Structural Fabric of Data Shapes the Geometry of LLM Training

Anonymous ACL submission

Abstract

The performance of large language models is closely tied to the quality and organization of their training data. However, existing data metrics largely emphasize surface-level properties, such as sequence length or model-based quality scores, and provide limited insight into how the internal organization of data relates to optimization. In this paper, we study the structural fabric of training data, defined as the pattern of semantic dependencies within a sequence. We quantify this property using dependency density, a normalized measure of effective segment-level dependencies. Across data length, domain, and format, we find that dependency density varies systematically with data configuration: code shows higher dependency density than natural language within long-context data, and instruction-tuning mappings show higher dependency density than raw intra-input modeling. We further show that coherent structural fabric is associated with a lower gradient effective rank, indicating a more concentrated optimization geometry with reduced gradient interference. Finally, we observe that this low-rank optimization regime is associated with a more stable intrinsic dimension, preserving a compact representation manifold that may mitigate representation noise and support downstream performance. These results provide a geometric perspective on how data structure relates to LLM training, linking data configuration, optimization dynamics, and model behavior through the lens of structural fabric.

1 Introduction

Large language models (LLMs) have demonstrated exceptional performance across a wide range of real-world tasks, including mathematical reasoning, dialogue, and code generation (Achiam et al., 2023; Guo et al., 2025). While it is established that model performance is highly driven by the quality and composition of training data (Zhou et al., 2023), the link between data properties and training dynamics remains under-explored. Prevailing perspectives often characterize data through surface-level metrics such as sequence length (Xia et al., 2025) or model-based scores (Li et al., 2024), overlooking the intrinsic structural fabric of the data. These limitations motivate a structural view of data that connects intra-sequence dependency organization to the geometry of training.

In this paper, we ask: *through what geometric mechanism does the intrinsic organization of data influence the final performance of LLMs?* To answer this, we establish an analytical framework bridging the **structural fabric of data**, **optimization geometry**, and **model behavior**. We quantify structural fabric with **dependency density**, which measures the average concentration of effective dependencies after normalizing away length-induced accumulation. Our investigation shows that coherent structural fabric fosters a **low-rank gradient manifold**, which acts as a geometric regularizer that stabilizes the model's **intrinsic dimension**. This chain of influence helps compress information into a coherent manifold, thereby improving performance on complex downstream tasks.

We analyze structural fabric across three dimensions: **length, domain, and format**. The resulting profile reveals that density is configuration-dependent rather than a monotonic function of length: code is denser than natural language within long-context data, and instruction-tuning mappings are substantially denser than raw intra-input modeling. This distinction matters because two samples with similar surface quality can provide very different training signals: one may expose explicit causal links across segments, whereas another may contain fluent but weakly coupled text.

We next examine the resulting **optimization geometry**. More coherent structural fabric is reflected in a more concentrated gradient spectrum and a lower **gradient effective rank**. This low-rank regime indicates that optimization is driven by coherent signals while suppressing gradient conflicts from semantic ambiguity or input noise. Finally, to characterize the resulting **model behavior**,

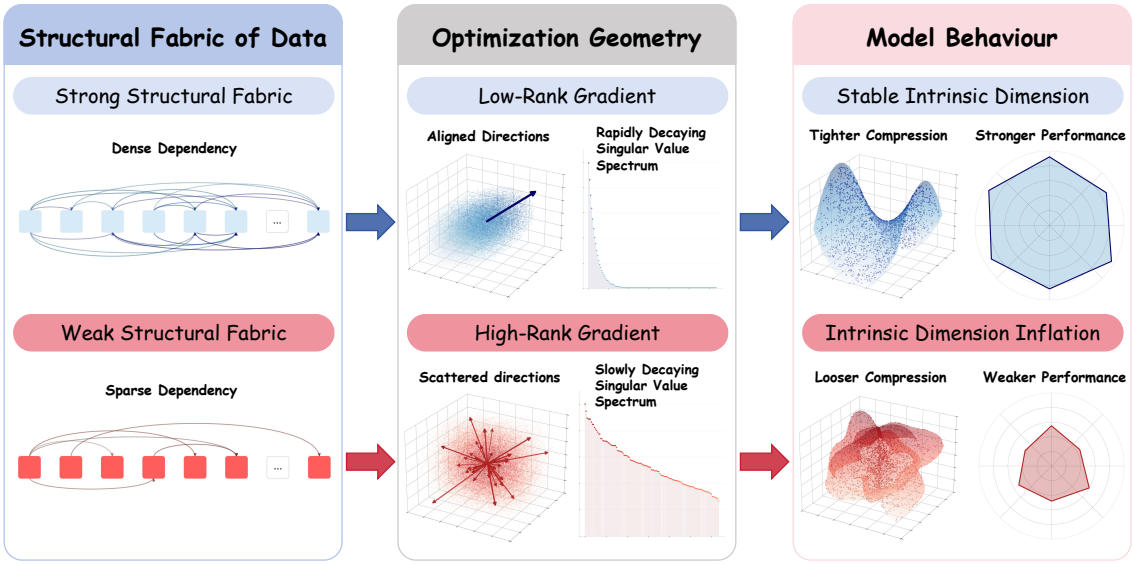


Figure 1: **Our analytical framework bridging data structural fabric, optimization geometry, and model behavior.** The structural fabric of training data (left) shapes the optimization geometry (middle), where strong dependency structure is associated with low-rank, coherent gradient updates. This process may act as a geometric regularizer that helps stabilize the model's intrinsic dimension (right), reducing representation blurring and supporting performance across downstream tasks.

we monitor the evolution of the model's **intrinsic dimension**. Data with coherent dependencies stabilizes this dimension throughout training, whereas less coherent update patterns lead to dimensional inflation, a sign of representation noise that degrades downstream performance.

In summary, our contributions are as follows:

- We establish an analytical framework that bridges the **structural fabric** of data, **optimization geometry**, and **model behavior**. This framework offers a geometric perspective for explaining how intrinsic data dependencies shape the training trajectory, moving beyond surface-level statistics.
- We adopt dependency density to quantitatively characterize training data. We show that structural fabric varies systematically with data length, domain, and format.
- We demonstrate that coherent structural fabric induces a lower gradient effective rank, implying reduced gradient interference during training. This produces a more stable intrinsic dimension and improves performance across benchmarks.

2 Related Work

2.1 Data Quality Assessment

Data quality is increasingly recognized as a more critical factor than scale in driving LLM performance. Early approaches relied on **heuristic rules**, such as regex-based cleaning (Rae et al., 2021) and perplexity thresholds (Wenzek et al., 2020), which have limitations in capturing logical coherence. Recent paradigms have shifted towards model-driven metrics. Winata et al. (2021) employs trainable classifiers to score documents based on similarity to high-quality references, while Lu et al. (2024) leverages ChatGPT to assign fine-grained intention tags, enabling direct quantification of data diversity and complexity.

However, these model-based metrics primarily focus on **semantic content** or **inter-document diversity**, leaving **intra-document structure** largely unexamined. While Chen et al. (2024) proposed an explicit formula to quantify long-range dependencies, they did not systematically evaluate structural disparities across diverse data configurations. Our work fills this gap by treating structural fabric as a complementary view of data quality: rather than asking whether a document is high quality in isolation, we ask how its internal dependency structure shapes training geometry.

2.2 Training Dynamics of LLMs

The training of deep neural networks is often characterized as an optimization process in a high-dimensional parameter space, yet extensive evidence suggests that effective learning occurs in a much lower-dimensional geometry. Prior work analyzes this from three perspectives: loss curvature,

low-rank gradient structure, and intrinsic dimension.

Hessian spectral analysis reveals that optimization trajectories typically evolve on low-dimensional manifolds (Fernandez et al., 2025). This is complemented by the strong low-rank bias observed in gradients, even in overparameterized models (Huh et al., 2021). Theoretical results further relate gradient rank to the eigenvalue spectrum of the input data covariance matrix (Sonthalia et al., 2025). Meanwhile, research on intrinsic dimension explains why models can be effectively optimized within low-dimensional subspaces (Aghajanyan et al., 2021), providing a theoretical basis for low-rank adaptation methods like LoRA (Hu et al., 2022).

Recent research has begun to explore how specific data characteristics influence optimization. Li et al. (2025) employs layer-wise gradient analysis, demonstrating that high-quality instruction and reasoning data consistently exhibit higher effective ranks driven by semantic diversity. Our work is complementary rather than contradictory: semantic diversity can broaden the span of useful updates, whereas structural coherence can concentrate updates around causal dependencies. We therefore focus on a different source of gradient geometry, showing that coherent structural fabric suppresses interference and yields a lower gradient effective rank in our experimental settings.

3 Analytical Framework

As illustrated in Figure 1, we establish a unified analytical framework connecting intrinsic data organization to downstream model behavior. We introduce **dependency density** to quantitatively capture the data's structural fabric (Section 4). We then investigate how this property shapes training dynamics by measuring the **gradient effective rank** (Section 5). Finally, we use the **intrinsic dimension** as a measure of geometric robustness, showing that a stable, compressed manifold is associated with reduced representation noise and stronger downstream performance (Section 6). The framework separates three related but distinct questions: how dependencies are distributed in data, how those dependencies concentrate gradient updates, and whether the resulting representations remain geometrically compact during training.

3.1 Quantifying Structural Fabric: Dependency Density

Following Chen et al. (2024), we adopt the Long Dependency Score (LDS) framework to quantify structural dependencies within training data. For a sequence equally divided into N segments $\{c_1, \ldots, c_N\}$, the dependency between a context segment c_j and a target segment c_i ($j < i$) is evaluated through three factors:

- **Dependency Strength ($\mathbf{DST}_{i,j}$):** Measures semantic correlation via relative perplexity (PPL) reduction:

$$\mathrm{DST}_{i,j} = \frac{\mathrm{PPL}(c_i) - \mathrm{PPL}(c_i|c_j)}{\mathrm{PPL}(c_i)}. \quad (1)$$

- **Dependency Distance ($\mathbf{DDI}_{i,j}$):** Represents the normalized positional gap as $\mathrm{DDI}_{i,j} = \frac{i-j}{N-1}$.
- **Dependency Specificity ($\mathbf{DSP}_i$):** An entropy-based term that penalizes dependencies from repetitive contexts to prioritize genuine semantic connections.

The comprehensive dependency score for a segment pair is then defined as:

$$\mathrm{LDS}_{i,j} = (\alpha \mathrm{DST}_{i,j} + \beta \mathrm{DDI}_{i,j}) \cdot \mathrm{DSP}_i, \quad (2)$$

where α and β are hyperparameters.

The original evaluation of LDS primarily relies on cumulative scores that sum all $\mathrm{LDS}_{i,j}$ pairs exceeding a threshold τ. However, such cumulative metrics naturally scale with sequence length, creating a bias where long-context corpora appear to have a stronger structural fabric simply due to their volume. To address this, we introduce a normalized density metric. This normalization is essential for our study because we compare datasets whose average contexts, domains, and supervision formats differ substantially. Without it, length-induced accumulation would obscure whether dependencies are truly concentrated or merely numerous.

Definition 3.1 (Dependency Density)**.** To isolate the intrinsic structural density from length-induced bias, we define **dependency density** (ρ) as the average LDS over the set of effective segment pairs $\mathcal{P}$:

$$\rho = \frac{1}{|\mathcal{P}|} \sum_{(j,i) \in \mathcal{P}} \mathrm{LDS}_{i,j}, \quad (3)$$

where $\mathcal{P} = \{(j,i) \mid j < i \text{ and } \mathrm{DST}_{i,j} > \tau\}$ represents pairs exceeding a dependency threshold τ. This normalized metric enables a consistent comparison of structural fabric across varying sequence lengths and data configurations.

3.2 Measuring Optimization Geometry: Gradient Effective Rank

To investigate how differences in data structural fabric shape the optimization geometry, we utilize the concept of effective rank to measure the spectral concentration of gradient updates. Specifically, we analyze the gradients of projection matrices within the Attention modules, $W \in \{W_Q, W_K, W_V, W_O\}$, and the Feed-Forward Network modules, $W \in \{W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}\}$. For a weight matrix W in the ℓ-th Transformer layer, let $G_{W,\ell} \in \mathbb{R}^{m \times n}$ denote its gradient matrix. We perform Singular Value Decomposition (SVD) such that $G_{W,\ell} = U\Sigma V^T$, where Σ contains the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$.

Definition 3.2 (Gradient Effective Rank)**.** The **gradient effective rank** $\mathcal{R}_{W,\ell}$ is defined as the exponential of the spectral entropy (Roy and Vetterli, 2007):

$$\mathcal{R}_{W,\ell} = \exp\left(-\sum_{k=1}^{\min(m,n)} \tilde{\sigma}_k \ln(\tilde{\sigma}_k)\right), \quad (4)$$

where $\tilde{\sigma}_k = \sigma_k / \sum_j \sigma_j$ represents the normalized singular values of the gradient matrix $G_{W,\ell}$.

A higher effective rank indicates that gradients are distributed isotropically across the parameter space, often suggesting uninformative updates or noise. In contrast, a lower effective rank suggests that the gradient mass is concentrated in a few dominant principal directions. We interpret this reduction as evidence of structured optimization, where updates are prioritized along directionally coherent paths aligned with the underlying structural fabric of the training data. Importantly, low rank is not treated as universally beneficial in isolation; it is meaningful here because it emerges together with explicit dependency structure, indicating compression around useful constraints rather than a collapse of learning signal.

3.3 Assessing Model Behavior: Intrinsic Dimension

To establish a final link between optimization geometry and model intelligence, we utilize the **intrinsic dimension** to investigate how the low-rank gradient regime manifests in the model's internal representation space.

Definition 3.3 (Intrinsic Dimension)**.** Following Ruppik et al. (2025), we employ the **TwoNN (Two-Nearest Neighbor)** estimator to calculate the **intrinsic dimension**, denoted as $\hat{d}$:

$$\hat{d} = \frac{N}{\sum_{i=1}^{N} \ln(\mu_i)}, \quad (5)$$

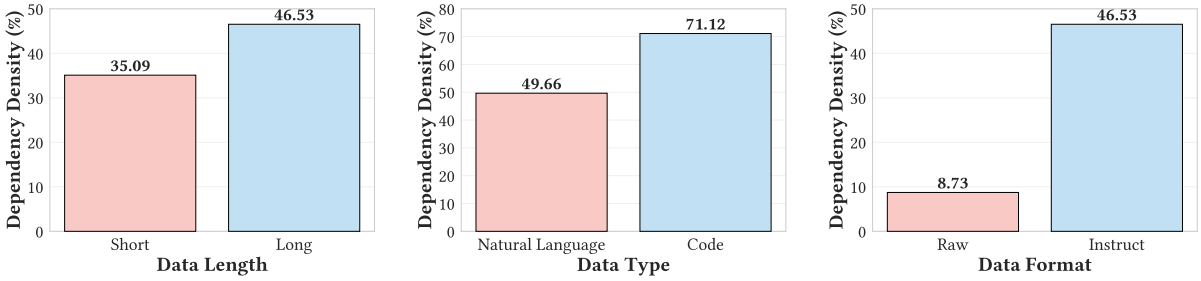
where, for a set of embedding vectors $\mathcal{X} = \{x_1, \ldots, x_N\}$, $\mu_i = r_{i,2}/r_{i,1}$ represents the ratio of the Euclidean distances from x_i to its second and first nearest neighbors, respectively.

While the hidden representations of LLMs reside in a high-dimensional ambient space, the actual data distribution typically lies on a much lower-dimensional manifold. This intrinsic structure reflects the complexity and compression of learned features, serving as a diagnostic tool for model health and representation quality. By tracking intrinsic dimension throughout the training process, we evaluate how the structural fabric of data and the resulting optimization dynamics shape the geometric stability of the embedding space. A lower intrinsic dimension indicates that the model has successfully compressed information into a structured manifold, whereas a rapidly rising dimension suggests *dimensional inflation*—a failure to maintain a compressed manifold that leads to the memorization of noise and degraded downstream performance.

4 Data Configuration Determines Dependency Structure

In this section, we utilize **dependency density** (ρ) to analyze the structural fabric of training data across three dimensions: length, domain, and format. We use two representative datasets: **UltraChat**, a short-context instruction-tuning dataset (Ding et al., 2023), and **ChatQA2**, a long-context dataset characterized by semantic coherence over extended spans (Xu et al., 2024).

By profiling data across three dimensions, we identify how data configurations distinctively shape the structural fabric. For **data length**, we sample sequences from ChatQA2 (long-context) and UltraChat (short-context). For **data domain**, we employ an LLM-based classifier to partition ChatQA2 into code tasks and natural language tasks. For **data format**, we compare the intrinsic dependencies of input alone against input-output pairs within ChatQA2. We keep these comparisons separated because length, domain, and format affect different aspects of dependency organization: length changes the span over which evidence is distributed,



(a) Dependency density across lengths. (b) Dependency density across domains. (c) Dependency density across formats.

Figure 2: **Quantitative profiling of structural fabric across three dimensions.** Dependency density (ρ) varies systematically with data length, domain, and format: long-context data is denser than short-context data, code is denser than natural language, and instruction-tuning mappings are denser than raw intra-input modeling.

domain changes the rigidity of semantic constraints, and format changes whether dependencies are expressed as open-ended continuation or directed input-output mapping.

4.1 Length Effects: Local Coupling versus Distributed Context

The relationship between sequence length and structural fabric indicates that extended contexts expose a denser dependency pattern after normalization. As shown in Figure 2a, the long-context data (ChatQA2) reaches a higher aggregate dependency density ($\rho = 46.53\%$) than the short-context instruction data (UltraChat; $\rho = 35.09\%$). This pattern suggests that long-context sequences contain more effective cross-segment dependencies, because evidence and constraints are distributed across a broader span. Short instruction examples still provide useful local input-output coupling, but the figure shows that their normalized dependency density is lower than that of long-context data. We therefore interpret length as a factor that can strengthen structural fabric when the additional context introduces coherent dependencies rather than merely adding unrelated tokens.

4.2 Domain Effects: Rigid Syntax Strengthens Dependencies

Beyond length, the semantic nature of the data domain shapes the concentration of the structural fabric. As illustrated in Figure 2b, code reaches a dependency density of 71.12%, substantially higher than natural language (49.66%).

This high density is a direct consequence of the **rigid logical constraints** inherent to programming languages. Unlike natural language, which allows for fluid and redundant expression, code is governed by strict syntactical rules where variable definitions, function calls, and control structures necessitate precise dependencies. This structural rigidity prevents semantic dilution, ensuring that most segments maintain high-strength connections. The comparison is therefore not simply a content distinction; it reflects how strongly the domain constrains valid continuations.

4.3 Format Effects: Task-Oriented Mappings Amplify Dependency

The format of data further reshapes the structural fabric by specifying the directionality and specificity of dependencies. By partitioning sequences into input and output regions, we observe that the mapping from inputs to outputs ($\rho_{out|in}$) creates a much stronger structural fabric than the autoregressive modeling of the input alone (ρ_{in}), as shown in Figure 2c. This comparison isolates an effect that is often conflated with dataset quality: supervision format can make the same underlying content more or less useful by clarifying which parts of the input should constrain the output.

The instruction-tuning format reaches a dependency density of 46.53%, compared with 8.73% for raw intra-input modeling. This disparity arises because the **instruction-following format** imposes a goal-oriented constraint: the generated response must be strictly grounded in specific segments of the input query. This requirement transforms latent semantic correlations into explicit, high-strength dependencies, effectively filtering out irrelevant intra-input noise and concentrating dependency mass on the causal link between the task and the response.

Takeaway 1

Data configuration shapes the structural fabric through dependency density, which varies systematically with length, domain, and format.

5 Structural Fabric Shapes Gradient Geometry

Following the profiling of data structural fabric, we investigate how these intrinsic properties shape the optimization geometry of LLMs. We employ **gradient effective rank** as a diagnostic probe to measure the dimensionality of the update space during training, aggregating the layer-wise gradient effective rank for both Attention (W_Q, W_K, W_V, W_O) and Feed-Forward Network ($W_{Gate}, W_{Up}, W_{Down}$) parameters across three experimental settings. This analysis asks whether dense and coherent dependencies merely describe data, or whether they also reshape the directions along which the model actually updates its parameters.

5.1 Low-Rank Gradient Geometry across Components

Our empirical analysis reveals that the structural fabric of data consistently influences the optimization geometry. As illustrated in Figure 3, a clear trend emerges: the gradient effective rank of every module and layer is systematically lower under configurations with more coherent structural constraints. Nearly all data points cluster significantly below the diagonal line ($y < x$), suggesting that coherent structural organization acts as a global geometric regularizer, guiding optimization along a more focused, low-rank manifold and suppressing gradient conflicts.

Within this global trend, we identify specific spectral characteristics. First, the **absolute effective rank increases monotonically with layer depth**: as layers progress from shallow (purple) to deep (yellow/green), points shift toward the upper-right region, indicating that deeper layers naturally occupy a higher-dimensional optimization manifold. Second, **FFN modules consistently exhibit a higher effective rank than Attention modules** within the same layer. Crucially, FFN modules also demonstrate **greater sensitivity to the structural fabric of data**, as evidenced by their more pronounced shifts across the axes. This suggests that while Attention updates are inherently more low-rank, dense structural dependencies are particularly vital for regularizing the optimization space of the feed-forward blocks. The module-level pattern also supports our interpretation that structural fabric acts globally: it changes not only isolated layers, but the spectral profile of the training trajectory across the network.

5.2 How Data Structure Constrains Gradient Manifolds

Building upon the universal trends, we now examine how each specific data configuration uniquely reshapes the optimization geometry.

Long-Context Coherence as an Optimization Signal. The analysis of **data length** shows a consistent link between dependency density and optimization geometry. As illustrated in the left panel of Figure 3, long-context data yields a lower effective rank than short-context data, matching the higher dependency density of long-context data in Figure 2a. This suggests that extended contexts provide coherent long-range constraints that concentrate gradient energy into a lower-rank subspace. Short data can still exhibit useful local coupling, but its weaker normalized dependency density translates into a less concentrated update geometry.

Rigid Syntax as a Low-Rank Domain Constraint. For the **data domain**, code consistently induces a lower effective rank than natural language across all layers. This disparity stems from the reduced semantic ambiguity in programming tasks; while natural language allows for high-variance completions that scatter gradient directions, the **rigid logical dependencies** of code force gradients to collapse into a focused, low-rank subspace.

Conditional Mapping Filters Input Noise in Instruction Format. The effect of **data format** demonstrates how isolating task-oriented dependencies purifies the gradient update. Instruction-tuning data induces a lower effective rank compared to raw text modeling, a shift most pronounced in the FFN modules. This confirms that the **high-density conditional mapping** $P(Y|X)$ refines gradient signals and aligns updates into a cleaner trajectory, avoiding high-variance updates from intra-input noise.

Takeaway 2

Coherent structural fabric induces a lower gradient effective rank, suppressing gradient interference to ensure directionally coherent updates focused on causal dependencies.

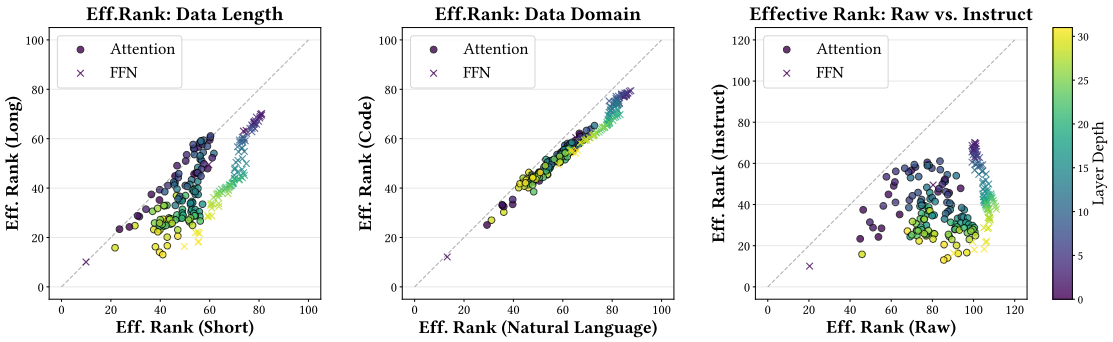


Figure 3: **Comprehensive Analysis of Gradient Effective Rank.** Each point represents the mean effective rank of a specific layer. Colors indicate the layer depth (0 to 31). Circles (○): Attention modules; crosses (×): FFN modules. Across all three dimensions, data points cluster below the diagonal ($y = x$), indicating that configurations with more coherent structural constraints induce a lower-rank optimization trajectory.

6 Geometric Stability Shapes Model Behavior

Building on the previous analyses, we now investigate how these structural properties manifest in the evolution of the model's **intrinsic dimension** and its final **downstream performance**. Our experiments are conducted using the Llama-3-8B architecture with a context window of 512K tokens. The training trajectory consists of two stages: first, the model undergoes Long-Context Continued Pre-training (LCPT) on the ProLong-data-512K dataset (Gao et al., 2025) to establish extensive long-context capabilities. Subsequently, we branch the model into two SFT regimes: one using **UltraChat** and the other using **ChatQA2**. We utilize these two datasets as **proxies for the data length configurations** analyzed in Section 4.1.

6.1 Structural Fabric Governs Geometric Stability of the Internal Manifold

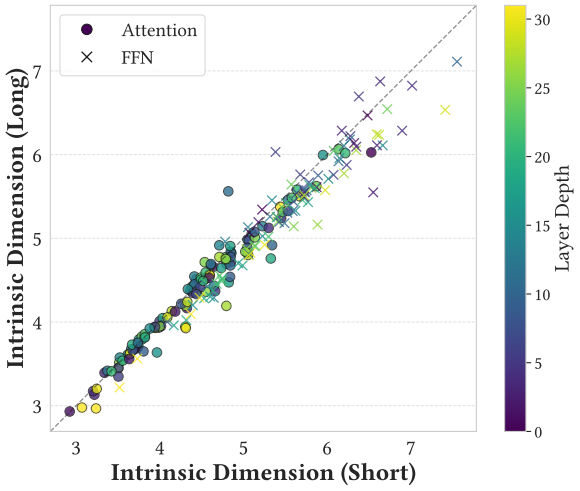
The evolution of intrinsic dimension illustrated in Figure 5 shows that the structural fabric of data shapes the geometric stability of internal representations by modulating the optimization manifold. During the LCPT phase, the model maintains a stable intrinsic dimension, reflecting a well-structured manifold inherited from large-scale pre-training.

We observe that the **short-context regime**, with weaker normalized dependency density, produces a higher gradient effective rank and immediately disrupts this stable manifold. This high-rank optimization regime triggers a sharp inflation of the intrinsic dimension from 4.733 to 4.861, indicating that isotropic gradient updates force the model to utilize redundant dimensions to encode noise. In stark contrast, the **long-context regime** exhibits superior geometric resilience. Its denser and more coherent long-range dependency structure keeps the gradient effective rank at a lower level, which in turn acts as a geometric regularizer that stabilizes the model's intrinsic dimensionality near 4.737. By ensuring directionally coherent updates, this long-context data allows the model to continue compressing knowledge into a structured manifold, reinforcing rather than destroying representational coherence. This result is consistent with the length analysis: stronger long-context dependency density aligns with lower-rank updates and better preservation of representation geometry.

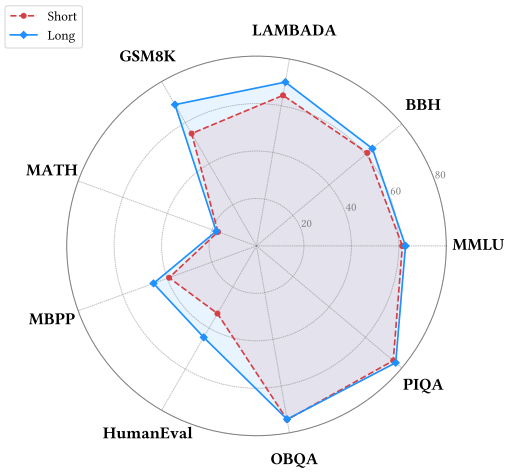
6.2 Compressed Internal Manifolds Improve Downstream Performance

We evaluate the final impact of data structural integrity by comparing the models resulting from the two SFT regimes. The layer-wise analysis in Figure 4a provides a comparative visualization of the intrinsic dimension across all Transformer projections. The x-axis represents the intrinsic dimension of the model tuned on the **short-context regime**, while the y-axis represents the model tuned on the **long-context regime**. Data points consistently cluster below the 45-degree diagonal line, indicating that the long-context model maintains a systematically lower intrinsic dimension across nearly all layers and modules. This consistent dimensional reduction is particularly pronounced in the deeper layers.

Critically, this difference in internal dimensionality serves as the geometric foundation for the observed performance gap. As illustrated in Figure 4b, the model with a more compressed intrinsic dimension consistently outperforms the one with a higher, more inflated dimensionality across various functional domains. The most substantial gains



(a) Comparison of intrinsic dimension across modules



(b) Downstream benchmark performance

Figure 4: Correlation between internal manifold compression and functional capabilities: (a) Data with coherent long-range structural constraints maintains a more compressed intrinsic dimension across model components; (b) this geometric stability is associated with stronger performance across benchmarks.

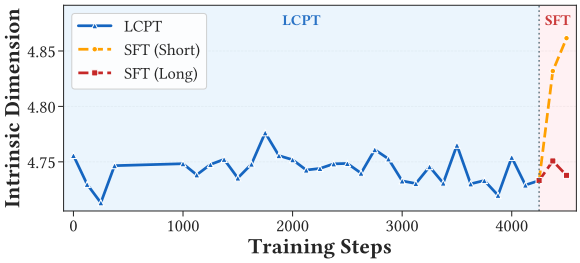


Figure 5: Evolution of intrinsic dimension during LCPT and SFT. The intrinsic dimension remains stable during LCPT and long-context tuning, but exhibits sharp inflation under the short-context tuning regime.

are observed in logic-heavy categories, such as mathematical reasoning and programming tasks. These results demonstrate that the dimensional expansion observed under the short-context regime is not merely a geometric artifact but a sign of representation noise that degrades the model's ability to handle precise logic and causal dependencies. The performance analysis therefore closes the loop of our framework: structural fabric shapes gradients, gradient geometry stabilizes representations, and stable representations support stronger functional behavior.

Takeaway 3

Coherent structural constraints induce a low-rank optimization geometry that stabilizes the model's intrinsic dimension, ensuring a compressed manifold that suppresses representation noise and improves complex reasoning performance.

7 Conclusion

This study establishes a unified geometric framework showing that the structural fabric of data shapes the training dynamics and functional capabilities of LLMs. We characterize this fabric with dependency density, showing that data configurations differ systematically in their semantic dependency patterns: long-context data exhibits denser dependencies than short-context data, code is denser than natural language within long-context data, and instruction-tuning mappings are denser than raw intra-input modeling. Across these configurations, coherent structural organization consistently induces a lower gradient effective rank, suppressing gradient interference and producing directionally coherent parameter updates. This low-rank optimization regime acts as a geometric regularizer that stabilizes the model's intrinsic dimension, preserving a compact representation manifold and improving performance on logic-intensive tasks. Overall, our results suggest that data quality should be understood not only through semantic content or surface statistics, but also through the dependency structures that make optimization coherent and representations stable.

Limitations

This work focuses on an analytical characterization of data structural fabric and its relationship to optimization geometry. While our results reveal consistent patterns across the datasets and configurations studied, we have not yet evalu-

ated how dependency-density-based analysis transfers to real-world data curation or deployment pipelines. Future work may study how these signals can guide practical training-data selection in application-specific settings.

Ethical Considerations

This paper presents an analytical study on the training dynamics and optimization geometry of LLMs. The experiments use standard, publicly available datasets and do not involve personally identifiable information or sensitive user data. We believe this work does not pose immediate ethical or societal risks.

We used AI assistants mainly for writing support, including language polishing, grammar correction, and improving clarity of presentation. The research ideas, experimental design, analysis, and final claims were developed, checked, and approved by the authors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online. Association for Computational Linguistics.

Longze Chen, Ziqiang Liu, Wanwei He, Yinhe Zheng, Hao Sun, Yunshui Li, Run Luo, and Min Yang. 2024. Long Context is Not Long at All: A Prospector of Long-Dependency Data for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8222–8234, Bangkok, Thailand. Association for Computational Linguistics.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Andres Fernandez, Frank Schneider, Maren Mahsereci, and Philipp Hennig. 2025. Connecting Parameter Magnitudes and Hessian Eigenspaces at Scale Using Sketched Methods. *arXiv preprint arXiv:2504.14701*.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2025. How to Train Long-Context Language Models (Effectively). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7376–7399, Vienna, Austria. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning. *Nature*, 645(8081):633–638.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. 2021. The Low-Rank Simplicity Bias in Deep Networks. *arXiv preprint arXiv:2103.10427*.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024. DataComp-LM: In Search of the Next Generation of Training Sets for Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pages 14200–14282. Curran Associates, Inc.

Ming Li, Yanhong Li, Ziyue Li, and Tianyi Zhou. 2025. How Instruction and Reasoning Data Shape Post-Training: Data Quality Through the Lens of Layer-wise Gradients. In *First Workshop on Foundations of Reasoning in Language Models*.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2024. #InsTag: Instruction Tagging for Analyzing Supervised Fine-tuning of Large Language Models. In *International Conference on Learning Representations*, volume 2024, pages 36456–36474.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, and 1 others. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv preprint arXiv:2112.11446*.

Olivier Roy and Martin Vetterli. 2007. The Effective Rank: A Measure of Effective Dimensionality. In *2007 15th European Signal Processing Conference*, pages 606–610.

Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, and 1 others. 2025. Less is More: Local Intrinsic Dimensions of Contextual Language Models. *arXiv preprint arXiv:2506.01034*.

Rishi Sonthalia, Michael Murray, and Guido Montúfar. 2025. Low Rank Gradients and Where to Find Them. *arXiv preprint arXiv:2510.01303*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language Models are Few-shot Multilingual Learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. 2025. Rethinking Data Selection at Scale: Random Selection is Almost All You Need. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2698–2711, Suzhou, China. Association for Computational Linguistics.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2024. ChatQA 2: Bridging the Gap to Proprietary LLMs in Long Context and RAG Capabilities. *arXiv preprint arXiv:2407.14482*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

A Model Specifications

We employ two distinct models for data dependency profiling and training dynamics analysis, balancing computational efficiency with task relevance.

- **LDS Calculation:** We utilize **Qwen3-0.6B** (Yang et al., 2025) to estimate segment-wise perplexity. This lightweight model enables efficient processing of large-scale datasets while providing a reliable approximation of semantic dependencies.
- **Gradient Analysis:** We use **Llama-3-8B-ProLong-512k-Base** (Gao et al., 2025) to investigate training dynamics and gradient effective rank. We select this model for its architectural relevance to modern high-performance LLMs and its optimized capacity for long-context processing.

B Detailed Analysis of Gradient Dynamics

B.1 Data Configurations Shape Spectral Decay of Gradients

To further investigate the optimization geometry, we analyze the singular value spectrum of the gradient matrices. As illustrated in Figure 6, the gradient spectra vary systematically across the data configurations analyzed in the main text.

Specifically, in the length comparison, long-context data exhibits a sharper initial drop compared to short-context sequences, indicating that long-range dependencies concentrate gradient information into a smaller, high-energy subspace. In the domain comparison, the rigid syntactical constraints of code result in a steeper relative decay and a thinner spectral tail than natural language, as the low-entropy nature of programming syntax naturally constrains the gradient manifold. In the format comparison, the goal-oriented mapping in instruction tuning accelerates spectral compression compared to raw text modeling, effectively filtering out isotropic noise. These results suggest that coherent dependency structure can act as an implicit regularizer, encouraging optimization to prioritize a low-dimensional manifold of causal dependencies.

B.2 Layer-wise Gradient Effective Rank across Data Configurations

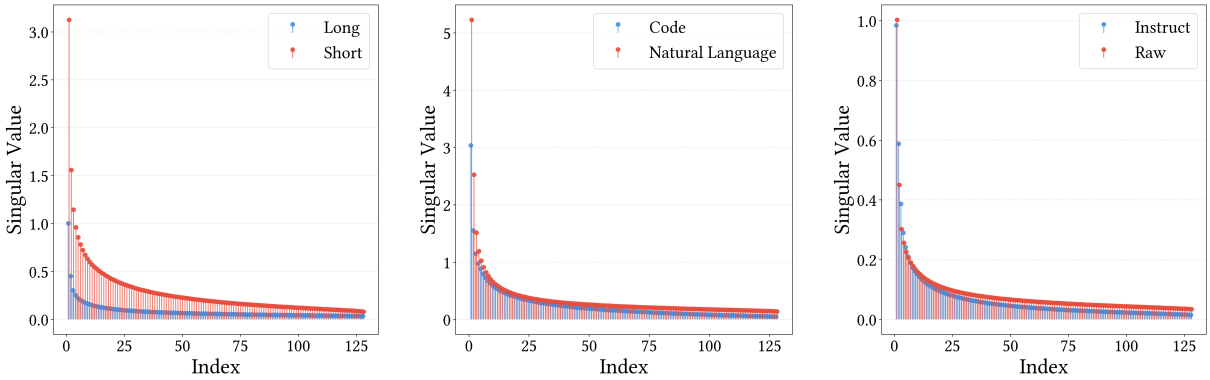
This section provides the complete layer-wise evolution of the gradient effective rank for the experiments discussed in Section 5. We report the effective rank for all projection matrices across both attention (W_Q, W_K, W_V, W_O) and feed-forward ($W_{gate}, W_{up}, W_{down}$) components.

In the figures below, solid lines represent the mean effective rank across training steps, with shaded regions denoting one standard deviation ($\pm 1\sigma$). The consistency of these patterns across all 32 layers demonstrates that the observed low-rank dynamics are systemic properties of the optimization process rather than localized artifacts.

Data Length (Figure 7). The comparison between short-context and long-context data shows similar rank behaviour in shallow layers (Layers 0–8), with a significant divergence emerging in deeper layers where long-context data exhibits a lower effective rank.

Data Domain (Figure 8). Code data induces a consistently lower gradient effective rank than natural language throughout the entire depth of the model, confirming the impact of rigid structural constraints on gradient dimensionality.

Data Format (Figure 9). The disparity between raw data and instruction-tuning data represents the most significant rank shift in our analysis. Instruction-tuning data consistently yields a lower-rank regime, isolating the structured conditional mapping $P(Y|X)$ from the high-entropy noise of raw text.



(a) Impact of Data Length (b) Impact of Data Domain (c) Impact of Data Format

Figure 6: **Spectral Decay of Gradients across different configurations.** The singular value spectra reveal how length, domain, and format reshape the concentration of gradient updates, with steeper decay corresponding to a more concentrated and low-rank optimization manifold.

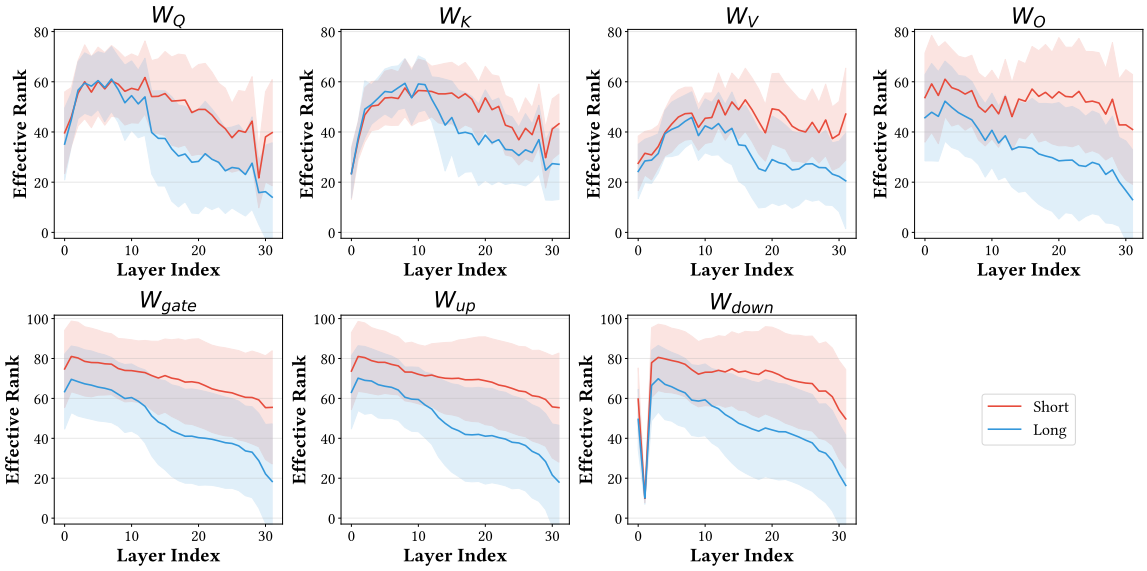


Figure 7: **Layer-wise Effective Rank: Short vs. Long.** Long-context data (blue) maintains a lower-rank manifold compared to short data (red). The divergence increases significantly in deeper layers (Layer > 10), confirming that long-range coherence primarily reshapes high-level representations.

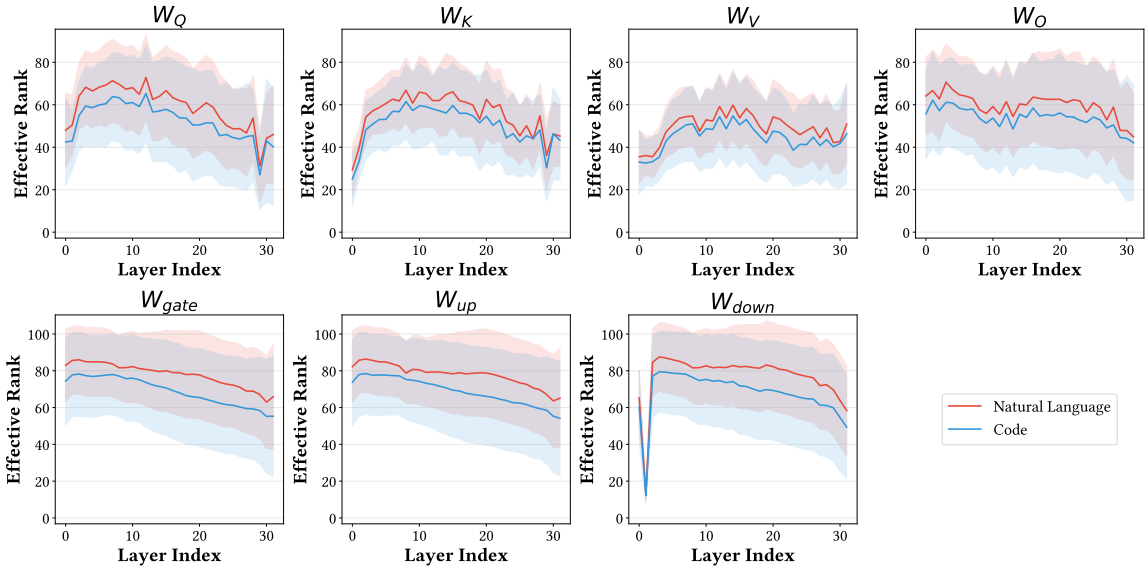


Figure 8: **Layer-wise Effective Rank: Natural Language vs. Code.** Code data (blue) induces a consistently lower rank than natural language (red) across all modules.

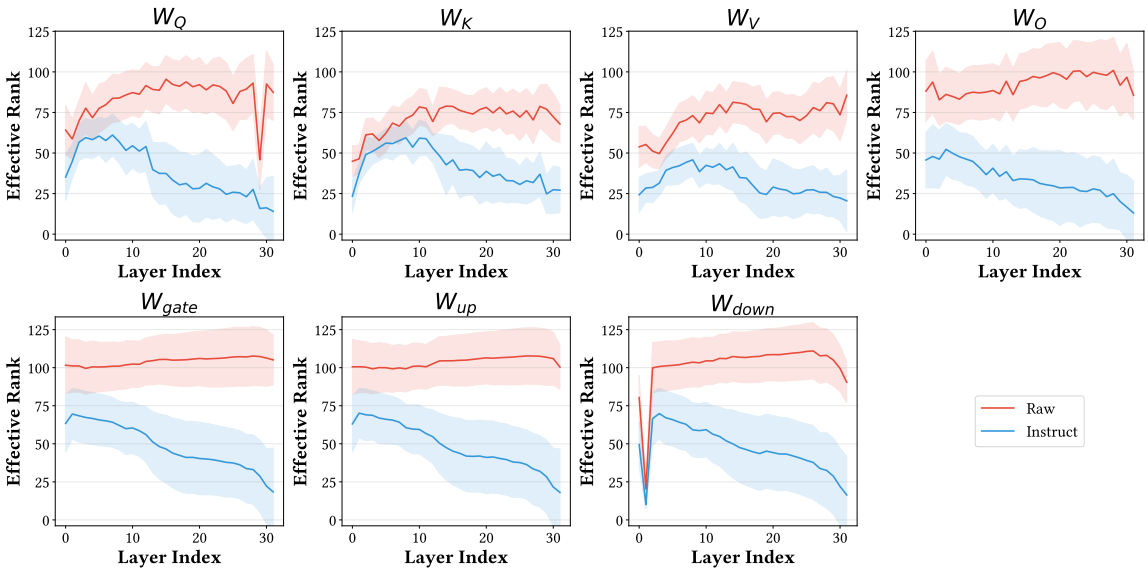


Figure 9: **Layer-wise Effective Rank: Raw Data vs. Instruction-tuning data.** Masking the input instruction loss (blue) leads to a dramatic collapse of the gradient rank compared to Joint Modeling (red).