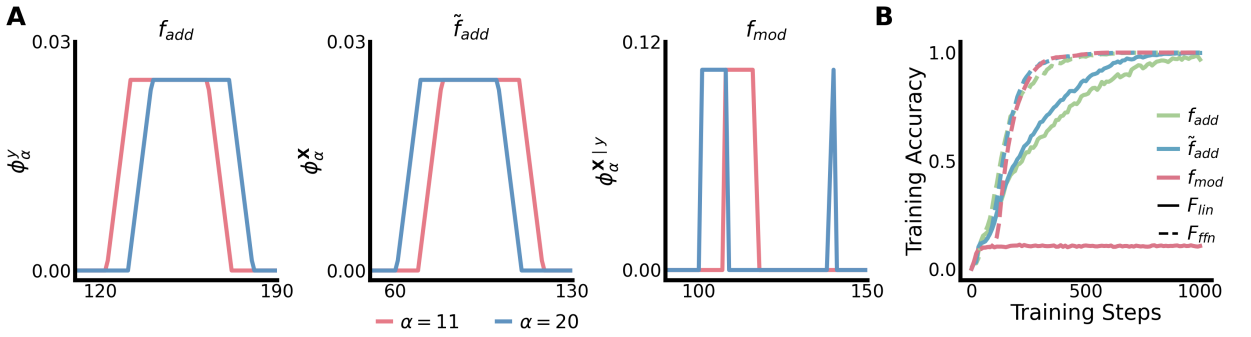



Figure 1: A: Probability signature which is distinct with varying $\alpha$ in each task ($f_{\text{add}} \to \phi_\alpha^y, \tilde{f}_{\text{add}} \to \phi_\alpha^{\boldsymbol{X}}, f_{\text{mod}} \to \phi_\alpha^{\boldsymbol{X}|y}$), $\alpha = 11$ (red) and 20 (blue). In $\phi_\alpha^{\boldsymbol{X}|y}$, it's displayed with $y = 160$. B: Training accuracy of the $F_{\text{lin}}$ and $F_{\text{ffn}}$ on the three addition tasks.

For notation convenience, we denote that $\boldsymbol{W}_{\mathcal{A}}^E = \left[\boldsymbol{W}_\alpha^E\right]_{\alpha\in\mathcal{A}}, \boldsymbol{\phi}_{\mathcal{A}} = [\boldsymbol{\phi}_\alpha]_{\alpha\in\mathcal{A}}$ for $\boldsymbol{\phi}_\alpha = \phi_\alpha^y, \phi_\alpha^{\boldsymbol{X}}, \phi_\alpha^{\boldsymbol{X}|y}$, and $\cos\left(\boldsymbol{W}_{\mathcal{A}}^E\right) := \left[\cos\left(\boldsymbol{W}_\alpha^E, \boldsymbol{W}_{\alpha'}^E\right)\right]_{\alpha,\alpha'\in\mathcal{A}}, \cos\left(\boldsymbol{\phi}_{\mathcal{A}}\right) := \left[\cos\left(\boldsymbol{\phi}_\alpha, \boldsymbol{\phi}_{\alpha'}\right)\right]_{\alpha,\alpha'\in\mathcal{A}}$. Similarly, $\cos\left(\boldsymbol{W}_{\mathcal{V}}^U\right) := \left[\cos\left(\boldsymbol{W}_\nu^U, \boldsymbol{W}_{\nu'}^U\right)\right]_{\nu,\nu'\in\mathcal{V}}$ and $\cos\left(\boldsymbol{\varphi}_{\mathcal{V}}^{\boldsymbol{X}}\right) := \left[\cos\left(\boldsymbol{\varphi}_\nu, \boldsymbol{\varphi}_{\nu'}\right)\right]_{\nu,\nu'\in\mathcal{V}}$.

### 5.1 Results

We train these addition tasks using the $F_{\text{lin}}$ and $F_{\text{ffn}}$ with $d = 200$. 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. The

complete training configurations are provided in Appendix A. Figure 1 B shows the training accuracy of $F_{\text{lin}}$ and $F_{\text{ffn}}$ on the three addition tasks. The results reveal that both $f_{\text{add}}$ and $\tilde{f}_{\text{add}}$ are learned well by the linear model, whereas $f_{\text{mod}}$ requires the nonlinear model to achieve an effective fit.

### 5.2 Embedding Matrix

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:

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

$R_{\text{order}}\left(\boldsymbol{W}_{\mathcal{A}}^{E}\right)$ 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 2 A represents the distribution of $\cos\left(\boldsymbol{W}_{\mathcal{A}}^{E}\right)$ for the three tasks with $F_{\text{lin}}$ and $F_{\text{ffn}}$, respectively, and Figure 2 B depicts the corresponding evolution of $R_{\text{order}}\left(\boldsymbol{W}_{\mathcal{A}}^{E}\right)$. In the case of $f_{\text{add}}$, anchor embeddings quickly form an ordered structure, where the cosine similarity gets smaller as the anchor distance gets larger. For the task $\tilde{f}_{\text{add}}$, the anchor embeddings also develop a similar hierarchical structure. However, its construction requires more steps, indicating that the driving factors of the structure in $f_{\text{add}}$ and $\tilde{f}_{\text{add}}$ are different. In $f_{\text{mod}}$, although the linear model fails to learn it effectively, the anchor embeddings still undergo noticeable changes from the initial stage. Specifically, all embedding vectors become aligned in nearly the same direction. Furthermore, the anchor embeddings of $f_{\text{mod}}$ in $F_{\text{ffn}}$ construct an ordered structure with more steps, suggesting that the activation provides another factor in deriving such an embedding structure.

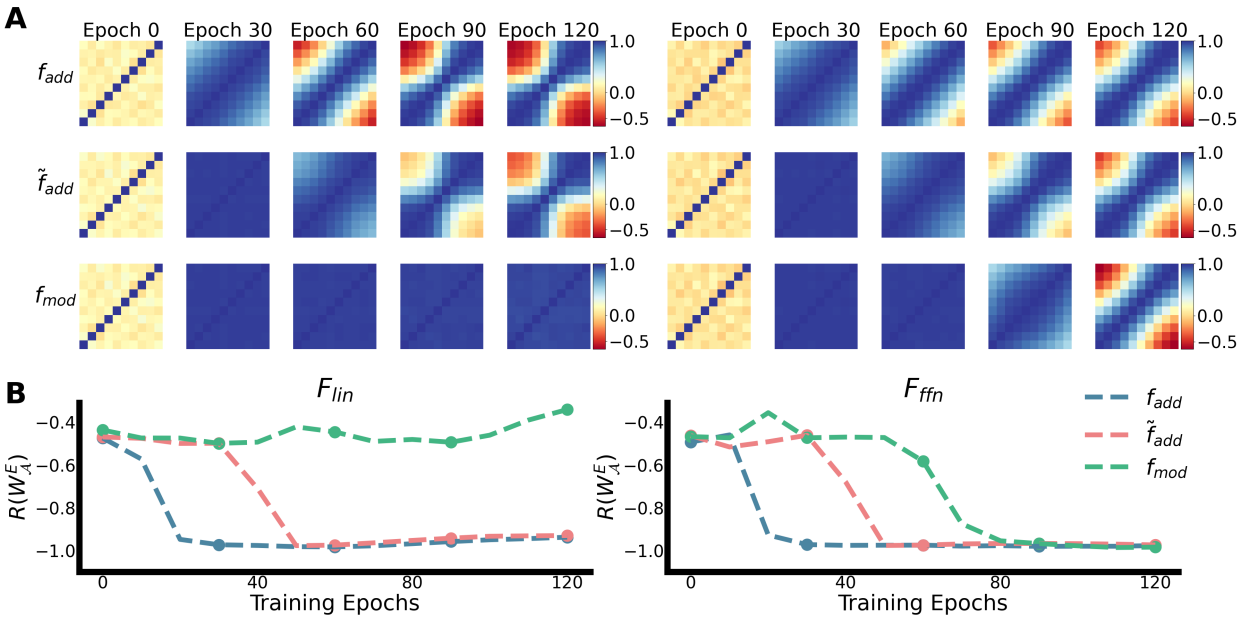



Figure 2: A: The heatmap of $\cos\left(\boldsymbol{W}_{\mathcal{A}}^{E}\right)$ in $F_{\text{lin}}$ (left) and $F_{\text{ffn}}$ (right) during the training process. Each row corresponds to $f_{\text{add}}$, $\tilde{f}_{\text{add}}$, and $f_{\text{mod}}$, respectively. B: Dynamics of $R_{\text{order}}\left(\boldsymbol{W}_{\mathcal{A}}^{E}\right)$ in $F_{\text{lin}}$ (left) and $F_{\text{ffn}}$ (right). Line colors represent task types.

We derive the reasons for the embedding structure in each task from the gradient flow. With Proposition 1, we obtain the following approximation:

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

$$\frac{d\boldsymbol{W}_{\alpha}^{E}}{dt}=\boldsymbol{W}^{U,T}r_{x}^{\text{in}}\left(\boldsymbol{\phi}_{\alpha}^{y}-\frac{1}{d_{\text{vob}}}\boldsymbol{W}^{U}\boldsymbol{W}^{E}\boldsymbol{\phi}_{\alpha}^{\boldsymbol{X}}+\boldsymbol{\eta}\right),\tag{1}$$

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

**Remark 1.** *Deep learning methods fundamentally comprise three essential components: the model, the data, and the optimization algorithm. Corollary 1 clearly illustrates how these three elements are coupled (model: $F_{\text{lin}}$; data: probability signatures; optimization algorithm: the gradient flow) and influence the embedding space, providing crucial insights for understanding and investigating their joint impact on the performance of deep learning approaches.*

Corollary 1 indicates that the dynamics of $\boldsymbol{W}_\alpha^E$ in $F_{\text{lin}}$ are primarily impacted by the probability signatures $\boldsymbol{\phi}_\alpha^y$ and $\boldsymbol{\phi}_\alpha^{\boldsymbol{X}}$, demonstrating the connection between data distribution and the embedding space. As we mentioned, the $\boldsymbol{\phi}_\alpha^y$ is distinct for different $\alpha$, while the $\boldsymbol{\phi}_\alpha^{\boldsymbol{X}}$ is identical for all $\alpha$ in $f_{\text{add}}$; the opposite holds for $\tilde{f}_{\text{add}}$. Figure 3 A depicts $\cos\left(\boldsymbol{\phi}_{\mathcal{A}}^y\right)$ (left) and $\cos\left(\boldsymbol{\phi}_{\mathcal{A}}^{\boldsymbol{X}}\right)$ (middle column) in $f_{\text{add}}$ (top) and $\tilde{f}_{\text{add}}$ (middle row), revealing that $\boldsymbol{\phi}_\alpha^y$ and $\boldsymbol{\phi}_\alpha^{\boldsymbol{X}}$ are significant for the formation of the hierarchy embedding structure in $f_{\text{add}}$ and $\tilde{f}_{\text{add}}$, respectively. Furthermore, (1) indicates that $\boldsymbol{\phi}_\alpha^y$ acts as a leading term and the effect of $\boldsymbol{\phi}_\alpha^{\boldsymbol{X}}$ is weaker in the early training process since it times a small magnitude term $\frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\boldsymbol{W}^E$. This results in the formation speed of the structure in $\tilde{f}_{\text{add}}$ being slower than $f_{\text{add}}$, which is consistent with the phenomenon in Figure 2.

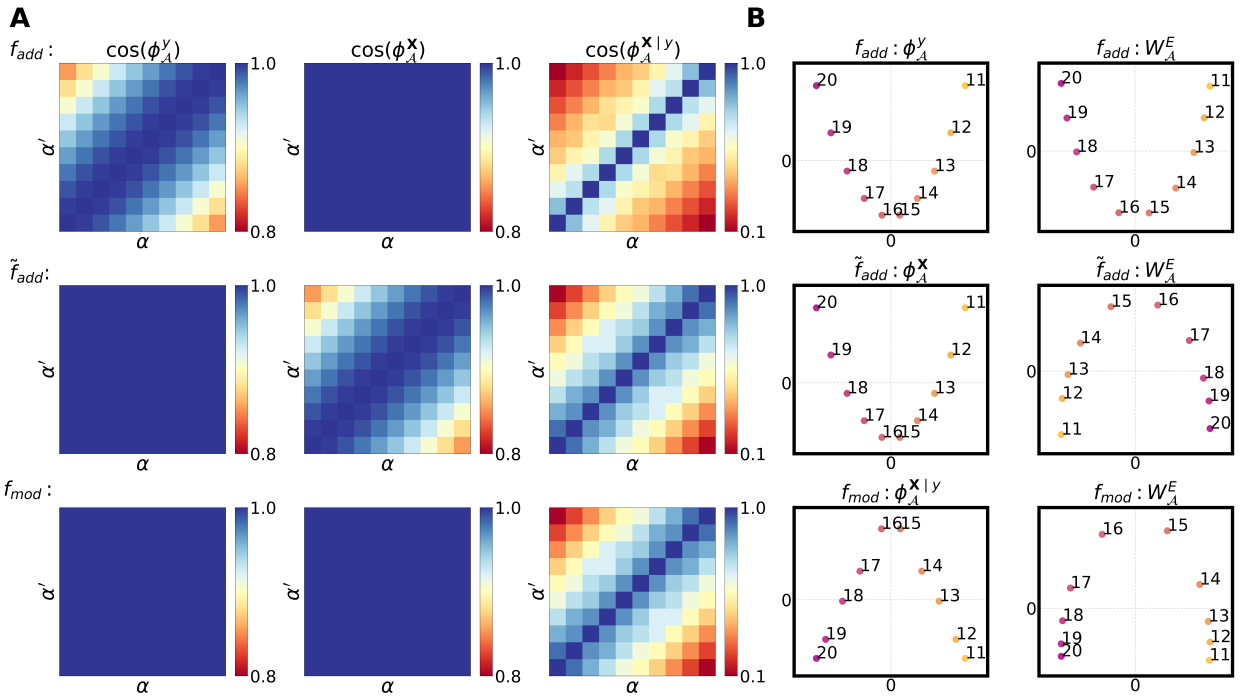



Figure 3: A: Cosine similarity among different anchor $\alpha$ of $\boldsymbol{\phi}_\alpha^y, \boldsymbol{\phi}_\alpha^{\boldsymbol{X}}, \boldsymbol{\phi}_\alpha^{\boldsymbol{X}|y}$ (see (4)) in each task. B: The PCA projection of the key factors ($f_{\text{add}} \to \boldsymbol{\phi}_\alpha^y$, $\tilde{f}_{\text{add}} \to \boldsymbol{\phi}_\alpha^{\boldsymbol{X}}$, $f_{\text{mod}} \to \boldsymbol{\phi}_\alpha^{\boldsymbol{X}|y}$) and the embedding vectors in different tasks ($F_{\text{ffn}}$, 120 epoch).

In task $f_{\text{mod}}$, $\boldsymbol{\phi}_\alpha^y$ and $\boldsymbol{\phi}_\alpha^{\boldsymbol{X}}$ are both identical across different anchors $\alpha$. Figure 3 A (bottom) indicates that $\cos\left(\boldsymbol{\phi}_{\mathcal{A}}^y\right)$ and $\cos\left(\boldsymbol{\phi}_{\mathcal{A}}^{\boldsymbol{X}}\right)$ are 1 for all anchor pairs, which leads these embedding vectors to converge to almost the same direction, consistent with the observation in $F_{\text{lin}}$. To identify the key factors that contribute to the formation of the ordered embedding structure for $f_{\text{mod}}$ in $F_{\text{ffn}}$, we perform a similar analysis of its gradient flow and obtain the following result.

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

$$\frac{d\boldsymbol{W}_\alpha^E}{dt} = \mathbb{T} \cdot \left(\boldsymbol{\phi}_\alpha^{\boldsymbol{X}|y}\right)^T + \boldsymbol{\eta}_{\boldsymbol{\phi}_\alpha^y} + \frac{1}{d_{\text{vob}}}\boldsymbol{\eta}_{\boldsymbol{\phi}_\alpha^{\boldsymbol{X}}} + \tilde{\boldsymbol{\eta}}, \tag{2}$$

*where $\mathbb{T} \in \mathbb{R}^{d \times d_{\text{vob}} \times d_{\text{vob}}}$, $\mathbb{T}_{:,:,\nu} = r_{\alpha,\nu}\text{diag}\left(\boldsymbol{W}_\nu^U\right)\boldsymbol{W}^E$ for $\nu \in \mathcal{V}$ and 0 otherwise. $\boldsymbol{\eta}_{\boldsymbol{\phi}_\alpha^y}$ and $\boldsymbol{\eta}_{\boldsymbol{\phi}_\alpha^{\boldsymbol{X}}}$ denotes the term related with $\boldsymbol{\phi}_\alpha^y$ and $\boldsymbol{\phi}_\alpha^{\boldsymbol{X}}$, respectively. $\tilde{\boldsymbol{\eta}}$ represents the higher-order term.*

Corollary 2 indicates that the ordered embedding structure of $f_{\text{mod}}$ primarily relies on probability signature $\boldsymbol{\phi}_{\alpha}^{\boldsymbol{X}|y}$. Figure 3 A (right) depicts the $\cos\left(\boldsymbol{\phi}_{\mathcal{A}}^{\boldsymbol{X}|y}\right)$ in tasks, which reveals that $\boldsymbol{\phi}_{\alpha}^{\boldsymbol{X}|y}$ in $f_{\text{mod}}$ constructs an ordered structure, resulting in the ordered embedding structure in $F_{\text{ffn}}$. Furthermore, Figure 3 C shows the PCA projection on probability signatures and the embedding space in $F_{\text{ffn}}$, revealing a high consistency. This comparison demonstrates the impact of the probability signatures in shaping the embedding space.

### 5.3 Unembedding Matrix

The $i$-th row of the unembedding matrix can also be viewed as the feature for the $i$-th token. As shown in Figure 4 A, a similar ordered structure emerges among the unembedding vectors with the label index in $F_{\text{lin}}$ across all tasks. Specifically, $\boldsymbol{W}_{\nu}^{U}$ in $f_{\text{mod}}$ constructs a ring where the similarity between small $\nu$ and large $\nu'$ is also large since $\mathcal{Z}_{\max}+1=\mathcal{Z}_{\min}$ in $f_{\text{mod}}$.

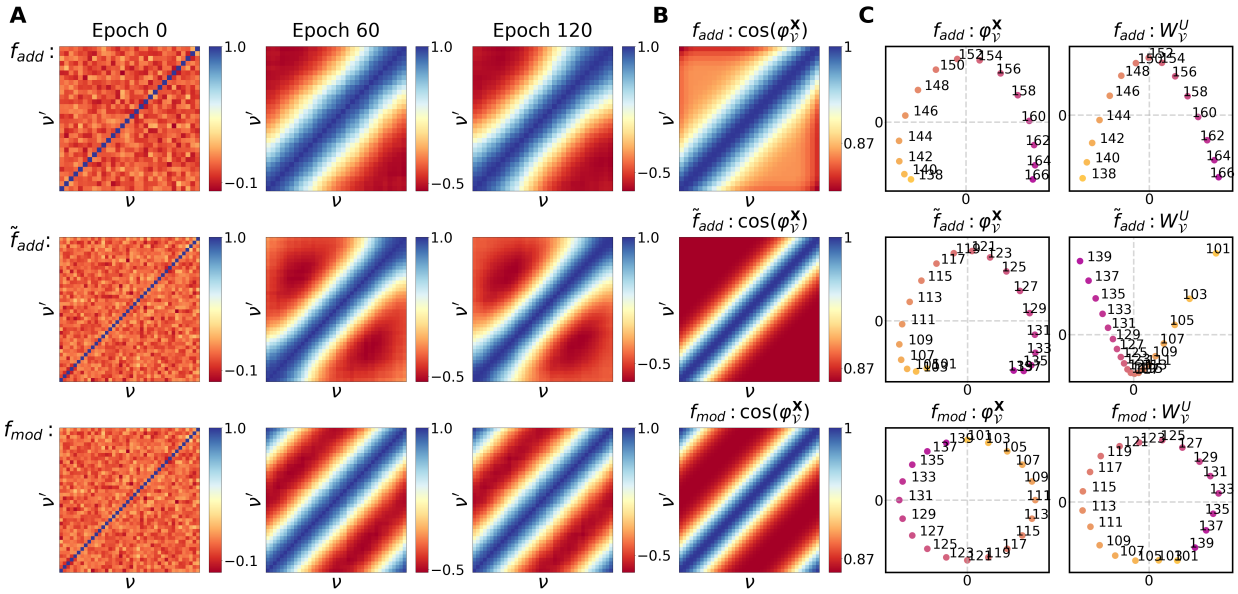



Figure 4: A: The heatmap of the $\cos\left(\boldsymbol{W}_{\nu}^{U}\right)$ with label index in $F_{\text{lin}}$ during the training process. B: The heatmap of $\cos\left(\boldsymbol{\varphi}_{\nu}^{\boldsymbol{X}}\right)$ across different tasks. C: PCA projection of $\boldsymbol{\varphi}_{\nu}^{\boldsymbol{X}}$ and $\boldsymbol{W}_{\nu}^{U}$ (epoch 120).

Similarly, we identify the driving factors of this specific structure by examining the gradient flow of $\boldsymbol{W}^{U}$. Since this phenomenon occurs in both $F_{\text{lin}}$ and $F_{\text{ffn}}$, it suffices to analyze the linear model. Based on Proposition 2, we derive the following result:

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

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

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

Corollary 3 illustrates that $\boldsymbol{\varphi}_{\nu}^{\boldsymbol{X}}$ plays a significant role in shaping the unembedding matrix. Figure 4 B depicts the distribution of $\cos\left(\boldsymbol{\varphi}_{\mathcal{V}}^{\boldsymbol{X}}\right)$, which is aligned with the distribution of the $\cos\left(\boldsymbol{W}_{\mathcal{V}}^{U}\right)$ in Figure 4 A. Furthermore, Figure 4 C compares the PCA projection of $\boldsymbol{\varphi}_{\mathcal{V}}^{\boldsymbol{X}}$ and $\boldsymbol{W}_{\mathcal{V}}^{U}$ in all tasks, revealing a high consistency and validating our analysis.

## 6 Language Model

We have demonstrated the influence of data distribution on the embedding space in the addition tasks. In this section, we explore how to extend this analysis to real-world language models. Most contemporary language models are built upon the Transformer decoder architecture. Assuming the input sequence is denoted as $\boldsymbol{X}$ with length $L$, we define a

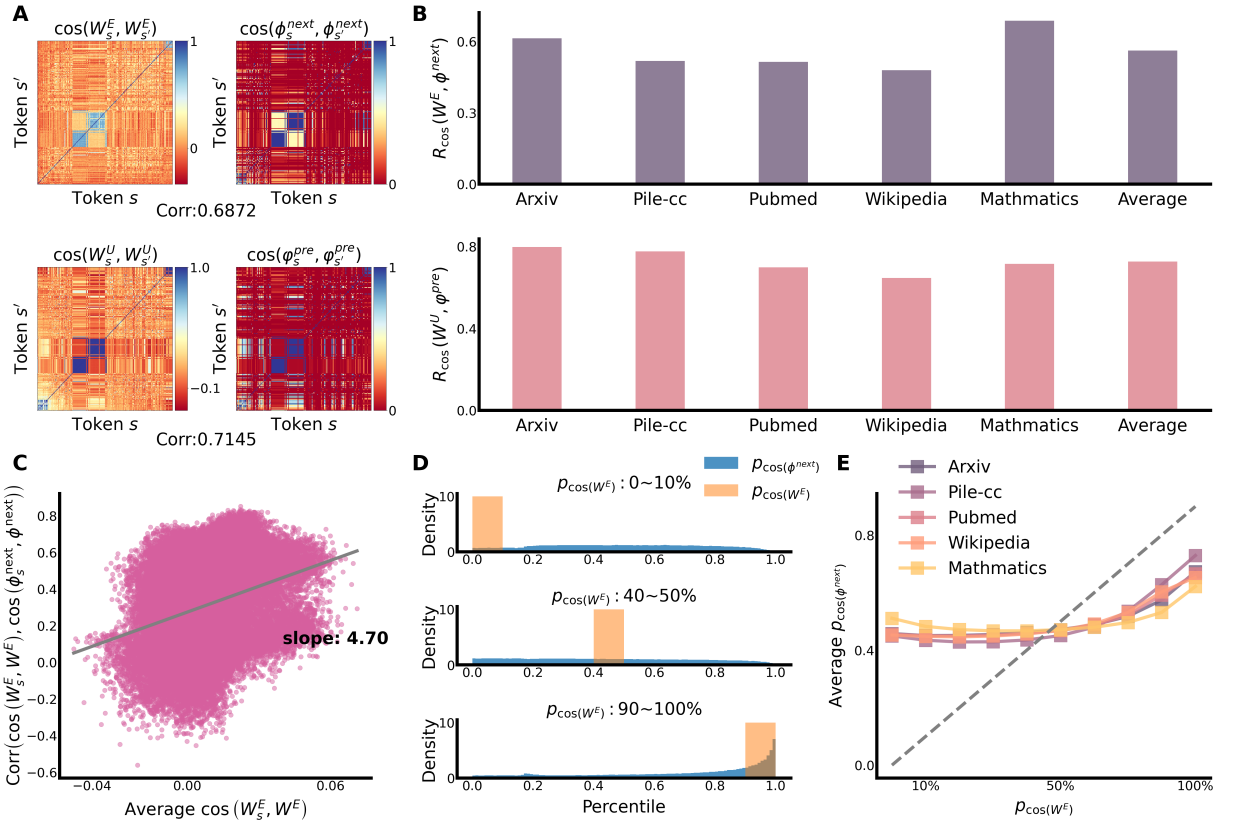



Figure 5: A: Heatmap of the cosine similarity of $\boldsymbol{W}^E, \boldsymbol{W}^U, \boldsymbol{\phi}^{\text{next}}$ and $\boldsymbol{\varphi}^{\text{pre}}$. B: $R_{\cos}\left(\boldsymbol{W}^E, \boldsymbol{\phi}^{\text{next}}\right)$ (top) and $R_{\cos}\left(\boldsymbol{W}^U, \boldsymbol{\varphi}^{\text{pre}}\right)$ (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 $p_{\cos(\boldsymbol{W}^E)}$.

language model $F_{\text{lan}}$ as:

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

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

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

and $\boldsymbol{\phi}^{\text{next}} = \left[\boldsymbol{\phi}_s^{\text{next}}\right]_{s\in\mathcal{V}}, \boldsymbol{\varphi}^{\text{pre}} = \left[\boldsymbol{\phi}_s^{\text{pre}}\right]_{s\in\mathcal{V}}$. We derive the following result:

**Corollary 4.** *Let $N \to \infty$, $\pi$ 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.*

Corollary 4 suggests that given any token $s$, the distributions of its next token and previous token significantly impact its embedding. We trained a group of Qwen2.5 models on different subsets of the Pile. Figure 5 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 $R_{\cos}\left(\boldsymbol{W}^E, \boldsymbol{\phi}^{\text{next}}\right) := \text{Corr}\left(\cos\left(\boldsymbol{W}^E\right), \cos\left(\boldsymbol{\phi}^{\text{next}}\right)\right)$, and similarly $R_{\cos}\left(\boldsymbol{W}^U, \boldsymbol{\varphi}^{\text{pre}}\right)$. Figure 5 B depicts both metrics across all subsets, suggesting that probability signatures significantly impact the structure of the embedding space and reflect the relationships among embeddings. Furthermore, we find that the probability signatures reflect the strong connections of embeddings more faithfully. As shown in Figure 5 C, the correlation 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 $\cos\left(\boldsymbol{W}_s^E, \boldsymbol{W}^E\right)$ 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 5 D displays the distribution of $p_{\cos(\boldsymbol{\phi}^{\text{next}})}$, conditioned on different intervals of the $p_{\cos(\boldsymbol{W}^E)}$, and Figure 5 E 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. In Appendix C, we provide a detailed method explanation, a specific case of the token group with large similarity, and an analysis with the Llama-2 architecture to validate the generalization of our analysis.

Since general-purpose pretrained base models are trained on broad corpora, we attempt to directly estimate their embedding structure with a subset of general text. We combine all datasets employed in Figure 5 and define $\tilde{\boldsymbol{\phi}} = \boldsymbol{\phi}^{\text{next}} + \boldsymbol{\varphi}^{\text{pre}}$ (Since the tied embedding, the detail is provided in Appendix C.1). We compare the $\cos\left(\tilde{\boldsymbol{\phi}}\right)$ with $\cos\left(\boldsymbol{W}^E\right)$ of Qwen2.5-3B-base. As shown in Figure 6 A, the structure of $\tilde{\boldsymbol{\phi}}$ could capture the main properties of the pre-trained model's embedding structure, particularly the presence of sub-blocks with high similarity. Furthermore, we examine the instance for the digits ranging from 0 to 9. Figure 6 B illustrates the $\cos\left(\boldsymbol{W}^E\right)$ and $\cos\left(\tilde{\boldsymbol{\phi}}\right)$ of such digits, both revealing an ordered organization that aligns with their numerical sequence. It should be noted that this estimation may not generalize across all open-source base models, as it is sensitive to both the initialization of the pre-trained model and the true training dataset.

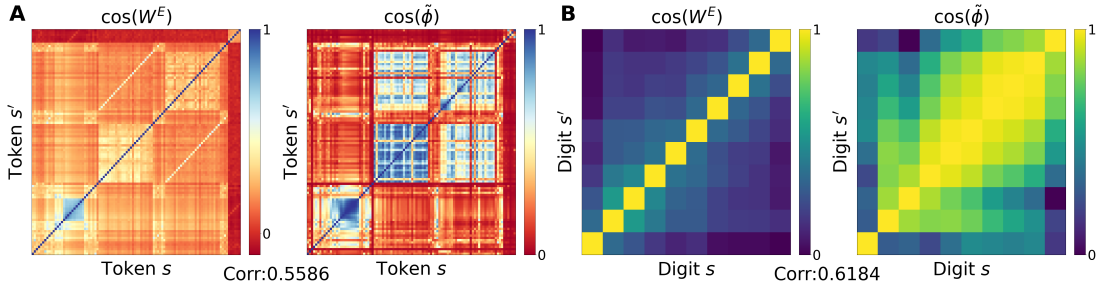



Figure 6: Cosine similarity of $\boldsymbol{W}^E$ of the Qwen2.5-3B-base and $\tilde{\boldsymbol{\phi}}$, respectively, with the frequently-appearing tokens (A) and the digits from 0 to 9 (B).

## 7 Conclusion

In this work, we investigate the formation of embedding structures in language models. By interpreting the relationship between embedding organization and semantic structure through the lens of data distribution, we propose the probability signatures and design the addition tasks to conduct variable-controlled experiments. Our findings demonstrate that probability signatures play a crucial role in shaping the embedding structure and reflecting underlying semantic relationships. An extended analysis of LLMs further confirms our analysis. This study establishes a bridge between data semantics and embedding space, offering new insights into the understanding of the joint impact of model, data and optimization method. For future work, we plan to extend our theoretical analysis into a comprehensive framework. Besides, we aim to incorporate the self-attention mechanism into our analysis of the LLMs, which is essential for capturing more subtle and complex relationships among embeddings that remain beyond the reach of our current methods.

## 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, Cal-

lum 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/.

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/.

Daniel L Ruderman. The statistics of natural images. *Network: computation in neural systems*, 5(4):517, 1994.

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.

David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

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.

# A Experimental Setups

**Addition tasks** For each type of addition task, we trained a linear model $F_{\text{lin}}$ and a Feedforward network $F_{\text{ffn}}$. 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 $10^{-5}$.

**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 $d_{\text{in}}$ 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 $U(\mathcal{A})$ and $U(\mathcal{Z})$ as the discrete uniform distribution over $\mathcal{A}$ and $\mathcal{Z}$, respectively. $A$ and $Z$ are the random variables following $U(\mathcal{A})$ and $U(\mathcal{Z})$. For the task $f_{\text{add}}$, 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 $\alpha$ or $\nu$. Figure 7 (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 $Y \sim U(\mathcal{Y})$ 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 $f_{\text{mod}}$, 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 7 depicts all these probability distributions.

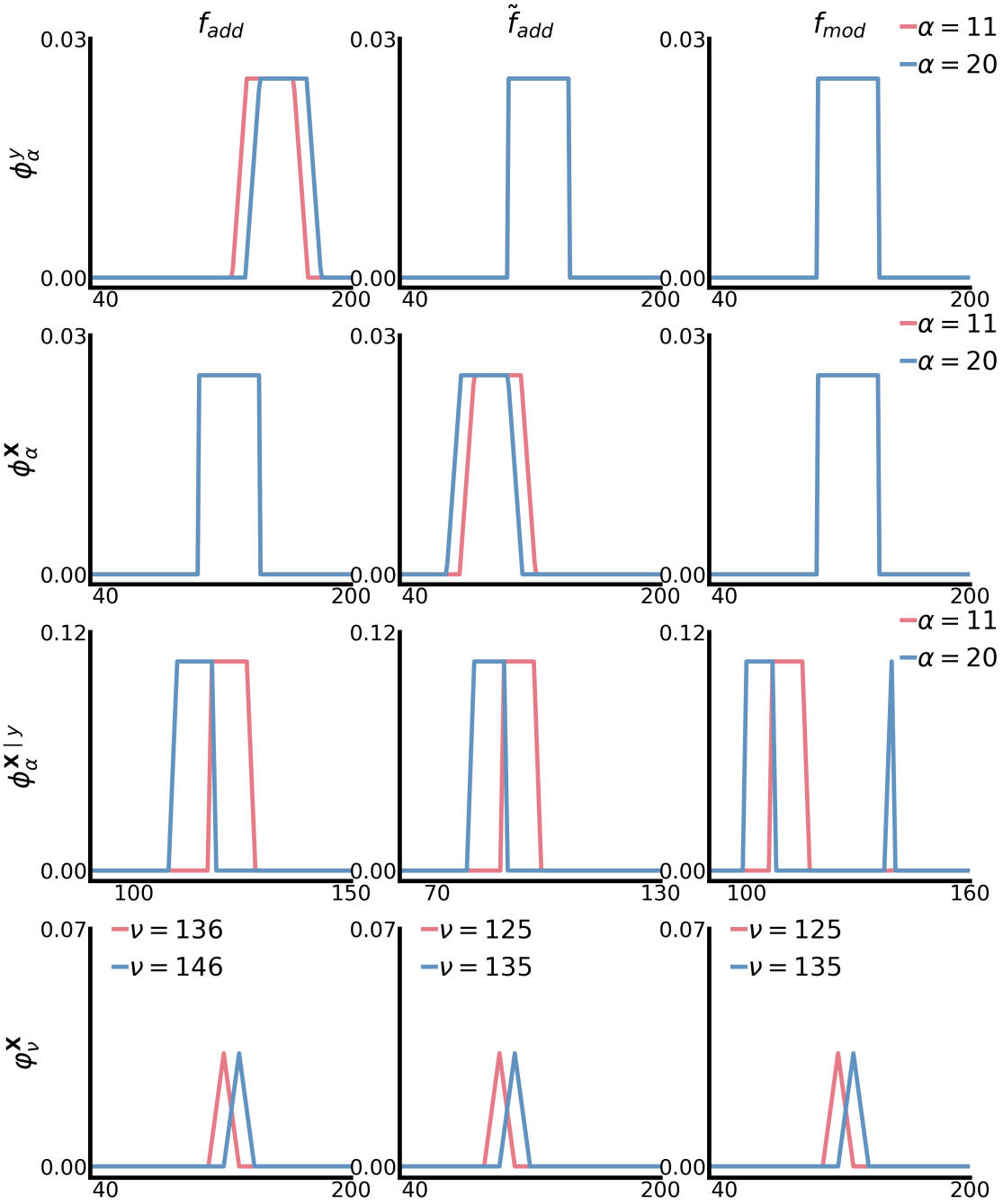



Figure 7: Probability signatures in each task under distinct $\alpha$ and $\nu$. In the distribution of $\phi_\alpha^{\mathbf{X}|y}$, $y = 150$ is displayed in $f_{\text{add}}$ and $y = 120$ in $\tilde{f}_{\text{add}}$ and $f_{\text{mod}}$, since 150 and 120 are the average label value in each task.

### B.2 Embedding matrix in Linear Model

Figure 8 depicts the PCA projection of the anchor embeddings in $F_{\text{lin}}$, revealing that $f_{\text{add}}$ and $\tilde{f}_{\text{add}}$ both establish an ordered structure while the anchor embeddings in $f_{\text{mod}}$ are chaotic.

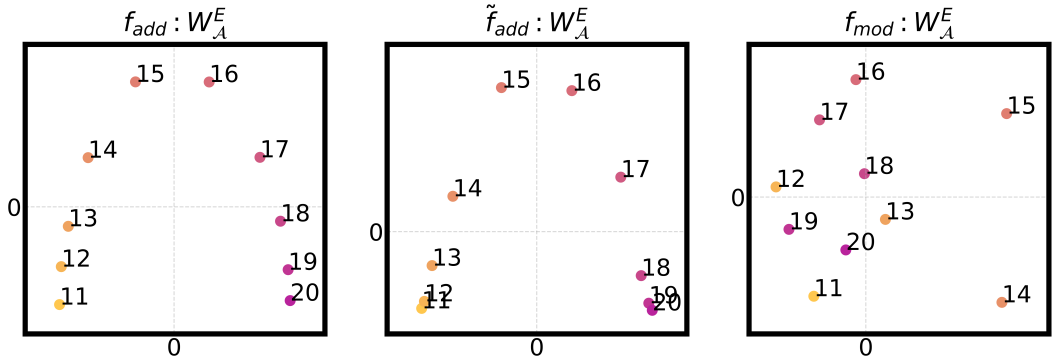



Figure 8: PCA projection of $\boldsymbol{W}_{\mathcal{A}}^{E}$ in $F_{\text{lin}}$ (epoch 120).

### B.3 Umembedding matrix in Feedforward Network

Figure 9 displays the structure of the unembedding matrix in $F_{\text{ffn}}$ 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.

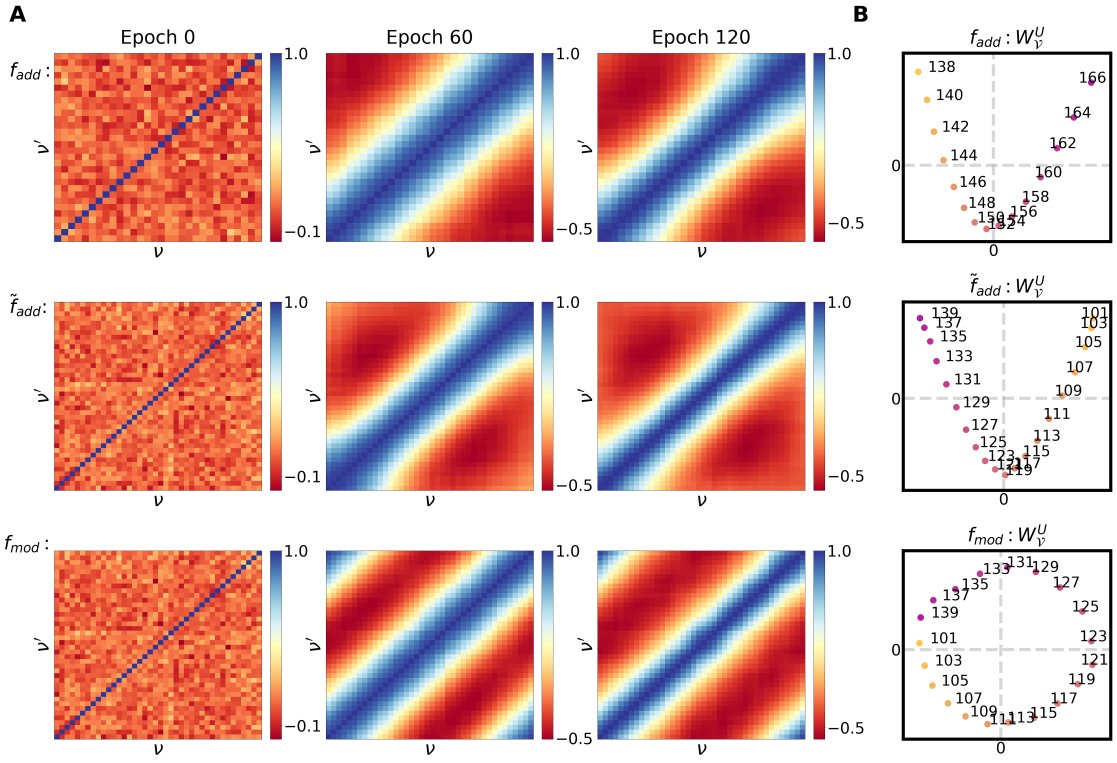



Figure 9: A: The heatmap of the $\cos\left(\boldsymbol{W}_{\mathcal{V}}^{U}\right)$ with label index in $F_{\text{ffn}}$ during the training process. B: PCA projection of $\boldsymbol{W}_{\mathcal{V}}^{U}$ in $F_{\text{ffn}}$ (epoch 120).

# C Language Models

## C.1 Technical Details

**Remark about Figure 5 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 $D_i$. 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

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

and the average embedding similarity as

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

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.

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

Figure 5 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 5 D & E** 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 5 D and E reveal the distribution and average value of $p_{\cos(\boldsymbol{\phi}^{\text{next}})}$, where $k\times 10\% \le p_{\cos(\boldsymbol{W}^E)} < (k+1)\times 10\%, k = 0, 1, 2, \cdots, 9$.

**Tied Embedding** In the Qwen2.5-3B-base model, the embedding matrix and unembedding matrix are the same one, which aims for computational source saving. Under this condition, we have that

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

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

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

### C.2 Complete results

Figure 10 represents the cosine similarity distribution of $\boldsymbol{W}^E, \phi^{\text{next}}, \boldsymbol{W}^U$ and $\varphi^{\text{pre}}$ in the other 4 subsets of Pile we selected, exhibiting an analogous phenomenon with the observation in Figure 5. The distribution representations $\phi^{\text{next}}$ and $\varphi^{\text{pre}}$ could effectively capture the high similarity among embedding vectors and unembedding vectors. Figure 11 displays the completed result of Figure 5 D.

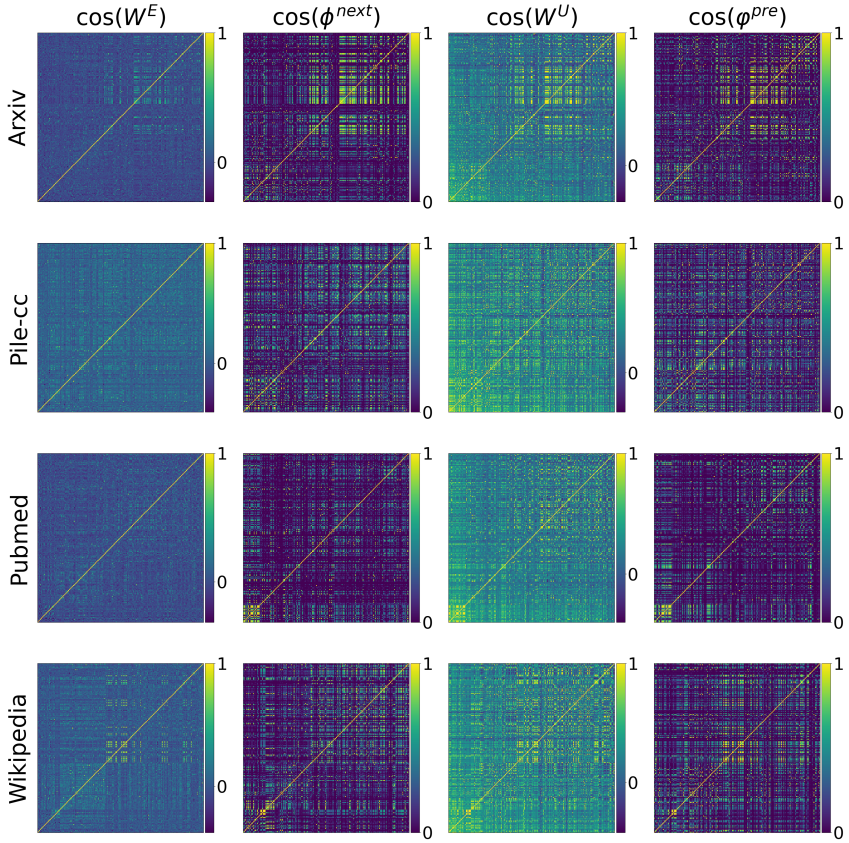



Figure 10: Cosine similarity distribution of $\boldsymbol{W}^E, \phi^{\text{next}}, \boldsymbol{W}^U, \varphi^{\text{pre}}$ in each experiment with distinct dataset. The tokens displayed are those with the most appearances in the dataset.

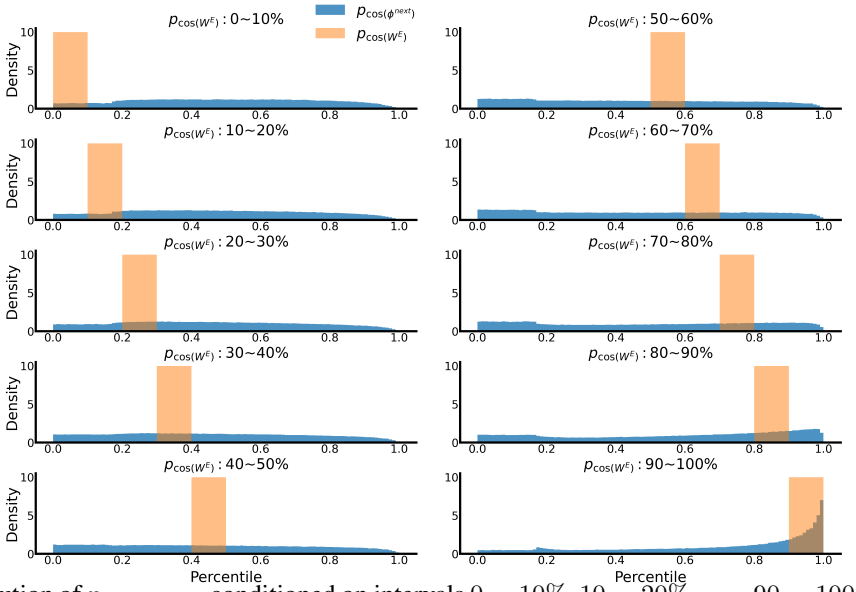
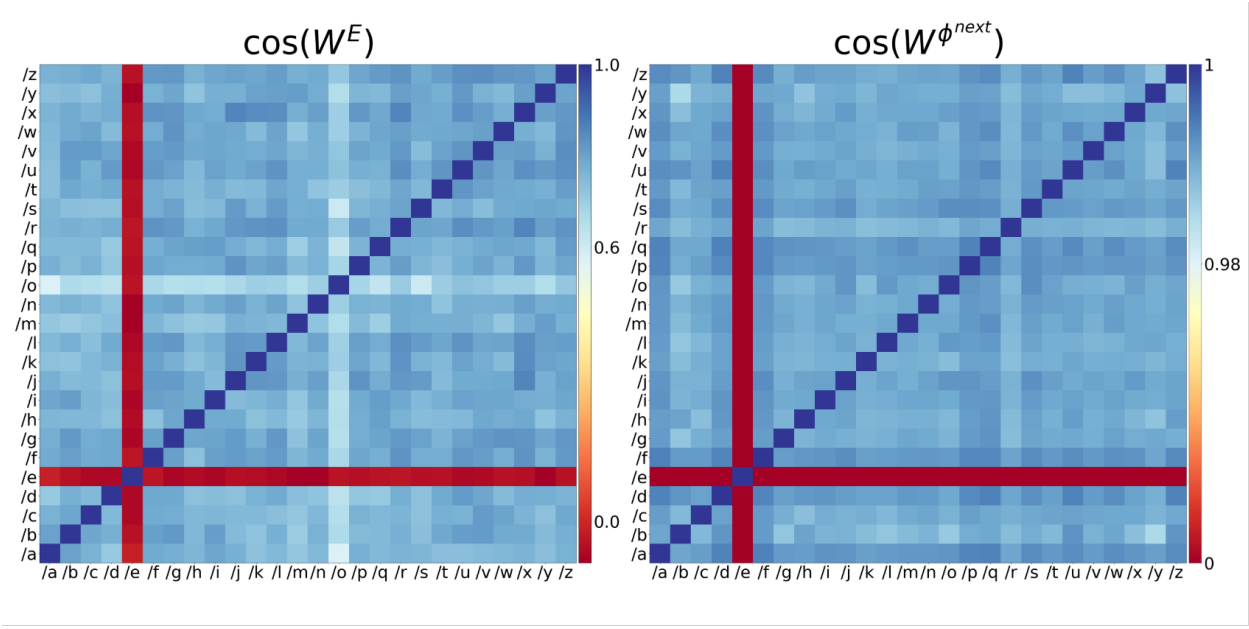



Figure 11: Distribution of $p_{\cos(\phi^{\text{next}})}$, conditioned on intervals $0 \sim 10\%, 10 \sim 20\%, \cdots, 90 \sim 100\%$ of the $p_{\cos(\boldsymbol{W}^E)}$.

### C.3 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 12 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**(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 12: 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 13, 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 5, indicating that our analytical approach is robust to variations in model architecture.

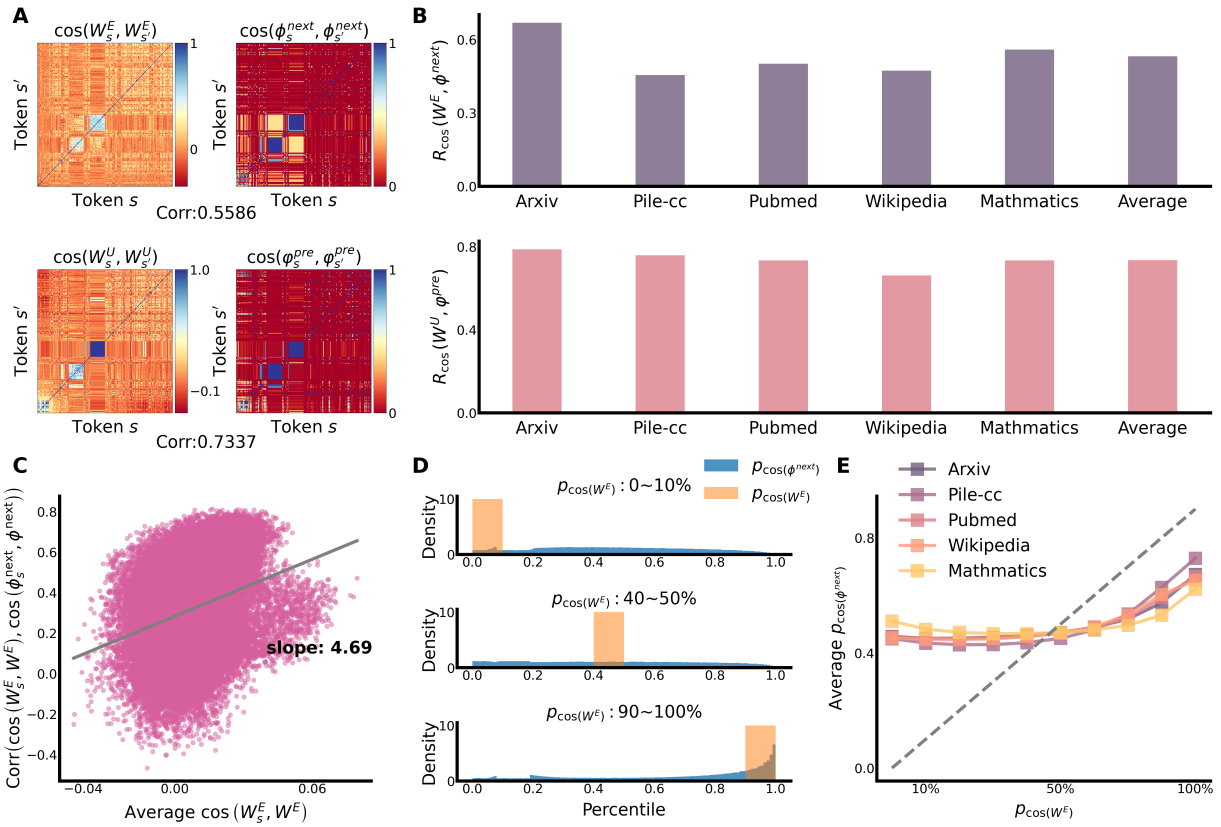



Figure 13: Results with Llama-2 architecture. A: Heatmap of the cosine similarity of $\boldsymbol{W}^E, \boldsymbol{W}^U, \boldsymbol{\phi}^{\text{next}}$ and $\boldsymbol{\varphi}^{\text{pre}}$. B: $R_{\cos}\left(\boldsymbol{W}^E, \boldsymbol{\phi}^{\text{next}}\right)$ (top) and $R_{\cos}\left(\boldsymbol{W}^U, \boldsymbol{\varphi}^{\text{pre}}\right)$ (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 $(\boldsymbol{X}, y) \in \mathbb{N}^{+,L} \times \mathbb{N}^{+}$, $\ell = -\log Softmax\left(F\left(\boldsymbol{X}\right)\right)_y$, we have that*

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

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

*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 $\ell$ 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 $vW_x^E$ 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

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

## 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

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

### D.3 Proof of Corollary 1

With proposition 1, we have that

$$\begin{aligned}\frac{d\boldsymbol{W}_\alpha^E}{dt} &= \boldsymbol{W}^{U,T}\left(\sum_{\nu\in\mathcal{V}} r_{\alpha,\nu}\boldsymbol{e}_\nu - \frac{r_\alpha^{\text{in}}}{N_\alpha^{\text{in}}}\sum_{i=1}^{N_x^{\text{in}}}\boldsymbol{p}_\alpha^i\right)\\ &= \boldsymbol{W}^{U,T} r_\alpha^{\text{in}}\left(\sum_{\nu\in\mathcal{V}} \frac{r_{\alpha,\nu}}{r_\alpha^{\text{in}}}\boldsymbol{e}_\nu - \frac{1}{N_\alpha^{\text{in}}}\sum_{i=1}^{N_x^{\text{in}}}\boldsymbol{p}_\alpha^i\right).\end{aligned}$$

Utilizing that $\text{softmax}(\boldsymbol{f}) = \frac{1}{d_{\text{vob}}}\mathbf{1} + \frac{1}{d_{\text{vob}}}\boldsymbol{f} + \mathcal{O}\left(d_{\text{vob}}^{-2}\boldsymbol{f}\right)$, we obtain that

$$\begin{aligned}\frac{d\boldsymbol{W}_\alpha^E}{dt} &= \boldsymbol{W}^{U,T} r_\alpha^{\text{in}}\left(\sum_{\nu\in\mathcal{V}} \frac{r_{\alpha,\nu}}{r_\alpha^{\text{in}}}\boldsymbol{e}_\nu - \frac{1}{N_\alpha^{\text{in}}}\sum_{i=1}^{N_\alpha^{\text{in}}}\left(\frac{1}{d_{\text{vob}}}\mathbf{1} - \frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\left(\boldsymbol{W}_{z_i}^E + \boldsymbol{W}_{\alpha_i}^E + \boldsymbol{W}_\alpha^E\right) + \mathcal{O}\left(d_{\text{vob}}^{-2}\boldsymbol{W}^U\boldsymbol{W}_\alpha^E\right)\right)\right)\\ &= \boldsymbol{W}^{U,T} r_\alpha^{\text{in}}\left(\sum_{\nu\in\mathcal{V}} \frac{r_{\alpha,\nu}}{r_\alpha^{\text{in}}}\boldsymbol{e}_\nu - \frac{1}{d_{\text{vob}}}\mathbf{1} + \frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\left(\frac{1}{N_\alpha^{\text{in}}}\sum_{i=1}^{N_\alpha^{\text{in}}}\left(\boldsymbol{W}_{z_i}^E + \boldsymbol{W}_{\alpha_i}^E\right) + \boldsymbol{W}_\alpha^E\right) + \mathcal{O}\left(d_{\text{vob}}^{-2}\boldsymbol{W}^U\boldsymbol{W}_\alpha^E\right)\right)\\ &= \boldsymbol{W}^{U,T} r_\alpha^{\text{in}}\left(\sum_{\nu\in\mathcal{V}} \frac{r_{\alpha,\nu}}{r_\alpha^{\text{in}}}\boldsymbol{e}_\nu - \frac{1}{d_{\text{vob}}}\mathbf{1} + \frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\left(\sum_{x\in(\mathcal{Z}\cap\mathcal{A})}\frac{N_{\alpha,x}^{\text{in}}}{N_\alpha^{\text{in}}}\boldsymbol{W}_x^E + \boldsymbol{W}_\alpha^E\right) + \mathcal{O}\left(d_{\text{vob}}^{-2}\boldsymbol{W}^U\boldsymbol{W}_\alpha^E\right)\right).\end{aligned}$$

Let $N\to\infty$, we have that

$$\begin{aligned}\frac{d\boldsymbol{W}_\alpha^E}{dt} =& \boldsymbol{W}^{U,T} r_\alpha^{\text{in}}\left(\sum_{\nu\in\mathcal{V}}\mathbb{P}_\pi\left(y=\nu \mid \alpha\in\boldsymbol{X}\right)\boldsymbol{e}_\nu - \frac{1}{d_{\text{vob}}}\mathbf{1}\right.\\ &\left. + \frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\left(\sum_{x\in(\mathcal{Z}\cap\mathcal{A})}\mathbb{P}_\pi\left(x\in\boldsymbol{X}\mid\alpha\in\boldsymbol{X}\right)\boldsymbol{W}_x^E + \boldsymbol{W}_\alpha^E\right) + \mathcal{O}\left(d_{\text{vob}}^{-2}\boldsymbol{W}^U\boldsymbol{W}_\alpha^E\right)\right)\\ =& \boldsymbol{W}^{U,T} r_\alpha^{\text{in}}\left(\boldsymbol{\phi}_\alpha^y + \frac{1}{d_{\text{vob}}}\boldsymbol{W}^U\boldsymbol{W}^E\boldsymbol{\phi}_\alpha^{\boldsymbol{X}}\right) + \boldsymbol{\eta},\end{aligned}$$

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

$$\begin{aligned}\frac{d\boldsymbol{W}_\alpha^E}{dt} =& \underbrace{\sum_{\nu\in\mathcal{V}}\frac{r_{\alpha,\nu}}{N_{\alpha,\nu}}\left(\boldsymbol{W}^{U,T}\boldsymbol{e}_\nu\right)\odot\sum_{i=1}^{N_{\alpha,\nu}}\left(\mathbf{1} + \sum_{x\in\boldsymbol{X}_{(\alpha,\nu)}^i}\boldsymbol{W}_{\boldsymbol{X}_{(\alpha,\nu)}^i}^E\right)}_{\boldsymbol{J}^y}\\ &\underbrace{- \frac{r_\alpha^{\text{in}}}{N_\alpha^{\text{in}}}\sum_{i=1}^{N_\alpha^{\text{in}}}\left(\mathbf{1} + \sum_{x\in\boldsymbol{X}_\alpha^i}\boldsymbol{W}_{\boldsymbol{X}_\alpha^i}^E\right)\odot\left(\boldsymbol{W}^{U,T}\boldsymbol{p}_\alpha^i\right)}_{\boldsymbol{J}^p}.\end{aligned}$$

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

$$\boldsymbol{J}^y = \boldsymbol{W}^{U,T} \sum_{\nu \in \mathcal{V}} r_{\alpha,\nu} \boldsymbol{e}_\nu + \sum_{\nu \in \mathcal{V}} r_{\alpha,\nu} \left( \boldsymbol{W}^{U,T} \boldsymbol{e}_\nu \right) \odot \left( \boldsymbol{W}^E_\alpha + \frac{1}{2N_{\alpha,\nu}} \sum_{i=1}^{2N_{\alpha,\nu}} \boldsymbol{W}^E_{x^i_{(\alpha,\nu)}} \right).$$

Let $N \to \infty$, we have that

$$\begin{aligned}
\boldsymbol{J}^y &= \boldsymbol{W}^{U,T} r^{\text{in}}_\alpha \boldsymbol{\phi}^y_\alpha \odot \left( \mathbf{1} + \boldsymbol{W}^E_\alpha \right) + \sum_{\nu \in \mathcal{V}} \text{diag}\left( \boldsymbol{W}^U_\nu \right) r_{\alpha,\nu} \sum_{x' \in \mathcal{V}} \mathbb{P}\left( x' \in \boldsymbol{X} \mid \alpha \in \boldsymbol{X}, y = \nu \right) \boldsymbol{W}^E_{x'} \\
&= \boldsymbol{W}^{U,T} r^{\text{in}}_\alpha \boldsymbol{\phi}^y_\alpha \odot \left( \mathbf{1} + \boldsymbol{W}^E_\alpha \right) + \sum_{\nu \in \mathcal{V}} \text{diag}\left( r_{\alpha,\nu} \boldsymbol{W}^U_\nu \right) \boldsymbol{W}^E \left( \boldsymbol{\phi}^{\boldsymbol{X}|y,T}_\alpha \right)^T_\nu \\
&= \boldsymbol{W}^{U,T} r^{\text{in}}_\alpha \boldsymbol{\phi}^y_\alpha \odot \left( \mathbf{1} + \boldsymbol{W}^E_\alpha \right) + \mathbb{T} \cdot \left( \boldsymbol{\phi}^{\boldsymbol{X}|y}_\alpha \right)^T,
\end{aligned}$$

where $\mathbb{T} \in \mathbb{R}^{d \times d_{\text{vob}} \times d_{\text{vob}}}$, $\mathbb{T}_{:,:,\nu} = r_{\alpha,\nu} \text{diag}\left( \boldsymbol{W}^U_\nu \right) \boldsymbol{W}^E$ for $\nu \in \mathcal{V}$ and 0 otherwise.
Similarly, for the term $\boldsymbol{J}^p$, we have that

$$\boldsymbol{J}^p = \boldsymbol{W}^{U,T} r^{\text{in}}_\alpha \left( \frac{1}{d_{\text{vob}}} \mathbf{1} - \frac{1}{d_{\text{vob}}} \boldsymbol{W}^U \left( \left( \boldsymbol{W}^E - \text{diag}\left( \boldsymbol{W}^{U,T} \mathbf{1} \right) \right) \boldsymbol{\phi}^{\boldsymbol{X}}_\alpha + \boldsymbol{W}^E_\alpha \right) + \boldsymbol{\epsilon} \right),$$

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

$$\frac{d\boldsymbol{W}^E_\alpha}{dt} = \mathbb{T} \cdot \left( \boldsymbol{\phi}^{\boldsymbol{X}|y}_\alpha \right)^T + \boldsymbol{\eta}_{\boldsymbol{\phi}^y_\alpha} + \frac{1}{d_{\text{vob}}} \boldsymbol{\eta}_{\boldsymbol{\phi}^{\boldsymbol{X}}_\alpha},$$

where $\boldsymbol{\eta}_{\boldsymbol{\phi}^y_\alpha} = \boldsymbol{W}^{U,T} r^{\text{in}}_\alpha \boldsymbol{\phi}^y_\alpha \odot \left( \mathbf{1} + \boldsymbol{W}^E_\alpha \right)$, $\boldsymbol{\eta}_{\boldsymbol{\phi}^{\boldsymbol{X}}_\alpha} = d_{\text{vob}} \boldsymbol{J}^p$. □

### D.5 Proof of Corollary 3

*Proof.* With Proposition 2, we have that

$$\begin{aligned}
\frac{d\boldsymbol{W}^U_\nu}{dt} =& \frac{r^{\text{out}}_\nu}{N^{\text{out}}_\nu} \sum_{i=1}^{N^{\text{out}}_\nu} \left( \sum_{x \in \boldsymbol{X}^i_{(\cdot,\nu)}} \boldsymbol{W}^E_x \right)^T - \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{p}^{i,\nu} \left( \sum_{x \in \boldsymbol{X}^i} \boldsymbol{W}^E_x \right)^T \\
=& L r^{\text{out}}_\nu \sum_{x \in \mathcal{V}} \mathbb{P}_\pi \left( x \in \boldsymbol{X} \mid y = \nu \right) \boldsymbol{W}^{E,T}_x - L \sum_{x \in \mathcal{V}} \mathbb{E}_\pi \left[ \boldsymbol{p}^\nu \mid x \in \boldsymbol{X} \right] \boldsymbol{W}^{E,T}_x \\
=& L r^{\text{out}}_\nu \left( \boldsymbol{W}^E \boldsymbol{\varphi}^{\boldsymbol{X}}_\nu \right)^T - \boldsymbol{\eta},
\end{aligned}$$

where $\boldsymbol{\eta} = L \left( \boldsymbol{W}^E \mathbb{E}_\pi \left[ \boldsymbol{p} \mid x \in \boldsymbol{X} \right] \right)^T$. □

### 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

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

Furthermore, we have that

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

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