

How Instruction and Reasoning Data shape Post-Training: Data Quality through the Lens of Layer-wise Gradients

Ming Li¹, Yanhong Li², Ziyue Li¹, Tianyi Zhou¹

¹University of Maryland; ²Allen Institute for AI

{minglii, tianyi}@umd.edu

Project: https://github.com/MingLiiii/Gradient_Unified

Abstract

As the post-training of large language models (LLMs) advances from instruction-following to complex reasoning tasks, understanding how different data affect finetuning dynamics remains largely unexplored. In this paper, we present a spectral analysis of layer-wise gradients induced by low/high-quality instruction and reasoning data for LLM post-training. Our analysis reveals that widely-studied metrics for data evaluation, e.g., IFD, InsTag, Difficulty, and Reward, can be explained and unified by spectral properties computed from gradients' singular value decomposition (SVD). Specifically, higher-quality data are usually associated with lower nuclear norms and higher effective ranks. Notably, effective rank exhibits better robustness and resolution than nuclear norm in capturing subtle quality differences. For example, reasoning data achieves substantially higher effective ranks than instruction data, implying richer gradient structures on more complex tasks. Our experiments also highlight that models within the same family share similar gradient patterns regardless of their sizes, whereas different model families diverge significantly. Providing a unified view on the effects of data quality across instruction and reasoning data, this work illuminates the interplay between data quality and training stability, shedding novel insights into developing better data exploration strategies for post-training.

1 Introduction

Large language models (LLMs) have shown remarkable potential for various complex tasks (Zhao et al., 2023; Xu et al., 2024a; Wu et al., 2025), yet their success in real-world applications hinges not just on model size but also on the quality of data used for training (Brown et al., 2020; Wang et al., 2023; Zhou et al., 2023). It is widely recognized that high-quality, diverse, and complex instruction-following examples during post-training (Xu et al., 2023; Ding et al., 2023; Li et al., 2023a; Liu et al.,

2023; Wu et al., 2024; Li et al., 2024b, 2025b), such as supervised fine-tuning (SFT), is crucial for eliciting generalization performance and reliability. There is growing recent interest in automated metrics that can evaluate data quality and select data for more efficient and effective post-training (Chen et al., 2023; Li et al., 2024e,d). Moreover, beyond instruction-following, the reasoning capability of LLMs (OpenAI et al., 2024b; DeepSeek-AI et al., 2025a) has also been proven to be largely dependent on data quality. For instance, s1.1 (Muenighoff et al., 2025) utilizes only 1k difficult math problems and DeepSeek-R1 generated responses to elicit LLM's strong reasoning capability.

Despite the verified importance of data quality to post-training, **how the quality of instruction/reasoning data affect the gradients during post-training** still remains largely unexplored. In addition, **can we unify different data quality metrics?** Prior work has mostly treated data quality filtering as a preprocessing step, evaluating its benefits in terms of end-task performance. But a systematic study is lacking to reveal the mechanism of how data quality affects the training dynamics. Meanwhile, there does not exist work that compares the learning dynamics induced by reasoning data and general instruction-following data or compares different data quality metrics' effects on post-training with data selection. A recent study of LLM post-training (Li et al., 2024c) on fast vs. slow thinking (Kahneman, 2011) for the first time analyzes the training dynamics on different data through the lens of layer-wise gradients. It discovers that learning detailed intermediate reasoning steps leads to smaller and more stable gradient updates than learning final answers only. Yet, this study focuses only on comparing "fast thinking" (a few answer tokens) vs. "slow thinking" (CoT paths) but does not extend to more challenging problems requiring more complex reasoning. The metric used to measure gradients is limited to magnitude

instead of more sophisticated spectral properties.

We address these gaps by conducting a **layer-wise, gradient-based spectral analysis** of LLM post-training when using **instruction/reasoning data of low/high quality**. Our study spans multiple diverse LLM families, including Qwen2.5 (Qwen et al., 2025), Llama3.1, Llama3.2 (Dubey et al., 2024), Gemma2 (Team et al., 2024)), and different sizes (1.5B - 14B), to ensure the generality of our findings. Inspired by s1 (Muennighoff et al., 2025), which selects data by difficulty, for reasoning data, we compare the s1.1 data and GSM8K (Cobbe et al., 2021) data (response generated by DeepSeek-R1) as low/high-quality data, respectively. For general instruction-following data, we adopt WizardLM (Xu et al., 2023), Magpie (Xu et al., 2024b), and OpenHermes 2.5 (Teknum, 2023) for our experiments and leverage automatic metrics for data quality, such as IFD (Li et al., 2024e), InsTag (Lu et al., 2023), Difficulty, and Reward to partition these datasets into low/high-quality subsets. While the first three evaluate the instructions, the reward directly measures the response quality.

We developed several novel metrics measuring the spectral properties of gradients revealed by Singular Value Decomposition (SVD) (**SVD-based Metrics**) and Gradient **Similarity-based Metrics**, which are applied to the projection layers for Query, Key, Value, and Output in transformer architectures (Vaswani et al., 2017): (1) *Nuclear Norm* measures the magnitude of the gradient, indicating the amount of changes and efforts required for post-training. (2) *Effective Rank* (Roy and Vetterli, 2007) captures the dimensionality of the gradient, indicating the diversity of the gradient directions. (3) *Same-layer Similarity* measures the alignment between gradients of different projections within the same layer. (4) *Adjacent-layer Similarity* measures the alignment of gradients between consecutive layers.

Main Contributions: In this study, we present a spectral analysis of layer-wise gradients in modern LLMs when finetuned on datasets of varying quality, namely high-quality, low-quality, instruction-following, and reasoning data. For the broad applicability of our conclusions, we conduct empirical investigations across diverse pretrained LLMs from multiple model families and on different datasets. Various automatic data evaluation metrics are used to split the data into low/high-quality partitions, and advanced reasoning data are also included for comparison. **Notably, we are the**

first to unify the effects of different data quality metrics and compare the instruction vs. reasoning data through the lens of layer-wise gradients.

Unlike existing work that focuses primarily on the gradient magnitude, we incorporate both SVD-based and similarity-based metrics, offering a more comprehensive analysis. These findings reveal previously overlooked gradient patterns and provide insights into enhancing the stability and efficiency of data synthesis and LLM training.

Our key findings:

1. Existing data quality metrics, e.g., IFD, InsTag, Difficulty, and Reward, can be *unified* due to consistent spectral properties of gradients, i.e., lower nuclear norms and higher effective ranks on high-quality data. This finding extends to both instruction and reasoning data, providing a unified view of the data quality effects.
2. Effective rank outperforms nuclear norm to distinguish low- vs. high-quality data. For reasoning data, s1.1 data yields the largest effective ranks across all experiments, suggesting high correlations between reasoning complexity and gradient diversity.
3. Within the same model family, layer-wise gradients’ spectral properties remain consistent across different model sizes. In contrast, the gradient patterns diverge significantly across distinct model families, reflecting the unique learning dynamics of each model family.
4. Cosine similarities between gradients from the same layer and adjacent layers remain nearly zero for different types of data, so they cannot reflect data quality.

2 Methodology

2.1 Preliminaries

We investigate gradient behaviors under the most widely adopted SFT approach. Each data point in an SFT dataset D consists of a pair (x, y) , where x is the instruction and y is the corresponding response. For the reasoning data, y concatenates both thinking tokens and response tokens. Let p_θ be a large language model with parameters θ . Under the SFT paradigm, p_θ is finetuned on each pair (x, y) by minimizing the following loss, where y_j denotes the j -th token of y , $y_{<j}$ denotes the preceding tokens, and l is the total length of y :

$$L_\theta = \frac{1}{l} \sum_{j=1}^l -\log(p_\theta(y_j | x, y_{<j})). \quad (1)$$

In this paper, we focus on the gradients of the layers associated with the attention mechanism (Vaswani et al., 2017), namely the *Query* (Q), *Key* (K), *Value* (V) projection layers and the final *Output* (O) projection layer. For simplicity, we denote the gradients of these projection layers by $G_{Q,i}$, $G_{K,i}$, $G_{V,i}$, and $G_{O,i}$ for each layer $i \in \{0, 1, \dots, N-1\}$ in the LLM.

2.2 Gradient Metrics from Spectral Analysis

To quantitatively analyze these gradients, we employ two categories of metrics: (1) *SVD-based Metrics* (the **Nuclear Norm** and the **Effective Rank**) and (2) *Similarity-based Metrics* (the **Same-layer Similarity** and the **Adjacent-layer Similarity**). While the SVD-based metrics describe properties of an individual gradient matrix, the similarity-based metrics reveal how gradients compare across different projections or layers.

2.2.1 SVD-based Metrics

Consider a gradient matrix $G_{X,i} \in \mathbb{R}^{m \times n}$, where $X \in \{Q, K, V, O\}$ indicates the projection (Query, Key, Value, or Output), and i represents the index the Transformer layer. We can write its Singular Value Decomposition as

$$G_{X,i} = U \Sigma V^T,$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix containing the singular values $\sigma_1, \dots, \sigma_{\min(m,n)}$, sorted in decreasing order. For simplicity, we omit all the subscripts (X and i) for SVD matrices and singular values.

Nuclear Norm. We measure the overall magnitude of the gradient matrix $G_{X,i}$ by its nuclear norm $\mathcal{N}_{X,i}$, which is formulated as

$$\mathcal{N}_{X,i} = \|G_{X,i}\|_* = \sum_{j=1}^{\min(m,n)} \sigma_j. \quad (2)$$

A higher nuclear norm indicates a larger overall gradient scale, implying that the model parameters at that layer are being updated more significantly, further indicating a potential distribution shift between the response and the model to be trained.

Effective Rank. We measure how uniformly the singular values of $G_{X,i}$ are distributed by the effective rank $\mathcal{R}_{X,i}$. We normalize the singular values and formulate the effective rank as

$$\mathcal{R}_{X,i} = \exp\left(-\sum_{j=1}^{\min(m,n)} \tilde{\sigma}_j \ln(\tilde{\sigma}_j)\right)$$

$$\tilde{\sigma}_j = \frac{\sigma_j}{\sum_{k=1}^{\min(m,n)} \sigma_k}$$

If only a few singular values are large (i.e., the gradient is concentrated in just a few directions), the effective rank is small. If many singular values all contribute significantly, the effective rank is relatively larger. It measures how diverse the directions of the gradient are. A higher effective rank indicates the gradient is spread out over more directions, suggesting richer updates, whereas a smaller value means that only a few directions dominate the gradient directions.

2.2.2 Similarity-based Metrics

While the SVD-based metrics above characterize individual gradient matrices in isolation, it can be equally insightful to analyze how gradients relate across projections or layers. To this end, we introduce cosine similarity measures at two levels: within the same layer and across adjacent layers.

3 Experimental Setup

3.1 Models

We evaluate our method on several pretrained LLMs across multiple families, covering a range of parameter scales. Specifically, we use Qwen2.5 (Qwen et al., 2025) in four configurations (1.5B, 3B, 7B, and 14B parameters), Llama3.1 (Dubey et al., 2024) with 8B parameters, Llama3.2 in 1B and 3B configurations, and Gemma2 (Team et al., 2024) in 2B and 9B configurations.

3.2 Datasets

Instruction-following data: **WizardLM** (Xu et al., 2023) is an instruction-following dataset created via an LLM-based “evolution” strategy that iteratively rewrites a set of initial prompts into more complex multi-step instructions, automatically generating high-complexity queries beyond what human annotators typically produce. **Magpie** (Xu et al., 2024b) is a fully synthetic alignment dataset obtained by prompting an aligned language model, and we utilize the 300k high-quality subset selected by the authors for our source data. **OpenHermes 2.5** is a large-scale curated compilation of roughly one million instruction–response samples drawn from a diverse range of open-source and GPT-4-generated datasets, designed to maximize diversity and task coverage for robust fine-tuning.

Reasoning data: We employ two sources of reasoning data, each paired with step-by-step traces and final solutions generated by DeepSeek-R1 (DeepSeek-AI et al., 2025a): **s1.1K** data is provided by s1, which contains difficult math problems with responses generated by DeepSeek-R1. This data can be viewed as high-quality reasoning data since it succeeds in eliciting LLMs’ reasoning capability with only 1,000 samples. To curate the relatively low-quality reasoning data, motivated by s1’s success in utilizing difficulty as the metric, we utilize the relatively easy math problems from **GSM8K** (Cobbe et al., 2021) with DeepSeek-R1 generated responses.

3.3 Data Quality Metrics to Partition Low/High-Quality Data

We adopt four automated data evaluation metrics to analyze instruction-following data quality. We focus on data evaluation metrics that do not rely on additional evaluation sets or training, which might lead to customized shifting, while we aim at the effects of original properties of the data instances.

IFD (Li et al., 2024e) quantifies the instruction-following difficulty of a sample by computing the ratio between the model’s perplexity when predicting the response without the instruction and with the instruction. A higher IFD indicates that the model fails to benefit from the instruction, suggesting that the instruction is either ambiguous or unhelpful. Following Li et al. (2024d), we use a small GPT2 model to efficiently compute IFD scores over large datasets.

InsTag (Lu et al., 2023) performs open-set multi-label tagging over instructions, capturing semantic attributes such as domain, task type, and intent. We use two derived metrics: (1) *instruction complexity*, defined by the number of tags per sample, and (2) *instruction diversity*, defined by tag vocabulary coverage across the dataset. Higher complexity scores typically correspond to more elaborate, multifaceted instructions. We use the per-sample complexity score as the filtering signal.

Reward Model Score uses *sfairXC/FsfairX-LLaMA3-RM-v0.1*, a reward model fine-tuned via preference modeling to predict human-aligned helpfulness (Xiong et al., 2024). Given a data pair, the model outputs a reward score reflecting predicted alignment with human preferences.

GPT-4o Difficulty Rating prompts the GPT-4o model to assign a difficulty score (*from 1 to 10*) to each instruction based on the perceived complex-

ity, ambiguity, and reasoning depth required. This method approximates a human-aligned evaluation of instruction difficulty.

For each dataset described above, and for each metric, we select 200 samples with the highest scores and 200 with the lowest scores for calculating the gradients. This allows us to isolate what each metric considers “good-quality” and “low-quality” data and to conduct gradient-based analysis across these contrasting subsets.

4 Empirical Analysis

4.1 Instruction-Following Data

4.1.1 Unifying Different Data Quality Metrics

In this section, we compare the effects of different data filtering metrics for general instruction-following data toward the gradient properties, including the averaged nuclear norms and effective ranks, as shown in Table 1. The nuclear norms in the table are calculated across the layers, $\mathcal{N}_X = \sum_{i=0}^{N-1} \mathcal{N}_{X,i}$, and the effective ranks are similar. Specifically, we use the WizardLM (Xu et al., 2023) data as the source data and calculate gradients on the Qwen2.5-7B model, and the data filtering metrics include IFD (Li et al., 2024e) (calculated on the Qwen2.5-7B model), InsTag (Lu et al., 2023), Difficulty, and Reward.

As shown in the table, across all metrics, the high-quality subsets exhibit substantially smaller averaged nuclear norms in the layer-wise gradients, represented by the consistent negative values (in red) for the Gap columns. Since the nuclear norm measures the overall magnitude of gradient updates; thus, a smaller value suggests the model requires less energy to adapt to high-quality data. It further indicates that high-quality data should be aligned with the learned knowledge of pretrained LLMs. At the same time, high-quality subsets also yield consistently larger effective ranks in their gradients represented by the large positive values (in green) for the Gap columns. A higher value suggests that more update directions are activated, which means high-quality data leads to richer, more multi-dimensional parameter updates, which likely improves the model’s ability to generalize and capture nuanced features of the instruction pairs.

Our finding illustrates that multiple different definitions of “data quality” can converge to overarching gradient properties, revealing a unified view by gradient-based spectral analysis.

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
WizardLM	IFD	k	1.3	6.1	-4.8	q	1.4	4.6	-3.2	k	88.5	14.2	74.3	q	131.9	12.7	119.2
		v	2.5	10.9	-8.4	o	2.7	8.3	-5.6	v	110.9	13.6	97.3	o	165.2	12.5	152.7
	InsTag	k	1.9	4.1	-2.2	q	1.9	3.2	-1.3	k	95.6	19.9	75.7	q	141.2	21.5	119.7
		v	3.1	7.6	-4.5	o	3.3	5.9	-2.6	v	120.7	20.8	99.9	o	180.5	22.3	158.2
	Difficulty	k	1.9	3.5	-1.6	q	1.9	2.7	-0.8	k	91.5	18.5	73.0	q	133.2	20.0	113.2
		v	3.1	6.8	-3.7	o	3.3	5.2	-1.9	v	114.9	19.1	95.8	o	167.8	20.7	147.1
	Reward	k	1.2	4.3	-3.1	q	1.2	3.6	-2.4	k	91.5	35.6	55.9	q	131.4	38.6	92.8
		v	2.1	7.8	-5.7	o	2.3	6.5	-4.2	v	113.2	36.9	76.3	o	166.8	41.4	125.4

Table 1: Nuclear norms and effective ranks of gradients calculated from high- or low-quality data selected by different metrics. *High* represents the high-quality subset, while *Low* represents the low-quality subset. *Gap* is calculated by $High - Low$. **All data quality metrics consistently identify the high-quality data, which show similar spectral properties of gradients: lower nuclear norms and higher effective ranks. Hence, the gradient properties can unify all metrics.**

4.1.2 Effects of Original Dataset Qualities

In this section, we aim to investigate the potential ineffectiveness of gradient properties for distinguishing high- or low-quality data. Similar to the previous settings, we split the data from different sources, WizardLM (Xu et al., 2023), OpenHermes 2.5 (Teknum, 2023), and Magpie (Xu et al., 2024b), into high- and low-quality subsets, calculating their effects on gradients based on Qwen2.5-7B. The nuclear norm and effective rank changes with layer indexes are shown in Figure 1. The shapes of gradient curves are almost kept the same for one specific model, whatever the data quality.

Moreover, we find that when an entire dataset is already composed of fairly clean and coherent instruction–response samples, e.g., the Magpie dataset shown in the third row, nuclear norms offer limited discriminative power in distinguishing it further, represented by the similar scales for both types of data. We hypothesize that once instructions and responses cross a certain threshold of clarity and consistency (i.e., minimal noise or confusion), the gradient magnitudes needed to adapt to these subsets no longer diverge sharply. In other words, the nuclear norm may effectively “saturate” and stop registering small but meaningful differences among these relatively high-quality groups. On the contrary, effective ranks remain sensitive to smaller quality disparities, even within a dataset already recognized for solid instruction–response fidelity. Consequently, effective rank acts as a finer-grained lens, revealing that while both subsets are indeed “good” enough to produce stable gradients, the higher-quality examples still manage to activate a broader range of update directions.

4.2 Reasoning Data

4.2.1 Unifying Quality Evaluation of Instruction and Reasoning Data

Beyond analyzing general instruction-following data, the recent surge in reasoning models encourages us to further explore the effects of reasoning data on LLM gradients. Motivated by recent success on eliciting models’ reasoning capabilities by distilling from stronger reasoning models through simple SFT, e.g., s1 (Muennighoff et al., 2025), LIMO (Ye et al., 2025), DeepSeek-R1 Distilled Qwen (DeepSeek-AI et al., 2025a), we formulate the exploration on reasoning data in the same structure as on instruction-following data. Compared with general instruction-following data, a key difference of advanced reasoning data is the utilization of dynamic long CoTs and the accordance with the test-time scaling law. To distinguish this data from general instruction-following data, we notate them as reasoning data. As illustrated by s1, 1,000 difficult math problems are sufficient to elicit LLMs’ reasoning capabilities; thus, we ask: **Does reasoning data have similar effects to gradients with previous general instruction data? Can the quality of advanced reasoning data be further distinguished by gradient properties like nuclear norm and effective rank?**

To formulate our experiments, we utilize the s1.1 data as the high-quality subset and GSM8K (Cobbe et al., 2021) (responses also generated by DeepSeek-R1 (DeepSeek-AI et al., 2025a)) as the low-quality subset for our reasoning data experiments. The results are shown in Table 2, in which the gradients are calculated based on Qwen2.5-7B, comparing the general instruction-following data with the reasoning data. Surprisingly, our experiments demonstrate that the same

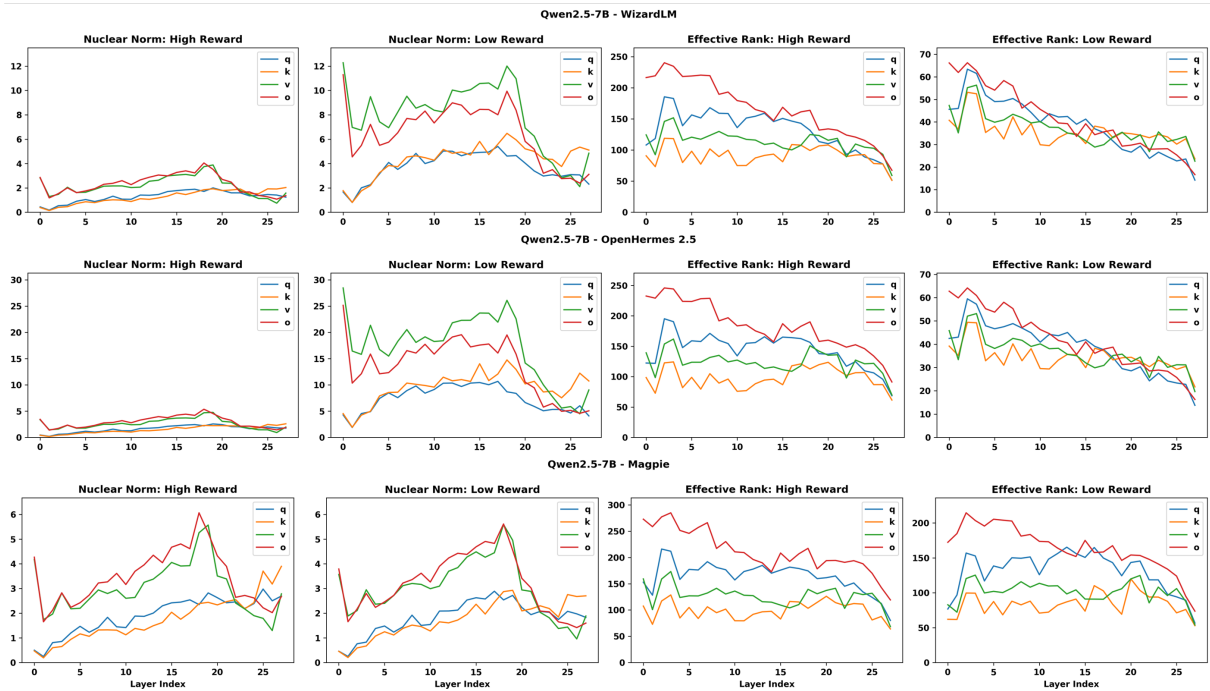


Figure 1: Low/high-quality data (measured by Reward) and their gradient properties (nuclear norms and effective ranks) across layers on diverse datasets including WizardLM, OpenHermes 2.5, and Magpie. The y-axis scales are kept the same for nuclear norms, while different for effective ranks, due to the large discrepancy. **For each specific model, the shapes of the gradient curves derived from different data sources are almost the same. The nuclear norm fails to reflect the quality discrepancies between datasets, while the effective ranks still works promisingly, e.g., Magpie has higher rank than others.**

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Magpie	Difficulty	k	1.7	1.7	0.0	q	1.8	1.8	0.0	k	95.9	83.8	12.1	q	153.3	124.4	28.9
		v	3.0	3.1	-0.1	o	3.3	3.3	0.0	v	118.0	102.7	15.3	o	195.1	151.7	43.4
Reasoning	Difficulty	k	1.0	1.3	-0.3	q	1.3	1.5	-0.2	k	138.8	106.1	32.7	q	361.2	203.3	157.9
		v	1.9	2.4	-0.5	o	2.5	2.8	-0.3	v	170.4	126.7	43.7	o	509.9	263.1	246.8

Table 2: Comparing the gradient properties between instruction data vs. reasoning data. For the reasoning data, *High* denotes data sampled from s1.1, and *Low* denotes sampled from GSM8K with DeepSeek-R1 responses. Reasoning data shows lower nuclear norms and higher effective ranks compared with instruction data. **Our analysis of gradients unifies the quality evaluation for reasoning and instruction data on both higher effective ranks reflecting higher quality. Moreover, the metric distinguishes low-/high- quality data by large gaps.**

gradient-derived metrics, nuclear norms and effective ranks, remain applicable for the advanced reasoning data: (i) *Reasoning vs. Instruction data*: Reasoning data, even the lower-quality subset, leads to lower nuclear norms and higher effective ranks compared with previous high-quality instruction data, suggesting the much higher data quality of recent reasoning data. (ii) *Higher- vs. Lower-quality reasoning data*: Even for the reasoning data with supreme data quality, there still exists a consistent and unified trend with the previous instruction data, i.e., smaller nuclear norms and larger effective ranks for higher-quality data. Moreover, the gaps in effective ranks between high- and low-quality subsets are more pronounced for reasoning data

than for general instruction data, which might provide a potential explanation on why s1.1 can reach such a promising performance with only 1000 data.

We are the first to investigate and compare the effects of general instruction-following and reasoning data toward LLM gradients in the training process. We reveal that both data types can be unified into a consistent pattern of gradient-based signals regarding quality, providing a unified view for understanding the quality effects.

4.2.2 Effects of Model Sizes

Figure 2 presents the gradient curves of LLMs with different sizes in the Qwen2.5 family on our reasoning data. The shapes of the gradient curves for rea-

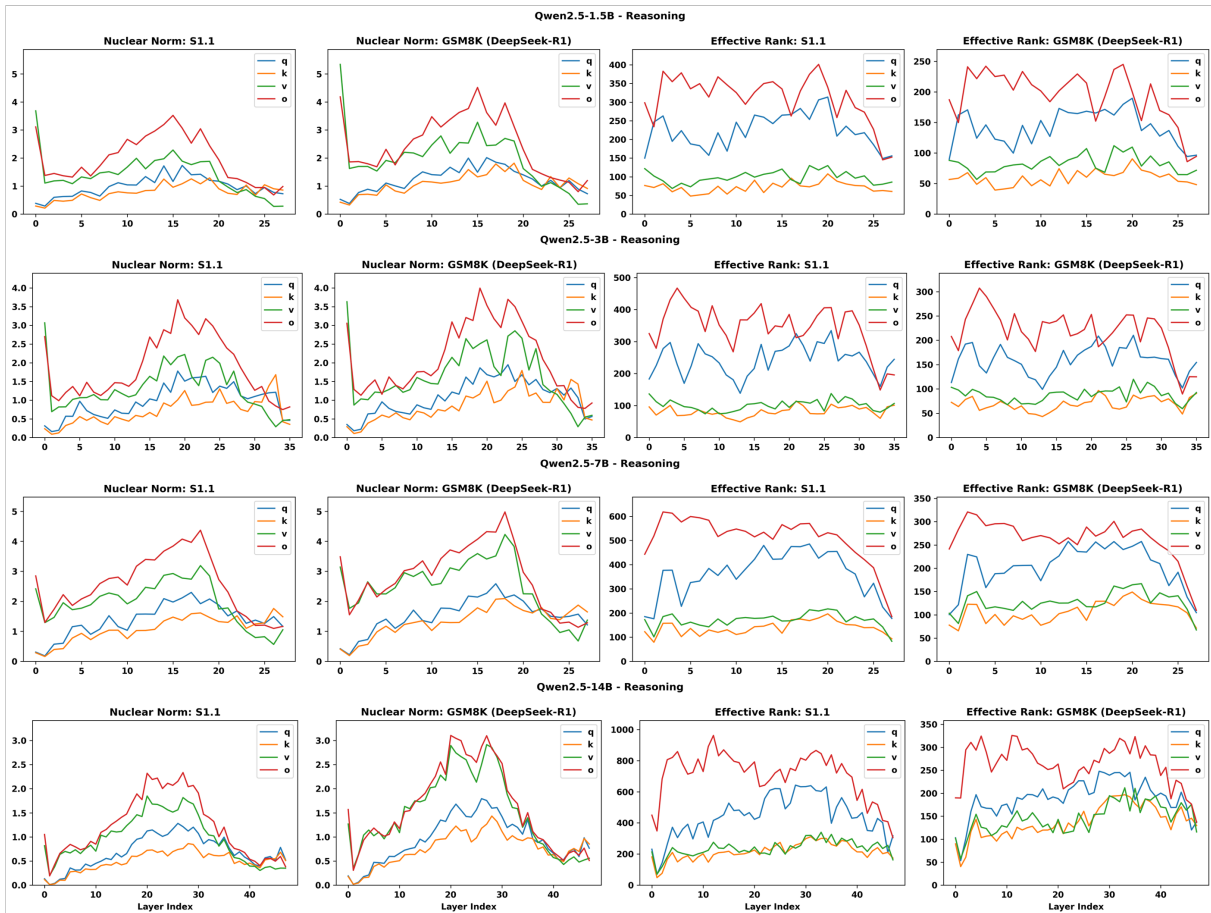


Figure 2: Model size scaling law for gradient properties. **Within the same model family, the layer-wise gradient statistics and dynamics are relatively consistent. Gradients on larger models exhibit better capabilities to distinguish data quality**, revealed by the increasing y-axis scales from the 1.5B model to 14B model.

soning data are almost the same as the instruction-following data, which further verifies our unified view on both types of data. Moreover, a consistent observation is that the overall shape of these layer-wise curves remains relatively (not strictly) stable as we move from smaller to larger models. For instance, GSM8K-based fine-tuning triggers a peak in nuclear norm at the mid layers for Qwen2.5-1.5B, a qualitatively similar, albeit scaled, peak in Qwen2.5-14B also can be identified.

Moreover, another interesting finding is that larger models tend to amplify the distinction between high- and low-quality subsets. Specifically, the gaps in effective rank between s1.1 and GSM8K grow accordingly when moving to bigger models. In other words, the larger model is more sensitive to whether the provided reasoning path is coherent and informative.

4.2.3 Effects of Model Families

In this section, we broaden the scope to compare entirely different model families, including

Qwen2.5-3B, Llama3.2-3B, and gemma2-2B, each possessing distinct pretraining recipes and architectural configurations. The first notable observation is that the layer-by-layer “shape” of the gradients can vary significantly among families. For instance, Llama3.2-3B might exhibit consistently higher nuclear norms in its early layers compared to Qwen2.5-3B, reflecting differences in embedding or attention initialization. Despite these baseline discrepancies, the relative gap between high- and low-quality reasoning data persists across all families. In other words, regardless of how each model is architected or pre-trained, high-quality data still yields smaller nuclear norms and larger effective ranks. This cross-family analysis suggests that reasoning data can be a valuable resource regardless of the specific LLM architectures. At the same time, our analysis shows the existence of family-specific “fingerprints” for nuclear norms and effective ranks, reflecting architectural and pretraining differences, which might potentially be useful for a better understanding of the LLM architectures.

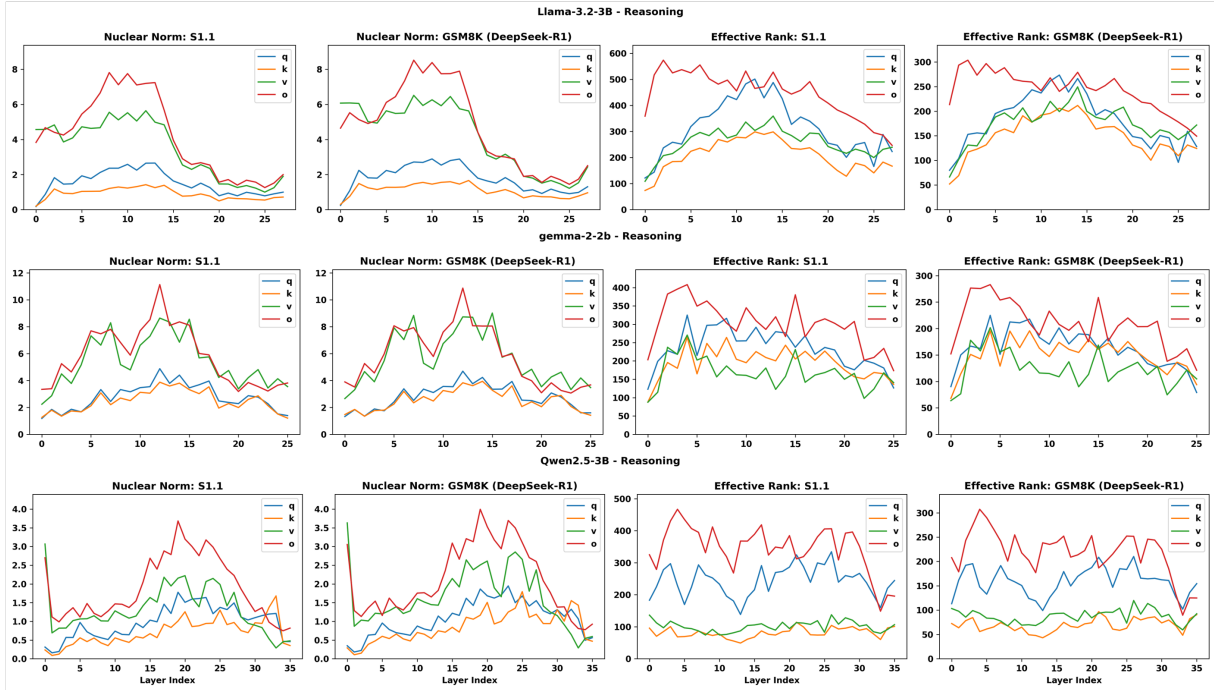


Figure 3: Gradient properties across different model families. **The gradient dynamics of the same data on different model families are largely different. This might be caused by their distinct model structures or training recipes and may reflect their different capabilities.**

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Magpie	Difficulty	k - v	-0.7e-3	-0.7e-3	q - o	0.0e-3	0.0e-3	k	-0.1e-3	-0.1e-3	q	0.1e-3	-0.1e-3
		v	0.3e-3	0.1e-3	o	0.4e-3	0.3e-3						
Reasoning	Difficulty	k - v	0.0e-3	0.0e-3	q - o	0.1e-3	-0.9e-3	k	-2.0e-3	-2.0e-3	q	0.0e-3	0.2e-3
		v	1.0e-3	2.0e-3	o	1.0e-3	2.0e-3						

Table 3: Gradient similarity metrics remain excessively small and cannot reflect the differences between instruction/reasoning data of low/high-quality. **It shows that layer-wise gradients in LLM post-training are nearly orthogonal, indicating that the similarity of gradients is not an effective indicator of data quality.**

4.2.4 Similarity-based Metrics

While our SVD-based metrics are proved consistently effective at capturing data-quality differences, similarity-based metrics, namely, same-layer similarity and adjacent-layer similarity, do not appear to yield meaningful signals in our experiments, as shown in Table 3. We keep the values in the same magnitude for easier comparison. In the table, we compare similarity measures for both instruction and reasoning data across high- and low-quality subsets. Regardless of dataset type or quality level, the reported cosine similarities remain extremely close to zero, with minimal observable variation. These low similarities suggest that the gradients for LLM SFT are nearly orthogonal, indicating that similarity on gradients is not an effective indicator of data quality.

5 Conclusion

We introduce a unified gradient-based framework for analyzing how varying data quality, ranging from general instruction-following to reasoning data, shapes the finetuning of LLMs. By examining the layer-wise gradients, we show that different quality metrics converge on remarkably similar gradient signatures, specifically, smaller gradient magnitudes and broader gradient directions for high-quality data. Notably, this pattern holds across multiple model families and parameter scales and, for the first time, reveals how reasoning data induce even higher effective ranks and thus richer parameter updates.

Limitation

The main limitation of this work is not conduct further experiments on whether these gradient spectral statistics can be directly utilized for follow-up data selection. However, as mentioned in the paper, the main contribution of this work lies in the first investigation on the unified effects of data quality metrics on gradients, rather than utilizing it as a new data selection method.

References

- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. [Large language models for mathematical reasoning: Progresses and challenges](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian’s, Malta. Association for Computational Linguistics.
- Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houlston, Tomasz Sternal, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, Łukasz Flis, Hannes Eberhard, Hubert Niewiadomski, and Torsten Hoefler. 2025. [Reasoning language models: A blueprint](#). *Preprint*, arXiv:2501.11223.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Alexander Bukharin and Tuo Zhao. 2023. [Data diversity matters for robust instruction tuning](#). *Preprint*, arXiv:2311.14736.
- Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. Instruction mining: Instruction data selection for tuning large language models. In *First Conference on Language Modeling*.
- Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2023. [Alpapasus: Training a better alpaca with fewer data](#). *Preprint*, arXiv:2307.08701.
- Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025a. [Towards reasoning era: A survey of long chain-of-thought for reasoning large language models](#). *Preprint*, arXiv:2503.09567.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025b. [Do not think that much for 2+3=? on the overthinking of o1-like llms](#). *Preprint*, arXiv:2412.21187.
- Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. [Gradnorm: Gradient normalization for adaptive loss balancing in deep multi-task networks](#). *Preprint*, arXiv:1711.02257.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, and 4 others. 2025. [Process reinforcement through implicit rewards](#). *Preprint*, arXiv:2502.01456.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. [Enhancing chat language models by scaling high-quality instructional conversations](#). *arXiv preprint arXiv:2305.14233*.
- Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. [Mods: Model-oriented data selection for instruction tuning](#). *Preprint*, arXiv:2311.15653.
- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. [GLM: General language model pretraining with autoregressive blank infilling](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 33 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. 2025. [Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill?](#) *Preprint*, arXiv:2504.06514.

- Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. 2019. [An investigation into neural net optimization via hessian eigenvalue density](#). *Preprint*, arXiv:1901.10159.
- Guy Hacohen and Daphna Weinshall. 2019. [On the power of curriculum learning in training deep networks](#). *Preprint*, arXiv:1904.03626.
- Alex Havrilla, Andrew Dai, Laura O’Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, Duy Phung, Maia Iyer, Dakota Mahan, Chase Blagden, Srishti Gureja, Mohammed Hamdy, Wen-Ding Li, Giovanni Paolini, Pawan Sasanka Amanamanchi, and Elliot Meyerson. 2024. [Surveying the effects of quality, diversity, and complexity in synthetic data from large language models](#). *Preprint*, arXiv:2412.02980.
- Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wen Yue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. [The impact of reasoning step length on large language models](#). *Preprint*, arXiv:2401.04925.
- Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hananeh Hajishirzi. 2020. [UNIFIEDQA: Crossing format boundaries with a single QA system](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.
- Pang Wei Koh and Percy Liang. 2020. [Understanding black-box predictions via influence functions](#). *Preprint*, arXiv:1703.04730.
- Chunyu Li, Heerad Farkhor, Rosanne Liu, and Jason Yosinski. 2018. [Measuring the intrinsic dimension of objective landscapes](#). *Preprint*, arXiv:1804.08838.
- Haoling Li, Xin Zhang, Xiao Liu, Yeyun Gong, Yifan Wang, Qi Chen, and Peng Cheng. 2025a. [Enhancing large language model performance with gradient-based parameter selection](#). *Preprint*, arXiv:2406.15330.
- Ming Li, Han Chen, Chenguang Wang, Dang Nguyen, Dianqi Li, and Tianyi Zhou. 2025b. [Ruler: Improving llm controllability by rule-based data recycling](#). *Preprint*, arXiv:2406.15938.
- Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. 2024a. [Selective reflection-tuning: Student-selected data recycling for LLM instruction-tuning](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16189–16211, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, and Tianyi Zhou. 2023a. [Reflection-tuning: Recycling data for better instruction-tuning](#). In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.
- Ming Li, Pei Chen, Chenguang Wang, Hongyu Zhao, Yijun Liang, Yupeng Hou, Fuxiao Liu, and Tianyi Zhou. 2024b. [Mosaic-it: Free compositional data augmentation improves instruction tuning](#). *Preprint*, arXiv:2405.13326.
- Ming Li, Yanhong Li, and Tianyi Zhou. 2024c. [What happened in llms layers when trained for fast vs. slow thinking: A gradient perspective](#). *Preprint*, arXiv:2410.23743.
- Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024d. [Superfiltering: Weak-to-strong data filtering for fast instruction-tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14255–14273, Bangkok, Thailand. Association for Computational Linguistics.
- Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024e. [From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7595–7628, Mexico City, Mexico. Association for Computational Linguistics.
- Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023b. [Self-alignment with instruction back-translation](#). *arXiv preprint arXiv:2308.06259*.
- Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025c. [From system 1 to system 2: A survey of reasoning large language models](#). *Preprint*, arXiv:2502.17419.
- Liangxin Liu, Xuebo Liu, Derek F Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024. [Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection](#). *Advances in Neural Information Processing Systems*, 37:97800–97825.
- Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. [What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning](#). *arXiv preprint arXiv:2312.15685*.
- Yue Liu, Jiaying Wu, Yufei He, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, and Bryan Hooi. 2025. [Efficient inference](#)

- for large reasoning models: A survey. *Preprint*, arXiv:2503.23077.
- Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models. *Preprint*, arXiv:2308.07074.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.
- OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024b. Openai o1 system card. *Preprint*, arXiv:2412.16720.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.
- Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. 2020. Estimating training data influence by tracing gradient descent. *Preprint*, arXiv:2002.08484.
- Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-Sheng Hua, Bowen Zhou, and Yu Cheng. 2025. A survey of efficient reasoning for large reasoning models: Language, modality, and beyond. *Preprint*, arXiv:2503.21614.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.
- Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE.
- Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. 2018. Empirical analysis of the hessian of over-parametrized neural networks. *Preprint*, arXiv:1706.04454.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.
- Weixi Song, Zuchao Li, Lefei Zhang, Hai Zhao, and Bo Du. 2024. Sparse is enough in fine-tuning pre-trained large language models. *Preprint*, arXiv:2312.11875.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 29 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.
- Teknum. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Gianis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krithika Doshi, Kuntal Kumar Pal, and 16 others. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yufei Wang, Wanjuan Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang,

- Xin Jiang, and Qun Liu. 2023. [Aligning large language models with human: A survey](#). *Preprint*, arXiv:2307.12966.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2024. [Lamini-lm: A diverse herd of distilled models from large-scale instructions](#). *Preprint*, arXiv:2304.14402.
- Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Lijie Hu, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, and Ninghao Liu. 2025. [Usable xai: 10 strategies towards exploiting explainability in the llm era](#). *Preprint*, arXiv:2403.08946.
- Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. [Less: Selecting influential data for targeted instruction tuning](#). In *International Conference on Machine Learning*, pages 54104–54132. PMLR.
- Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2024. [Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint](#). *Preprint*, arXiv:2312.11456.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. [Wizardlm: Empowering large language models to follow complex instructions](#). *Preprint*, arXiv:2304.12244.
- Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024a. [A survey on knowledge distillation of large language models](#). *arXiv preprint arXiv:2402.13116*.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024b. [Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing](#). *Preprint*, arXiv:2406.08464.
- Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. [CrossFit: A few-shot learning challenge for cross-task generalization in NLP](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. [Limo: Less is more for reasoning](#). *arXiv preprint arXiv:2502.03387*.
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. [Gradient surgery for multi-task learning](#). *Preprint*, arXiv:2001.06782.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. [A survey of large language models](#). *Preprint*, arXiv:2303.18223.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu 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](#). *Preprint*, arXiv:2305.11206.

Table of Contents

A Related Work	13
A.1 Gradient Analysis	13
A.2 SFT and Data Selection	14
A.3 Reasoning Capability	15
B Prompts for Data Evaluation	15
C Gradient Curves for Different Models and Metrics	17
C.1 Qwen2.5 1.5B	17
C.2 Qwen2.5 3B	22
C.3 Qwen2.5 7B	27
C.4 Llama 3.1 8B	32
C.5 Llama 3.2 1B	37
C.6 Llama 3.2 3B	42
C.7 Gemma 2 2B	47

A Related Work

A.1 Gradient Analysis

Empirical analyses of gradient dynamics have provided deep insights into how large language models (LLMs) learn during both pretraining and fine-tuning. A recurring observation is that LLM training tends to concentrate updates in a low-dimensional subspace, where only a small fraction of parameters account for the majority of the gradient magnitude (Song et al., 2024). This sparsity in meaningful directions underlies the effectiveness of methods like parameter-efficient fine-tuning and low-rank adaptation. For example, Li et al. (2025a) propose gradient masking to zero out low-impact updates and concentrate learning on parameters with high gradient magnitude. These observations are consistent with findings from spectral studies

of the Hessian (Sagun et al., 2018; Ghorbani et al., 2019), which reveal that deep networks—including transformers—tend to learn in a space governed by a few dominant gradient directions. This low effective dimensionality has been linked to both generalization and the robustness of optimization trajectories (Li et al., 2018).

Gradient-based perspectives have also proven valuable in understanding how different types of training data influence learning. In curriculum learning, for instance, it has been observed that ordering data from easy to hard helps align gradients in earlier phases of training, resulting in smoother convergence and better generalization (Hacohen and Weinshall, 2019). In multi-task and continual learning, techniques like GradNorm (Chen et al., 2018) and PCGrad (Yu et al., 2020) explicitly manipulate gradient magnitudes and directions to resolve conflicts between tasks. Furthermore, gradient-based influence functions (Koh and Liang, 2020; Pruthi et al., 2020) have been used to trace how individual data points affect model updates and predictions, enabling researchers to detect noisy or highly influential examples. In the context of instruction tuning and alignment, such analyses can reveal how instruction-following or reasoning-rich data differentially steer the model’s behavior.

A closely related work by Li et al. (2024c) studies how cognitive styles (fast vs. slow thinking) impact LLM gradient dynamics, showing that reasoning-rich (slow thinking) data yields smaller, stable gradients compared to direct-answer (fast thinking) data. Extending this, we introduce new metrics: effective rank for gradient complexity, and two similarity metrics (Same-layer Similarity, Adjacent-layer Similarity) to analyze gradient alignment within and between layers. These metrics provide deeper insights into how instruction-following and reasoning-rich data influence LLM training.

A.2 SFT and Data Selection

Supervised fine-tuning (SFT) is the subsequent step where an LLM is further trained on labeled examples to specialize it for specific tasks or to align it with human instructions. This process updates the model’s weights using task-specific input–output pairs (e.g., prompts and desired responses) (Wang et al., 2023) InstructGPT (Ouyang et al., 2022) demonstrated that fine-tuning GPT-3 on human-curated instruction datasets yields models that more

effectively follow user instructions compared to the base models. This approach, referred to as instruction tuning, enhances performance and exhibits robust generalizability.

In the instruction tuning, the concept “quality is all you need” is widely accepted (Zhou et al., 2023; Touvron et al., 2023; Havrilla et al., 2024). Earlier research focused on curating high-quality datasets through human experts or powerful LLMs (Khashabi et al., 2020; Ye et al., 2021; Wei et al., 2022; Wang et al., 2022; Du et al., 2022; Taori et al., 2023). Although these approaches yield high-quality datasets, they are time-consuming, expensive, and offer limited diversity.

However, LIMA (Zhou et al., 2023) demonstrates that as few as 1,000 high-quality, human-curated training instances can substantially enhance instruction-following performance. Building on this work, data selection has emerged as an increasingly critical stage in instruction tuning. InsTag (Lu et al., 2023) uses a powerful proprietary model to tag samples within SFT datasets based on semantics and intentions, subsequently selecting data that meet predefined criteria for complexity and diversity in their tags. Alpapas (Chen et al., 2023) utilizes proprietary LLMs chatGPT and Claude2 to automatically identify and filter out low-quality data within the dataset. Cherry LLM (Li et al., 2024e) introduces Instruction-Following Difficulty (IFD) scores—a self-guided metric for evaluating instruction difficulty without relying on additional LLMs—and uses these scores to select datasets. Motivated by Humpback (Li et al., 2023b), Selective Reflection-Tuning (Li et al., 2024a) introduces a teacher-student collaborative pipeline that is guided by the IFD score and its reverse version to select the data based on the evaluation feasibility. Du et al. (2023) and Bukharin and Zhao (2023) utilize reward models as the metric for measuring data quality and subsequently select the data. DEITA (Liu et al., 2023) employs ChatGPT to diversify datasets, then applies various data selection metrics to construct a high-quality dataset. Superfiltering (Li et al., 2024d) demonstrates that both weak and strong language models consistently assess instruction difficulty, thereby streamlining the filtering process. Instruct Mining (Cao et al.) presents a method for automatically selecting high-quality instruction-following data using natural language quality indicators. SelectIT (Liu et al., 2024) proposes an uncertainty-aware self-filtering approach that leverages an LLM’s intrinsic

sic uncertainty to select high-quality instruction-tuning data LESS (Xia et al., 2024) constructs a low-dimensional “gradient datastore” for candidate data, subsequently selecting examples whose gradient influence closely matches that of a limited set of target demonstration examples. Collectively, these studies focus on distinguishing high-quality data samples from lower-quality ones for effective instruction tuning.

A.3 Reasoning Capability

Recent advancements in large language model architectures and training objectives have led to a paradigm shift from fast, heuristic (System-1) response generation to deliberate, multi-step (System-2) reasoning processes (Ahn et al., 2024; Li et al., 2024c; Jin et al., 2024; Besta et al., 2025; Li et al., 2025c; Chen et al., 2025a), marked by advanced reasoning models such as OpenAI’s o1/o3 (OpenAI et al., 2024b), QwQ (Qwen et al., 2025), and DeepSeek-R1 (DeepSeek-AI et al., 2025a). System 1 models (e.g., GPT-4o (OpenAI et al., 2024a), LLaMA-3 (Dubey et al., 2024), DeepSeek-V3 (DeepSeek-AI et al., 2025b)) generate intuitive, rapid responses but often struggle with more intricate tasks. In contrast, System 2 models adopt methodical analysis and iterative self-critique, which strengthens their performance on complex reasoning benchmarks.

Various post-training methods have been introduced to further enhance these capabilities. A number of studies employ reinforcement learning strategies to guide models toward superior reasoning approaches (Shao et al., 2024; Cui et al., 2025). Moreover, researchers have demonstrated that instruction tuning on meticulously curated, high-quality datasets can greatly boost a model’s reasoning performance (Ye et al., 2025; Muennighoff et al., 2025). However, despite the advancement in reasoning capability, the issue of overthinking (Chen et al., 2025b; Fan et al., 2025; Qu et al., 2025; Liu et al., 2025) brings further challenges to the area.

B Prompts for Data Evaluation

For our selected data quality evaluation metrics, In-Tag and Difficulty are the prompt-based methods. We use GPT4o for the evaluation and the prompts are provided in Figure 4 and Figure 5.

Prompt for InsTag

System Prompt

You are a helpful assistant.

User Prompt

You are a tagging system that provides useful tags for instruction intentions to distinguish instructions for a helpful AI assistant. Below is an instruction:

[begin]

{*instruction*}

[end]

Please provide coarse-grained tags, such as "Spelling and Grammar Check" and "Cosplay", to identify the main intentions of the above instruction. Your answer should be a list that includes the titles of tags and a brief explanation of each tag. You can provide several tags as you wish. Your response has to strictly follow this JSON format: [{"tag": str, "explanation": str}, {"tag": str, "explanation": str}, ...]. Please respond in English.

Figure 4: The prompt for InsTag.

Prompt for Difficulty

System Prompt

You are a helpful assistant.

User Prompt

You are a difficulty estimation system that can rate the difficulty level of instruction intentions. Below is an instruction:

[begin]

{*instruction*}

[end]

The instruction can be tagged with a difficulty level from 1 to 10, where 1 is the easiest and 10 is the hardest. Please rate the difficulty level of the instruction. Please first output a single line containing the difficulty score. Then, provide a brief explanation of why you rated the instruction with that difficulty score.

Figure 5: The prompt for Difficulty.

C Gradient Curves for Different Models and Metrics

C.1 Qwen2.5 1.5B

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
WizardLM	Difficulty	k	1.6	4.2	-2.6	q	1.6	3.4	-1.7	k	56.2	14.0	42.2	q	104.5	16.4	88.1
		v	2.7	7.9	-5.2	o	3.2	6.5	-3.3	v	81.2	16.9	64.4	o	142.3	18.3	124.0
	IFD (GPT2)	k	1.8	4.7	-3.0	q	1.8	3.6	-1.8	k	44.4	15.7	28.7	q	74.7	17.3	57.5
		v	3.2	7.9	-4.7	o	3.6	6.7	-3.2	v	62.9	18.6	44.3	o	96.6	19.9	76.8
	InsTag	k	1.6	5.0	-3.4	q	1.7	4.0	-2.4	k	57.9	15.0	42.8	q	108.9	17.8	91.1
		v	2.7	9.1	-6.4	o	3.2	7.6	-4.3	v	83.5	18.5	65.0	o	148.8	20.1	128.7
	Reward	k	1.1	4.4	-3.3	q	1.2	3.8	-2.6	k	56.8	24.7	32.1	q	103.9	30.1	73.8
		v	1.9	7.8	-5.8	o	2.3	7.1	-4.7	v	80.7	30.3	50.4	o	140.4	35.2	105.2

Table 4: Qwen2.5-1.5B - WizardLM - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
WizardLM	Difficulty	k - v	-6.9e-04	-1.3e-03	q - o	1.8e-06	7.1e-06	k	6.0e-04	6.1e-04	q	-1.6e-04	4.9e-04
		v			o			v	3.9e-04	-8.1e-04	o	-7.3e-04	2.2e-03
	IFD (GPT2)	k - v	-7.0e-04	-9.9e-04	q - o	-4.3e-06	3.6e-06	k	2.5e-04	3.6e-04	q	1.3e-04	5.9e-04
		v			o			v	3.8e-04	-9.5e-05	o	-5.5e-04	3.5e-03
	InsTag	k - v	-1.2e-03	-6.4e-04	q - o	1.1e-05	1.1e-05	k	1.1e-03	2.4e-04	q	2.3e-04	6.9e-04
		v			o			v	2.9e-04	-3.0e-05	o	-8.6e-04	3.7e-03
	Reward	k - v	-4.7e-04	-6.6e-04	q - o	-9.3e-06	-7.0e-06	k	7.5e-04	1.1e-03	q	7.5e-05	-7.6e-05
		v			o			v	3.7e-04	-1.9e-04	o	-1.4e-04	1.6e-03

Table 5: Qwen2.5-1.5B - WizardLM - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
OpenHermes 2.5	Difficulty	k	2.0	4.2	-2.2	q	2.0	3.4	-1.5	k	55.2	21.1	34.1	q	101.2	26.6	74.6
		v	3.4	7.3	-4.0	o	3.8	6.6	-2.8	v	79.4	26.6	52.9	o	136.5	30.5	106.0
	IFD (GPT2)	k	1.6	11.0	-9.4	q	1.7	8.7	-7.0	k	49.0	21.0	28.0	q	85.7	24.3	61.4
		v	3.0	18.8	-15.8	o	3.5	16.3	-12.8	v	69.3	25.1	44.1	o	110.7	28.9	81.8
	InsTag	k	1.6	4.1	-2.5	q	1.7	3.4	-1.7	k	57.6	20.3	37.4	q	107.6	25.7	81.9
		v	2.8	7.1	-4.4	o	3.3	6.5	-3.2	v	82.5	25.8	56.8	o	145.2	29.9	115.3
	Reward	k	1.2	8.7	-7.5	q	1.4	6.9	-5.6	k	58.6	24.0	34.6	q	108.5	30.6	77.9
		v	2.2	15.4	-13.2	o	2.8	13.4	-10.6	v	84.2	29.5	54.8	o	147.1	36.2	110.9

Table 6: Qwen2.5-1.5B - OpenHermes 2.5 - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
OpenHermes 2.5	Difficulty	k - v	-4.1e-04	-5.2e-04	q - o	8.9e-06	1.1e-05	k	6.8e-04	-3.4e-04	q	7.3e-05	5.5e-05
		v			o			v	4.4e-04	1.1e-04	o	-4.8e-04	1.7e-03
	IFD (GPT2)	k - v	-1.1e-03	6.3e-04	q - o	-1.0e-05	5.2e-06	k	1.5e-04	6.4e-04	q	-1.3e-04	7.9e-05
		v			o			v	6.5e-05	-4.0e-05	o	-5.4e-04	2.3e-03
	InsTag	k - v	-7.8e-04	-1.5e-04	q - o	1.2e-06	2.2e-06	k	1.7e-03	-2.0e-04	q	1.4e-04	4.0e-04
		v			o			v	-3.1e-04	2.0e-04	o	-6.1e-04	2.0e-03
	Reward	k - v	-3.0e-04	-5.8e-04	q - o	-4.1e-07	-3.8e-06	k	6.1e-04	1.1e-03	q	-3.0e-04	-2.1e-04
		v			o			v	2.2e-06	-1.8e-04	o	-1.2e-03	3.0e-03

Table 7: Qwen2.5-1.5B - OpenHermes 2.5 - Similarity-based Metrics

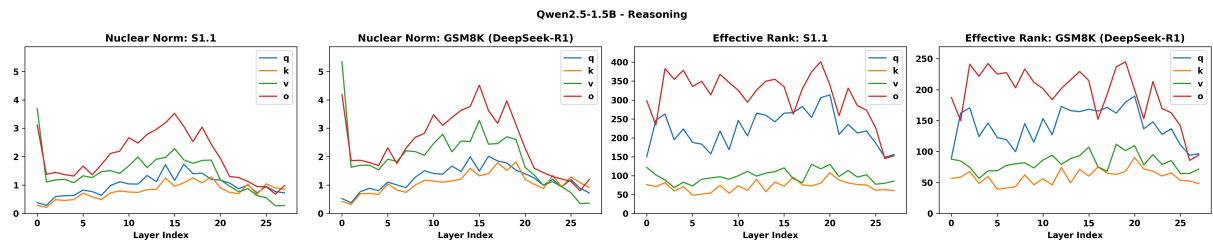


Figure 6: Qwen2.5 1.5B - Reasoning Data

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Magpie	Difficulty	k	1.5	1.5	-0.0	q	1.6	1.6	-0.0	k	55.8	51.7	4.1	q	110.0	92.3	17.7
		v	2.6	2.8	-0.1	o	3.2	3.3	-0.1	v	78.6	73.2	5.4	o	149.2	121.8	27.4
	IFD (GPT2)	k	1.4	1.2	0.2	q	1.6	1.2	0.3	k	58.5	55.7	2.8	q	117.5	99.1	18.4
		v	2.6	2.1	0.5	o	3.3	2.5	0.8	v	83.6	77.3	6.3	o	158.7	135.4	23.3
	InsTag	k	1.5	1.5	0.0	q	1.6	1.6	0.1	k	61.6	52.1	9.5	q	124.1	96.7	27.5
		v	2.6	2.7	-0.0	o	3.2	3.2	0.0	v	86.8	73.9	13.0	o	170.5	128.9	41.6
	Reward	k	1.4	1.4	-0.0	q	1.6	1.5	0.1	k	58.1	48.8	9.3	q	118.7	94.9	23.8
		v	2.5	2.5	-0.1	o	3.2	2.9	0.3	v	84.4	67.5	16.9	o	164.2	128.3	35.9

Table 8: Qwen2.5-1.5B - Magpie - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Magpie	Difficulty	k - v	-2.4e-03	-1.5e-03	q - o	-1.1e-05	7.9e-06	k	-2.7e-04	3.6e-04	q	-2.0e-04	-6.8e-05
		v	1.5e-03	3.3e-04	o	1.4e-05	-1.1e-03						
	IFD (GPT2)	k - v	-1.6e-03	-1.6e-03	q - o	1.2e-05	-1.2e-05	k	-6.7e-04	1.4e-03	q	-1.4e-04	-6.0e-05
		v	3.1e-04	2.2e-04	o	-9.4e-04	-1.1e-04						
	InsTag	k - v	-1.3e-03	-1.5e-03	q - o	-4.0e-06	1.0e-05	k	1.5e-03	-2.2e-04	q	-6.5e-05	3.0e-04
		v	-4.2e-05	1.1e-03	o	-2.7e-04	-1.5e-03						
	Reward	k - v	-1.5e-03	-3.3e-03	q - o	-1.5e-05	7.8e-06	k	6.1e-04	-6.4e-04	q	-3.0e-04	-6.0e-04
		v	6.0e-04	2.0e-03	o	-2.1e-03	-1.2e-03						

Table 9: Qwen2.5-1.5B - Magpie - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Reasoning	Reasoning	k	0.8	1.1	-0.3	q	1.0	1.2	-0.3	k	72.1	59.5	12.6	q	223.8	141.7	82.1
		v	1.4	2.0	-0.6	o	2.0	2.5	-0.5	v	98.7	82.6	16.1	o	311.7	194.2	117.5

Table 10: Qwen2.5-1.5B - Reasoning - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Reasoning	Reasoning	k - v	4.8e-04	6.6e-04	q - o	-4.2e-06	-1.2e-05	k	1.6e-03	3.6e-04	q	5.3e-04	-1.2e-04
		v	9.1e-04	4.2e-03	o	5.2e-04	-9.7e-04						

Table 11: Qwen2.5-1.5B - Reasoning - Similarity-based Metrics

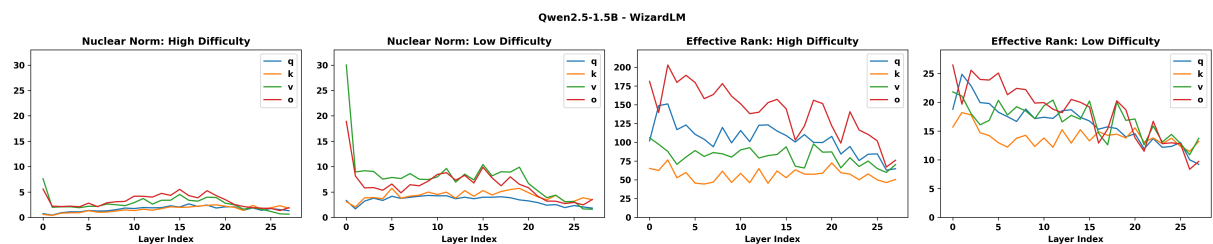


Figure 7: Qwen2.5 1.5B - WizardLM with Difficulty Metric

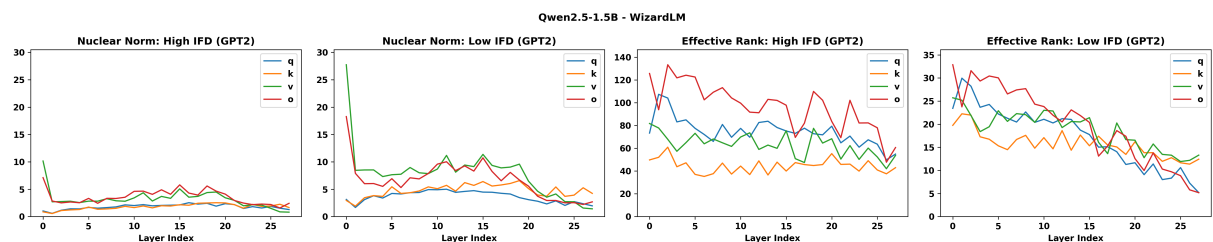


Figure 8: Qwen2.5 1.5B - WizardLM with IFD (GPT-2) Metric

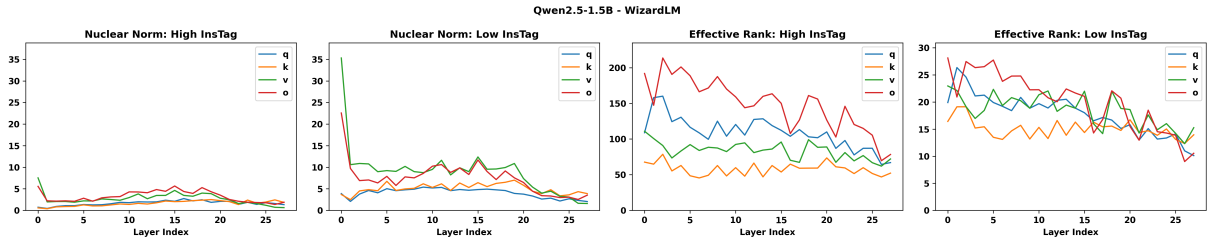


Figure 9: Qwen2.5 1.5B - WizardLM with InsTag Metric

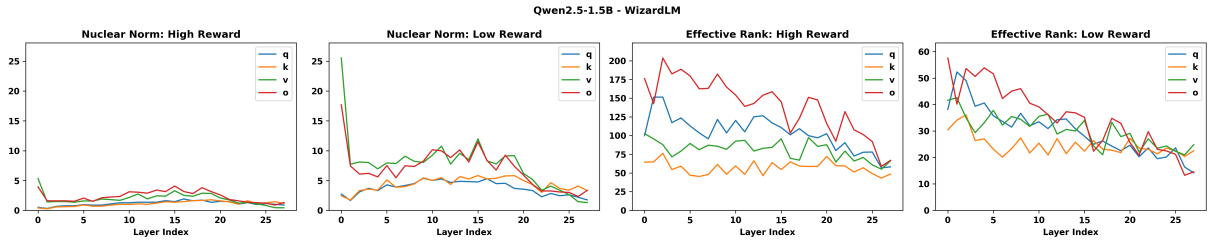


Figure 10: Qwen2.5 1.5B - WizardLM with Reward Model Metric

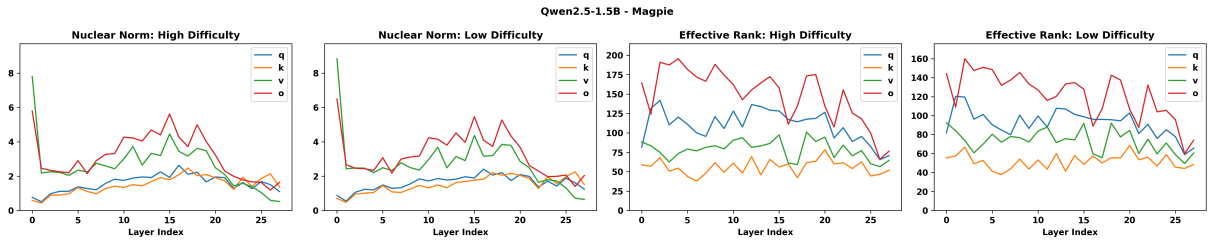


Figure 11: Qwen2.5 1.5B - Magpie with Difficulty Metric

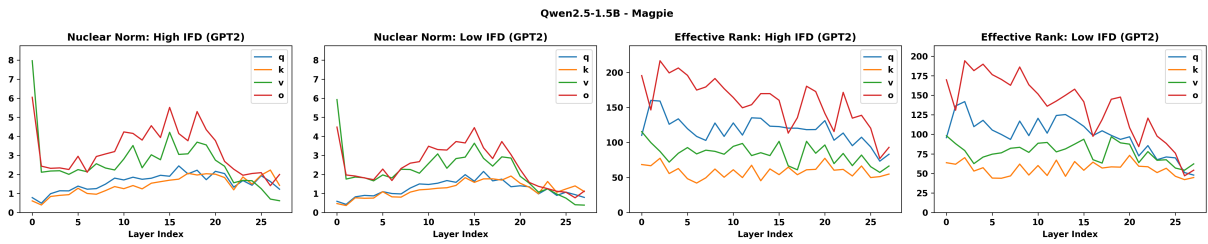


Figure 12: Qwen2.5 1.5B - Magpie with IFD (GPT-2) Metric

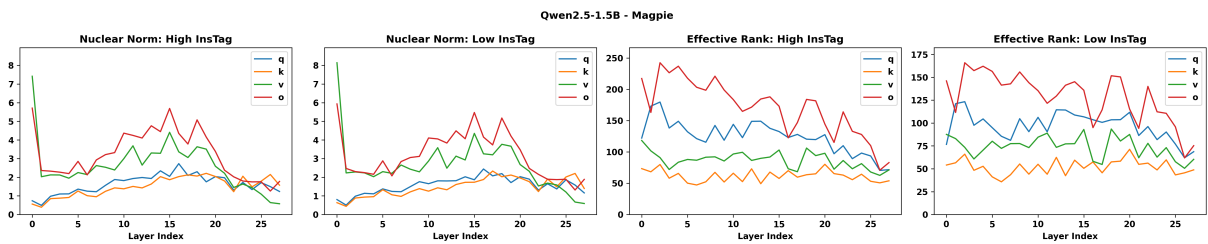


Figure 13: Qwen2.5 1.5B - Magpie with InsTag Metric

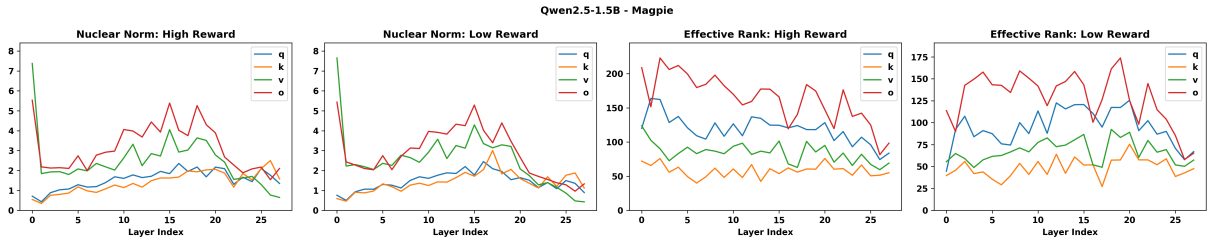


Figure 14: Qwen2.5 1.5B - Magpie with Reward Model Metric

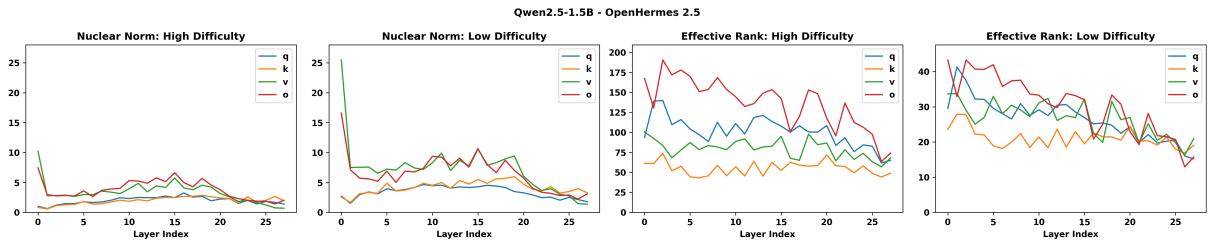


Figure 15: Qwen2.5 1.5B - OpenHermes with Difficulty Metric

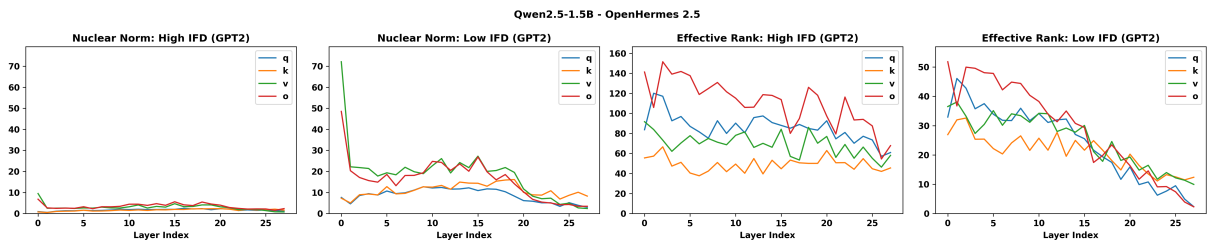


Figure 16: Qwen2.5 1.5B - OpenHermes with IFD (GPT-2) Metric

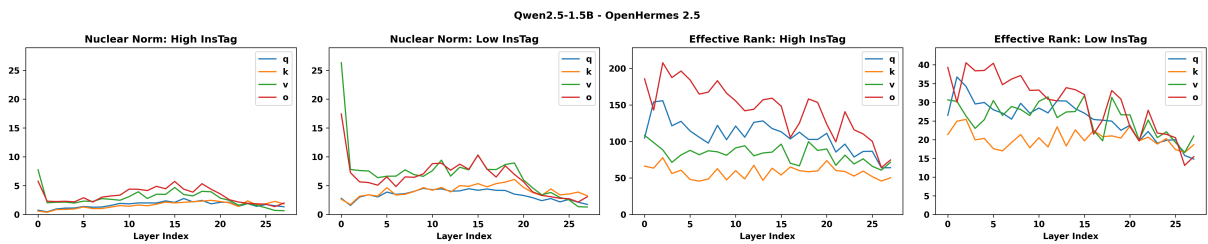


Figure 17: Qwen2.5 1.5B - OpenHermes with InsTag Metric

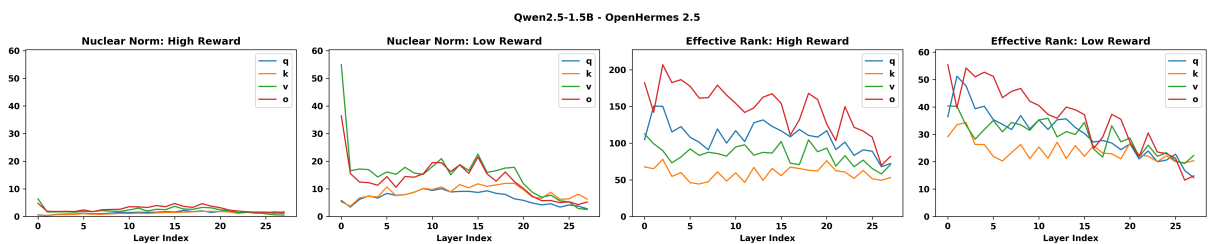


Figure 18: Qwen2.5 1.5B - OpenHermes with Reward Model Metric

C.2 Qwen2.5 3B

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
WizardLM	Difficulty	k	1.2	2.9	-1.7	q	1.3	2.5	-1.2	k	65.6	15.9	49.7	q	118.7	18.0	100.7
		v	2.1	5.7	-3.6	o	2.5	4.9	-2.4	v	87.3	18.3	69.1	o	160.4	19.8	140.7
	IFD (GPT2)	k	1.3	3.2	-2.0	q	1.4	2.8	-1.4	k	51.9	18.2	33.8	q	84.0	19.5	64.5
		v	2.4	5.7	-3.3	o	2.7	5.0	-2.3	v	68.6	20.3	48.3	o	106.3	22.0	84.3
	InsTag	k	1.2	3.6	-2.4	q	1.3	3.2	-1.9	k	67.2	17.1	50.2	q	123.8	19.5	104.3
		v	2.1	6.8	-4.8	o	2.5	6.0	-3.5	v	89.5	19.7	69.8	o	168.2	21.5	146.7
	Reward	k	0.8	3.2	-2.4	q	0.9	3.0	-2.1	k	66.0	28.8	37.1	q	119.3	34.0	85.3
		v	1.5	6.0	-4.6	o	1.8	5.5	-3.8	v	86.9	33.2	53.6	o	161.7	38.6	123.1

Table 12: Qwen2.5-3B - WizardLM - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
WizardLM	Difficulty	k - v	-5.6e-04	-5.4e-04	q - o	9.1e-06	-3.9e-06	k	-1.3e-03	-7.8e-04	q	2.5e-05	1.0e-04
		v	-5.6e-04	-5.4e-04	o	9.1e-06	-3.9e-06	v	-5.1e-04	-6.4e-04	o	1.4e-03	2.1e-03
	IFD (GPT2)	k - v	-2.7e-04	-8.4e-04	q - o	1.9e-06	-1.2e-05	k	-1.3e-03	-1.1e-03	q	-2.3e-04	2.7e-05
		v	-2.7e-04	-8.4e-04	o	1.9e-06	-1.2e-05	v	3.9e-04	3.4e-04	o	1.9e-03	2.4e-03
InsTag	k - v	-6.3e-04	-8.8e-05	q - o	4.0e-06	1.3e-07	k	-1.2e-03	-7.4e-04	q	3.0e-05	-1.5e-05	
	v	-6.3e-04	-8.8e-05	o	4.0e-06	1.3e-07	v	-5.1e-04	6.8e-06	o	1.5e-03	1.7e-03	
Reward	k - v	-1.1e-04	-1.2e-04	q - o	7.9e-06	4.6e-06	k	-6.5e-04	-1.1e-03	q	-2.5e-04	-1.2e-04	
	v	-1.1e-04	-1.2e-04	o	7.9e-06	4.6e-06	v	-1.7e-04	-8.0e-04	o	9.4e-04	1.4e-03	

Table 13: Qwen2.5-3B - WizardLM - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
OpenHermes 2.5	Difficulty	k	1.3	3.0	-1.8	q	1.4	2.8	-1.3	k	64.7	24.1	40.6	q	115.5	29.2	86.3
		v	2.2	5.6	-3.4	o	2.7	5.2	-2.5	v	85.8	28.7	57.1	o	153.7	32.7	121.0
	IFD (GPT2)	k	1.2	7.5	-6.3	q	1.4	6.6	-5.2	k	57.0	25.0	32.0	q	95.4	28.5	67.0
		v	2.3	13.2	-10.9	o	2.7	11.8	-9.1	v	74.9	28.3	46.6	o	121.0	33.2	87.7
	InsTag	k	1.2	3.0	-1.8	q	1.4	2.7	-1.3	k	67.7	22.9	44.8	q	124.6	27.9	96.7
		v	2.1	5.4	-3.3	o	2.6	5.0	-2.4	v	89.0	27.4	61.7	o	167.0	31.7	135.3
	Reward	k	0.9	6.5	-5.6	q	1.1	5.7	-4.6	k	67.8	28.2	39.6	q	122.5	35.2	87.2
		v	1.7	12.0	-10.3	o	2.2	10.7	-8.5	v	90.4	32.5	57.9	o	163.5	40.7	122.8

Table 14: Qwen2.5-3B - OpenHermes 2.5 - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
OpenHermes 2.5	Difficulty	k - v	2.5e-04	-1.2e-04	q - o	2.2e-08	-6.8e-06	k	-1.7e-03	-1.5e-03	q	-2.0e-04	-9.2e-05
		v	2.5e-04	-1.2e-04	o	2.2e-08	-6.8e-06	v	9.3e-05	-2.7e-04	o	8.0e-05	1.5e-03
	IFD (GPT2)	k - v	-9.5e-05	-5.5e-04	q - o	2.9e-06	-6.5e-06	k	-1.3e-03	2.7e-05	q	-2.3e-04	-5.9e-04
		v	-9.5e-05	-5.5e-04	o	2.9e-06	-6.5e-06	v	-2.0e-04	-9.0e-04	o	1.6e-03	2.1e-03
InsTag	k - v	-7.0e-04	8.5e-05	q - o	-4.5e-06	-3.6e-06	k	-1.7e-03	-1.5e-03	q	1.6e-04	2.3e-04	
	v	-7.0e-04	8.5e-05	o	-4.5e-06	-3.6e-06	v	-1.1e-04	-4.9e-04	o	9.8e-04	1.7e-03	
Reward	k - v	-2.2e-04	2.8e-05	q - o	5.4e-06	-4.2e-07	k	-1.7e-03	-1.5e-03	q	-2.1e-04	-4.3e-04	
	v	-2.2e-04	2.8e-05	o	5.4e-06	-4.2e-07	v	5.4e-04	-1.6e-04	o	-3.6e-05	1.2e-03	

Table 15: Qwen2.5-3B - OpenHermes 2.5 - Similarity-based Metrics

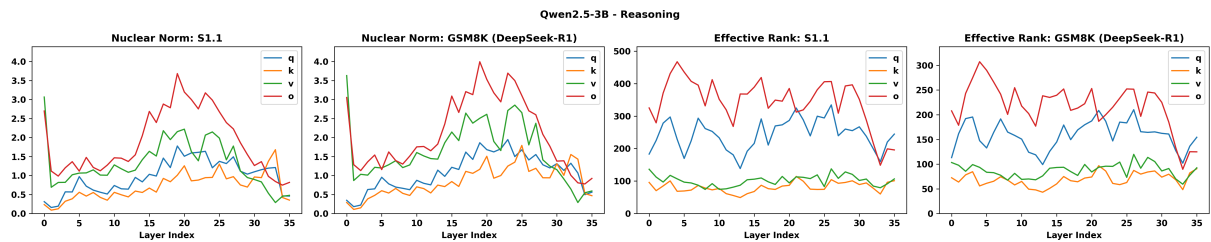


Figure 19: Qwen2.5 3B - Reasoning Data

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Magpie	Difficulty	k	1.1	1.2	-0.0	q	1.3	1.3	-0.0	k	63.7	59.6	4.1	q	122.4	101.8	20.6
		v	2.1	2.2	-0.1	o	2.5	2.6	-0.1	v	82.3	78.5	3.9	o	165.4	131.6	33.9
	IFD (GPT2)	k	1.1	0.9	0.1	q	1.3	1.1	0.3	k	68.0	63.9	4.1	q	131.9	111.7	20.2
		v	2.1	1.7	0.3	o	2.6	2.0	0.6	v	90.0	81.0	9.1	o	175.3	153.0	22.3
	InsTag	k	1.2	1.1	0.0	q	1.4	1.3	0.1	k	70.6	59.8	10.9	q	140.0	106.6	33.4
		v	2.1	2.1	0.0	o	2.6	2.5	0.1	v	91.3	78.5	12.9	o	192.1	139.8	52.3
	Reward	k	1.1	1.1	-0.0	q	1.3	1.2	0.1	k	67.3	55.2	12.1	q	133.2	104.0	29.3
		v	2.0	2.0	-0.0	o	2.5	2.3	0.2	v	90.5	69.6	21.0	o	181.8	140.0	41.7

Table 16: Qwen2.5-3B - Magpie - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Magpie	Difficulty	k - v	-1.8e-04	-4.3e-04	q - o	-1.1e-05	4.3e-06	k	-6.9e-04	-1.6e-03	q	-8.0e-04	-5.8e-04
		v	-1.3e-04	1.5e-04	o	5.2e-04	1.7e-03						
	IFD (GPT2)	k - v	-2.8e-04	-7.9e-05	q - o	-7.6e-06	1.5e-06	k	-2.1e-03	-1.7e-03	q	-2.7e-04	-4.9e-04
		v	-4.4e-04	3.8e-05	o	1.3e-03	-9.8e-05						
InsTag	k - v	-3.9e-04	-2.8e-04	q - o	1.1e-06	-1.0e-06	k	-1.8e-03	-1.6e-03	q	-5.0e-04	-5.7e-04	
	v	1.7e-04	2.2e-04	o	5.0e-04	1.1e-03							
Reward	k - v	-1.3e-03	3.4e-05	q - o	-2.1e-06	-1.0e-05	k	-1.2e-03	1.8e-04	q	-3.5e-04	-1.7e-03	
	v	4.8e-05	1.7e-03	o	1.8e-03	1.5e-03							

Table 17: Qwen2.5-3B - Magpie - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Reasoning	Reasoning	k	0.7	0.9	-0.1	q	1.0	1.1	-0.1	k	81.4	69.3	12.1	q	243.4	158.6	84.8
		v	1.3	1.6	-0.3	o	1.9	2.1	-0.2	v	101.2	87.5	13.8	o	344.6	217.6	127.0

Table 18: Qwen2.5-3B - Reasoning - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Reasoning	Reasoning	k - v	-3.0e-04	1.7e-04	q - o	8.3e-07	5.4e-06	k	-2.7e-03	-3.1e-03	q	-6.9e-04	5.4e-05
		v	5.8e-04	-2.3e-03	o	-3.0e-03	-1.9e-03						

Table 19: Qwen2.5-3B - Reasoning - Similarity-based Metrics

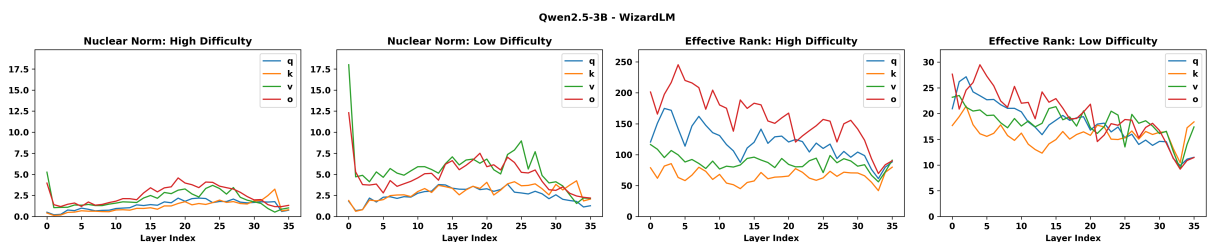


Figure 20: Qwen2.5 3B - WizardLM with Difficulty Metric

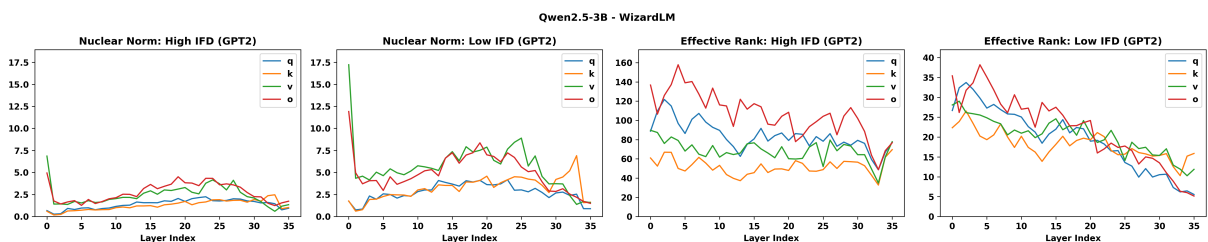


Figure 21: Qwen2.5 3B - WizardLM with IFD (GPT-2) Metric

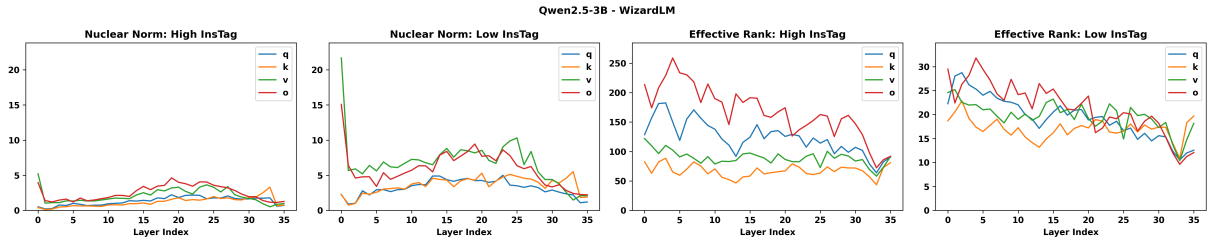


Figure 22: Qwen2.5 3B - WizardLM with InsTag Metric

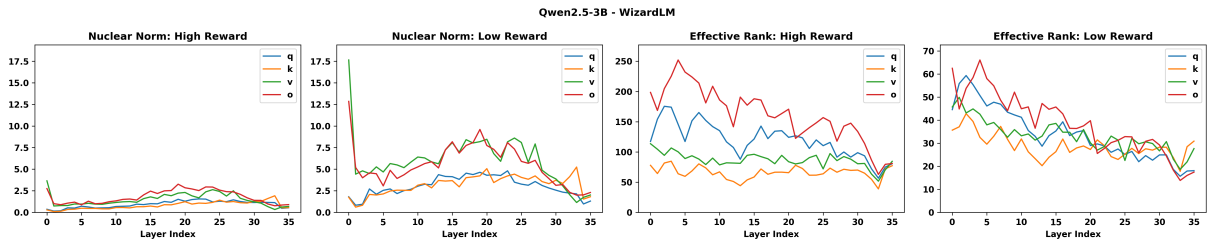


Figure 23: Qwen2.5 3B - WizardLM with Reward Model Metric

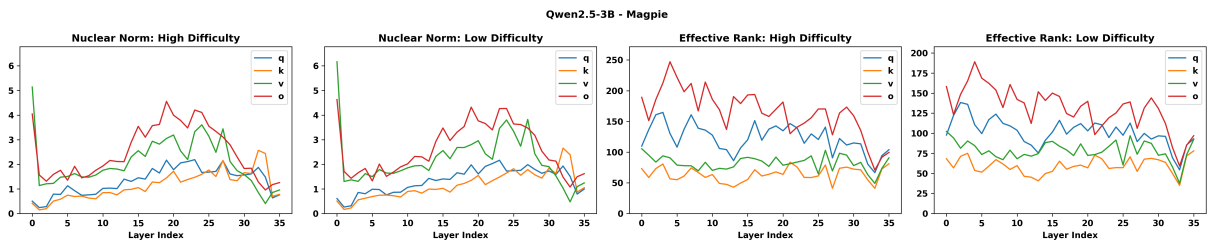


Figure 24: Qwen2.5 3B - Magpie with Difficulty Metric

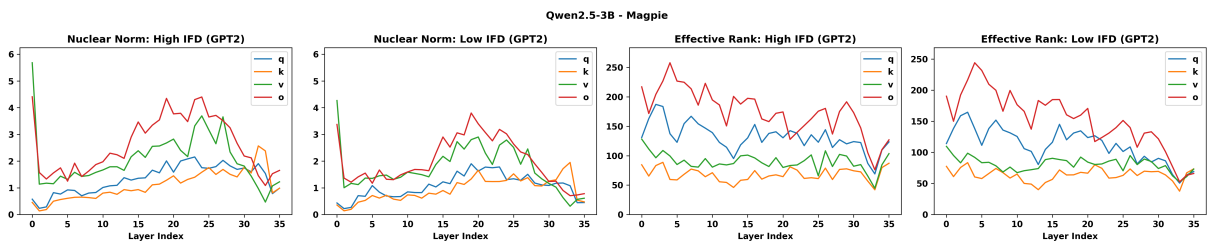


Figure 25: Qwen2.5 3B - Magpie with IFD (GPT-2) Metric

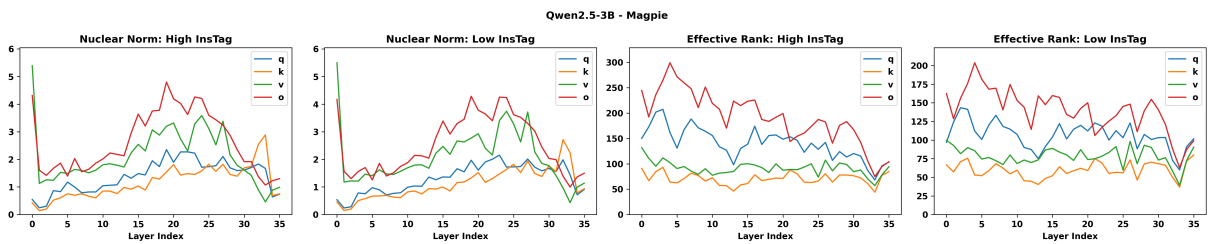


Figure 26: Qwen2.5 3B - Magpie with InsTag Metric

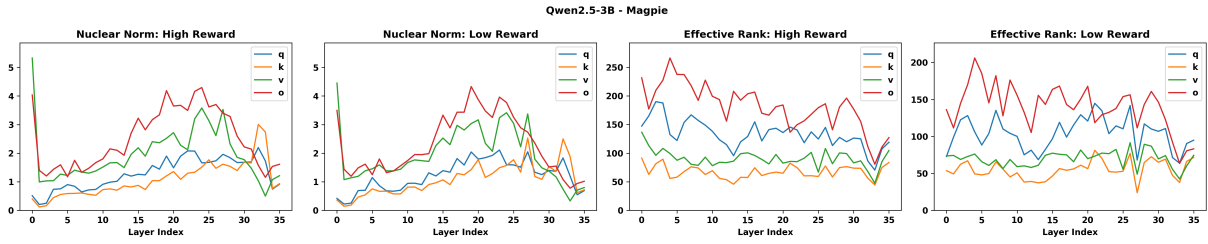


Figure 27: Qwen2.5 3B - Magpie with Reward Model Metric

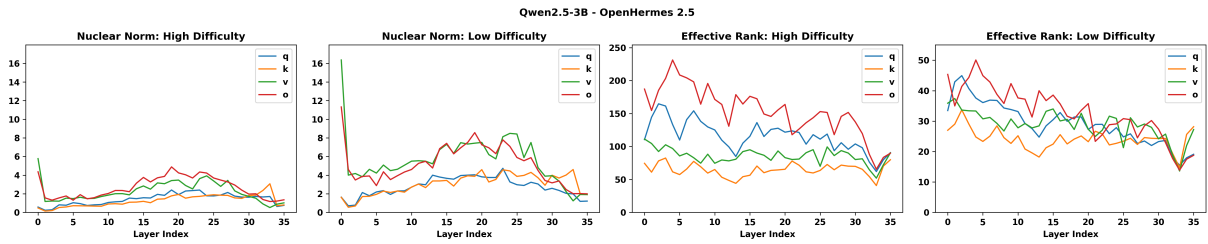


Figure 28: Qwen2.5 3B - OpenHermes with Difficulty Metric

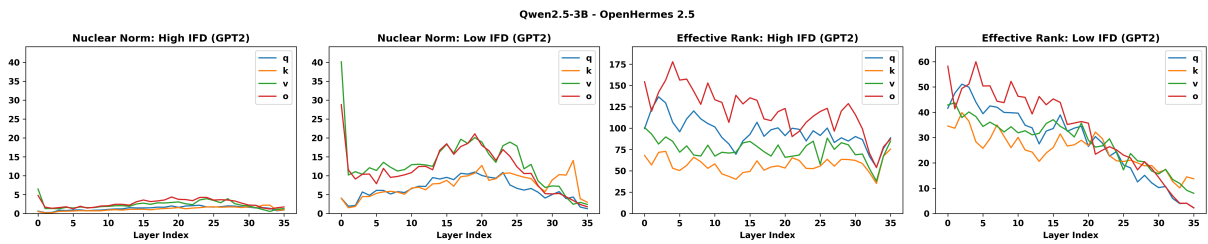


Figure 29: Qwen2.5 3B - OpenHermes with IFD (GPT-2) Metric

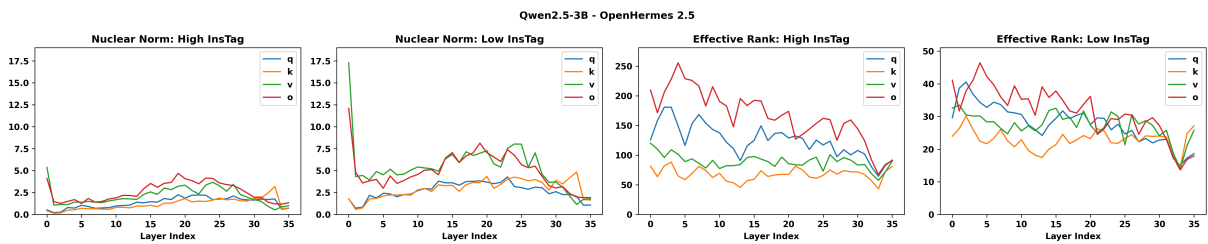


Figure 30: Qwen2.5 3B - OpenHermes with InsTag Metric

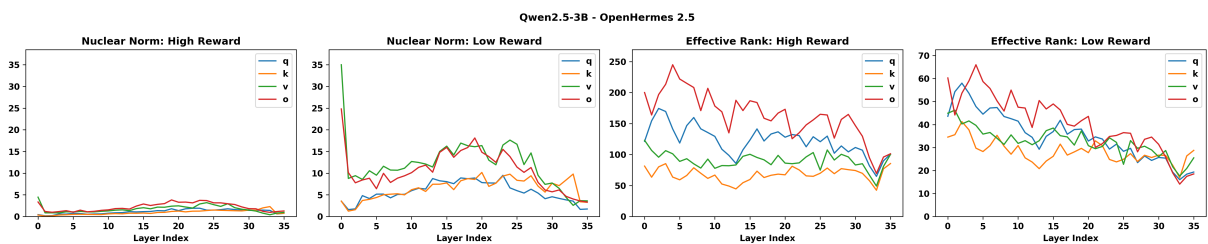


Figure 31: Qwen2.5 3B - OpenHermes with Reward Model Metric

C.3 Qwen2.5 7B

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
WizardLM	Difficulty	k	1.9	3.5	-1.6	q	1.9	2.8	-0.9	k	91.6	18.6	73.0	q	133.2	20.1	113.2
		v	3.2	6.9	-3.7	o	3.4	5.2	-1.9	v	114.9	19.2	95.7	o	167.9	20.7	147.2
	IFD (GPT2)	k	1.9	4.4	-2.5	q	1.9	3.3	-1.4	k	68.6	22.0	46.6	q	97.8	21.7	76.1
		v	3.4	7.6	-4.2	o	3.5	5.9	-2.5	v	84.2	22.2	62.0	o	116.4	23.1	93.3
	InsTag	k	1.9	4.1	-2.2	q	1.9	3.3	-1.4	k	95.6	20.0	75.6	q	141.2	21.6	119.6
		v	3.2	7.6	-4.5	o	3.4	6.0	-2.6	v	120.7	20.8	99.9	o	180.6	22.4	158.2
	Reward	k	1.3	4.3	-3.1	q	1.3	3.7	-2.4	k	91.5	35.6	55.9	q	131.5	38.6	92.9
		v	2.2	7.9	-5.7	o	2.3	6.5	-4.2	v	113.3	36.9	76.4	o	166.9	41.5	125.4

Table 20: Qwen2.5-7B - WizardLM - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
WizardLM	Difficulty	k - v	-7.1e-04	-1.4e-04	q - o	-3.9e-06	7.7e-06	k	-5.0e-04	-2.4e-04	q	-8.7e-05	-2.3e-04
		v			o			v	-3.2e-04	2.4e-04	o	1.1e-03	-7.8e-05
	IFD (GPT2)	k - v	-5.1e-04	-1.8e-04	q - o	-4.3e-06	-5.9e-06	k	-1.3e-03	-6.0e-04	q	-9.1e-06	2.5e-04
		v			o			v	1.0e-04	-2.3e-05	o	7.7e-04	-7.3e-04
	InsTag	k - v	-1.1e-03	-5.0e-05	q - o	8.8e-07	1.5e-06	k	-5.2e-04	-7.3e-04	q	-4.4e-05	1.6e-04
		v			o			v	1.3e-04	4.6e-04	o	1.2e-03	-1.6e-04
	Reward	k - v	-6.2e-04	-3.8e-04	q - o	2.3e-06	1.3e-06	k	-5.8e-04	-7.5e-04	q	-1.3e-04	1.9e-04
		v			o			v	1.7e-04	9.3e-05	o	5.5e-04	-4.5e-04

Table 21: Qwen2.5-7B - WizardLM - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
OpenHermes 2.5	Difficulty	k	2.0	4.0	-2.0	q	2.0	3.3	-1.2	k	92.1	29.0	63.1	q	135.7	33.3	102.4
		v	3.4	7.2	-3.8	o	3.6	5.9	-2.3	v	115.7	31.1	84.6	o	170.8	35.1	135.7
	IFD (GPT2)	k	1.8	8.5	-6.7	q	1.9	6.5	-4.7	k	78.5	32.7	45.8	q	114.0	33.7	80.3
		v	3.3	14.8	-11.5	o	3.5	11.8	-8.3	v	96.6	32.5	64.0	o	136.6	36.5	100.1
	InsTag	k	1.8	4.1	-2.3	q	1.9	3.3	-1.5	k	98.4	27.4	70.9	q	145.9	31.5	114.4
		v	3.1	7.2	-4.1	o	3.3	6.0	-2.7	v	123.1	29.7	93.5	o	183.8	33.7	150.1
	Reward	k	1.5	9.5	-8.1	q	1.6	7.4	-5.8	k	98.0	34.4	63.5	q	143.9	38.5	105.4
		v	2.6	17.2	-14.6	o	2.9	13.6	-10.6	v	123.4	36.3	87.1	o	182.6	41.9	140.7

Table 22: Qwen2.5-7B - OpenHermes 2.5 - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
OpenHermes 2.5	Difficulty	k - v	-8.8e-04	-2.7e-04	q - o	2.0e-06	-1.6e-06	k	-6.2e-04	-5.6e-04	q	-8.3e-05	7.2e-05
		v			o			v	-2.3e-04	2.6e-04	o	7.1e-04	-1.0e-04
	IFD (GPT2)	k - v	-4.0e-04	-3.0e-04	q - o	-2.4e-06	7.5e-06	k	-5.3e-04	-5.9e-04	q	-1.5e-04	-1.8e-05
		v			o			v	-2.9e-04	4.4e-04	o	3.6e-04	-1.6e-03
	InsTag	k - v	-1.1e-03	-3.1e-04	q - o	-2.7e-06	-3.0e-06	k	-7.9e-04	-3.7e-04	q	-3.2e-05	2.1e-04
		v			o			v	1.5e-05	3.0e-04	o	1.1e-03	-6.0e-04
	Reward	k - v	-5.4e-04	7.9e-05	q - o	-2.1e-06	2.9e-06	k	-5.9e-04	-3.3e-04	q	-2.3e-04	-1.2e-04
		v			o			v	2.2e-04	-5.2e-05	o	-3.6e-05	-9.4e-05

Table 23: Qwen2.5-7B - OpenHermes 2.5 - Similarity-based Metrics

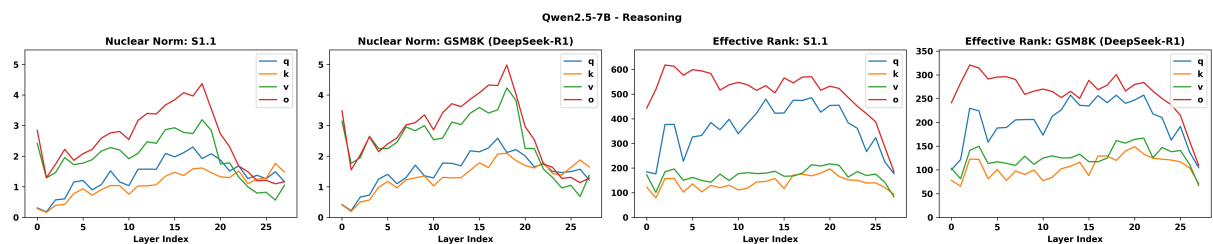


Figure 32: Qwen2.5 7B - Reasoning Data

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Magpie	Difficulty	k	1.8	1.8	0.0	q	1.9	1.9	0.0	k	95.9	83.9	12.1	q	153.3	124.4	28.9
		v	3.1	3.1	-0.1	o	3.4	3.4	-0.0	v	118.1	102.8	15.3	o	195.2	151.8	43.4
	IFD (GPT2)	k	1.8	1.6	0.2	q	1.9	1.6	0.3	k	100.6	93.5	7.1	q	163.7	133.3	30.4
		v	3.1	2.8	0.3	o	3.5	2.8	0.7	v	125.8	110.9	14.9	o	207.2	170.2	37.0
	InsTag	k	2.0	1.7	0.3	q	2.2	1.8	0.3	k	108.4	86.0	22.4	q	174.1	132.0	42.1
		v	3.5	3.0	0.4	o	3.8	3.3	0.5	v	133.8	105.7	28.1	o	225.0	163.3	61.7
	Reward	k	1.8	1.8	0.0	q	1.9	1.8	0.1	k	99.4	83.9	15.5	q	163.5	131.2	32.3
		v	3.0	3.0	-0.0	o	3.4	3.2	0.2	v	126.6	101.1	25.5	o	212.0	164.1	47.9

Table 24: Qwen2.5-7B - Magpie - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Magpie	Difficulty	k - v	-7.2e-04	-6.9e-04	q - o	-1.6e-06	7.9e-06	k	-1.2e-04	-7.1e-05	q	1.0e-04	-1.0e-04
		v			o			v	3.1e-04	-1.0e-04	o	4.1e-04	2.5e-04
	IFD (GPT2)	k - v	-5.9e-04	-8.9e-04	q - o	4.6e-06	6.9e-06	k	-7.9e-06	-7.9e-04	q	-2.7e-04	9.7e-05
		v			o			v	-1.1e-04	-1.8e-04	o	8.6e-04	1.1e-03
	InsTag	k - v	-1.1e-03	-8.3e-04	q - o	4.1e-06	2.1e-06	k	-3.3e-04	-1.6e-04	q	-9.4e-05	1.2e-05
		v			o			v	5.3e-04	-2.7e-04	o	1.5e-03	2.7e-04
	Reward	k - v	-1.5e-03	-1.2e-03	q - o	4.1e-06	3.9e-06	k	-6.9e-04	-2.8e-04	q	-1.8e-04	2.0e-05
		v			o			v	-2.9e-04	6.5e-04	o	1.3e-03	-1.6e-04

Table 25: Qwen2.5-7B - Magpie - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Reasoning	Reasoning	k	1.1	1.3	-0.3	q	1.4	1.5	-0.2	k	138.8	106.2	32.7	q	361.3	203.4	157.9
		v	1.9	2.5	-0.6	o	2.5	2.8	-0.3	v	170.5	126.7	43.8	o	509.9	263.1	246.8

Table 26: Qwen2.5-7B - Reasoning - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Reasoning	Reasoning	k - v	-8.9e-04	8.6e-05	q - o	-2.4e-06	-1.1e-05	k	-2.4e-03	-2.7e-03	q	4.4e-05	1.7e-04
		v			o			v	9.7e-04	2.2e-03	o	1.6e-03	2.5e-03

Table 27: Qwen2.5-7B - Reasoning - Similarity-based Metrics

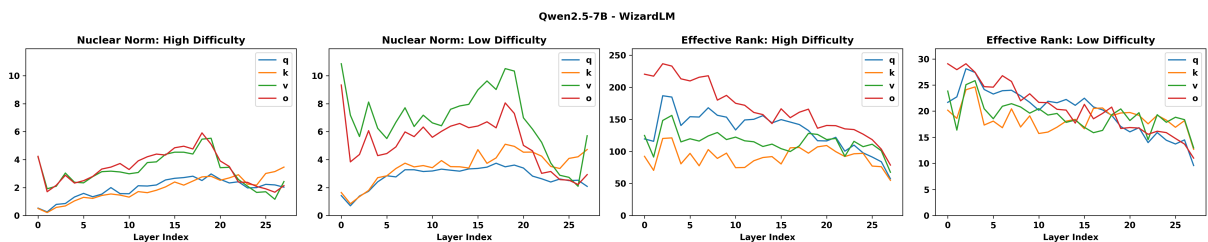


Figure 33: Qwen2.5 7B - WizardLM with Difficulty Metric

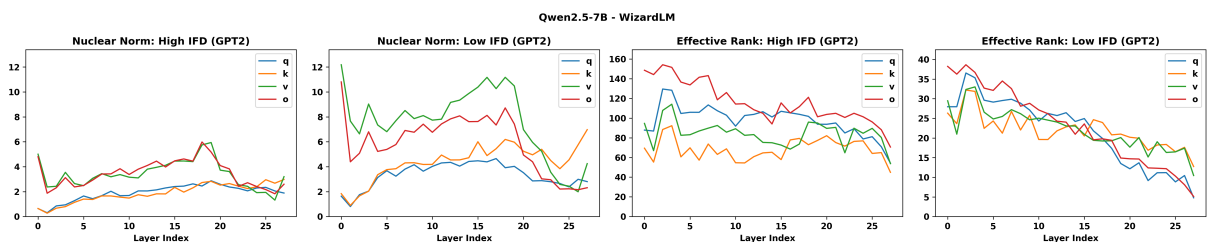


Figure 34: Qwen2.5 7B - WizardLM with IFD (GPT-2) Metric

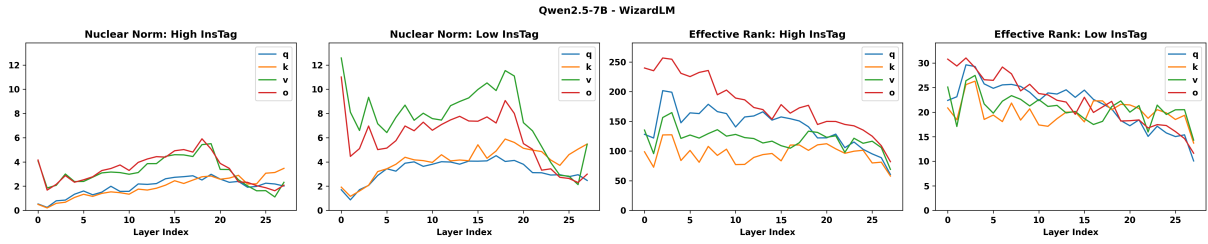


Figure 35: Qwen2.5 7B - WizardLM with InsTag Metric

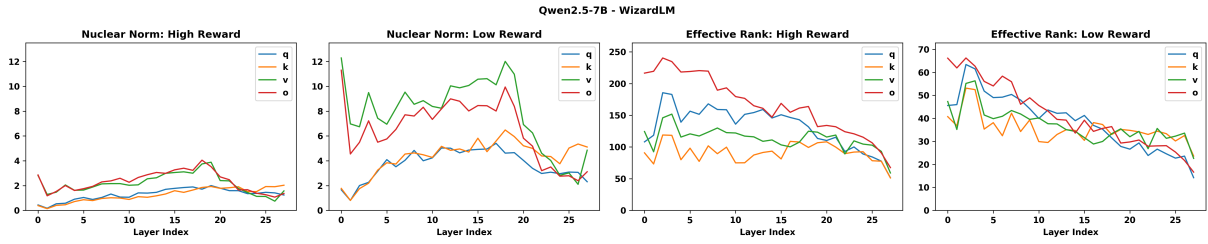


Figure 36: Qwen2.5 7B - WizardLM with Reward Model Metric

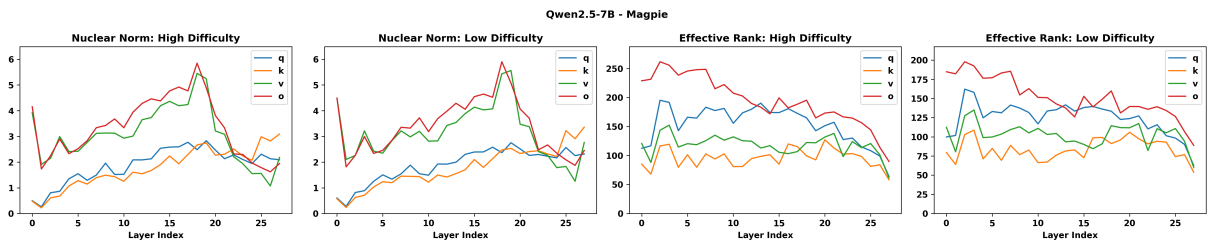


Figure 37: Qwen2.5 7B - Magpie with Difficulty Metric

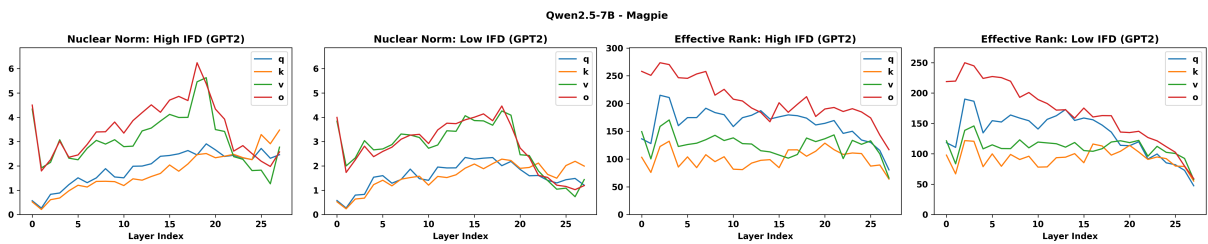


Figure 38: Qwen2.5 7B - Magpie with IFD (GPT-2) Metric

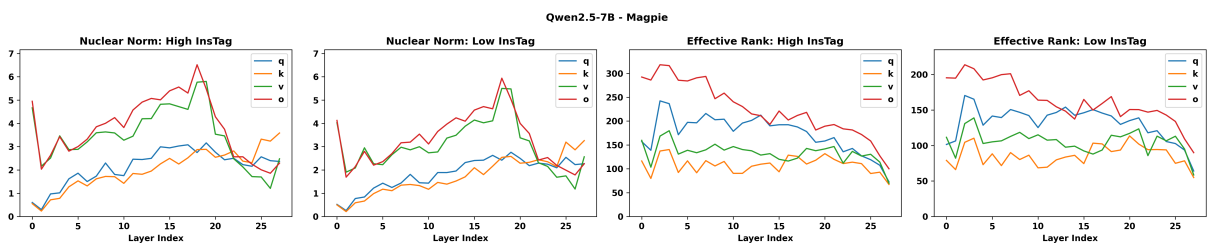


Figure 39: Qwen2.5 7B - Magpie with InsTag Metric

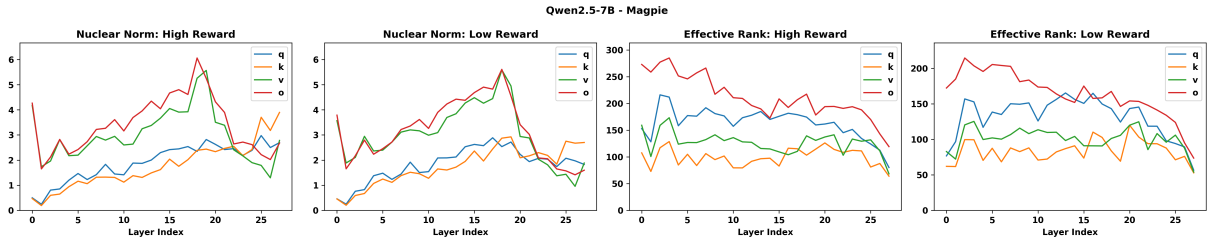


Figure 40: Qwen2.5 7B - Magpie with Reward Model Metric

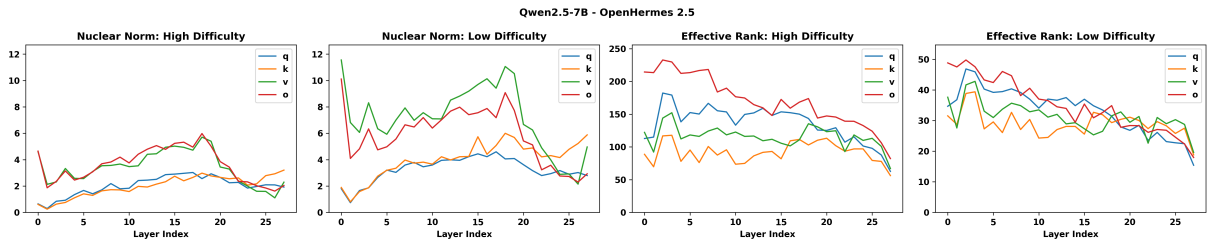


Figure 41: Qwen2.5 7B - OpenHermes with Difficulty Metric

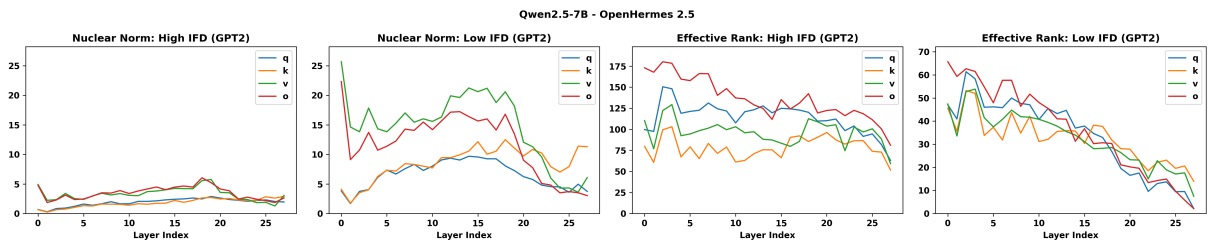


Figure 42: Qwen2.5 7B - OpenHermes with IFD (GPT-2) Metric

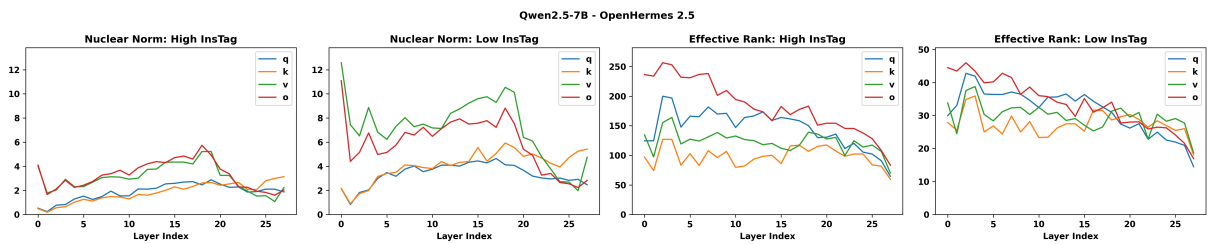


Figure 43: Qwen2.5 7B - OpenHermes with InsTag Metric

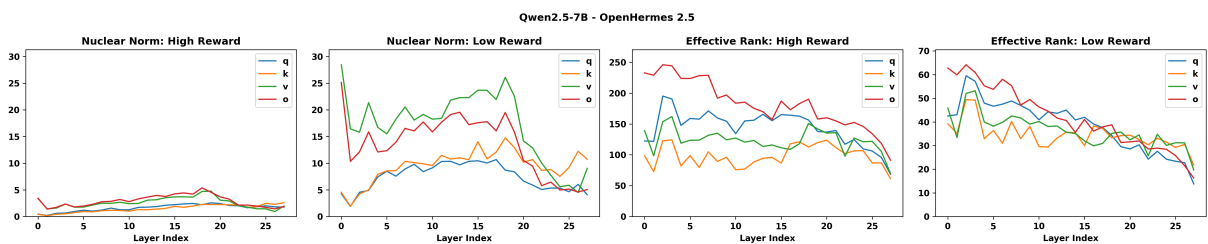


Figure 44: Qwen2.5 7B - OpenHermes with Reward Model Metric

C.4 Llama 3.1 8B

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
WizardLM	Difficulty	k	1.3	2.8	-1.5	q	2.5	4.9	-2.4	k	105.3	19.6	85.6	q	109.2	17.4	91.9
		v	6.5	15.5	-9.0	o	6.3	11.4	-5.0	v	112.1	17.6	94.5	o	151.8	19.1	132.6
	IFD (GPT2)	k	1.5	3.7	-2.3	q	2.9	6.1	-3.2	k	75.0	23.3	51.7	q	78.0	18.3	59.6
		v	7.7	20.0	-12.3	o	7.2	14.5	-7.3	v	78.0	20.3	57.7	o	102.3	20.5	81.8
	InsTag	k	1.4	3.3	-1.9	q	2.6	5.7	-3.1	k	107.8	21.3	86.5	q	111.7	18.8	92.8
		v	7.0	17.8	-10.8	o	6.7	13.3	-6.6	v	115.7	19.0	96.6	o	156.7	20.5	136.2
	Reward	k	1.2	3.2	-2.0	q	2.2	5.7	-3.5	k	100.7	37.6	63.1	q	104.2	32.2	72.0
		v	6.3	17.5	-11.2	o	5.7	13.7	-8.0	v	105.0	32.8	72.2	o	143.5	36.5	107.1

Table 28: Llama-3.1-8B - WizardLM - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
WizardLM	Difficulty	k - v	-2.7e-04	5.0e-06	q - o	-2.9e-06	7.0e-07	k	-2.5e-04	7.3e-04	q	-1.6e-04	-1.3e-05
		v			o			v	4.5e-04	2.8e-04	o	3.8e-05	9.6e-04
	IFD (GPT2)	k - v	5.0e-05	-2.6e-04	q - o	1.3e-06	2.3e-06	k	3.8e-04	2.9e-04	q	-3.2e-05	1.5e-04
		v			o			v	2.9e-04	-9.1e-05	o	9.4e-05	1.0e-03
InsTag	k - v	-2.0e-04	3.3e-04	q - o	-3.8e-06	1.1e-06	k	-2.9e-04	4.4e-04	q	-4.0e-04	-4.7e-05	
	v			o			v	6.2e-04	-3.3e-04	o	8.2e-05	1.2e-03	
Reward	k - v	-2.0e-05	-1.1e-04	q - o	1.1e-06	-1.8e-07	k	-1.0e-03	2.2e-04	q	-3.0e-04	-2.4e-04	
	v			o			v	4.9e-04	-5.0e-04	o	2.1e-04	4.0e-04	

Table 29: Llama-3.1-8B - WizardLM - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
OpenHermes 2.5	Difficulty	k	1.5	3.5	-1.9	q	2.9	5.8	-3.0	k	102.0	29.9	72.1	q	106.6	27.5	79.1
		v	7.9	18.9	-11.0	o	7.4	14.3	-7.0	v	107.0	27.4	79.6	o	143.8	31.0	112.8
	IFD (GPT2)	k	1.4	8.8	-7.4	q	2.8	13.5	-10.7	k	85.2	29.1	56.1	q	89.5	20.8	68.7
		v	7.4	48.9	-41.4	o	7.2	33.8	-26.6	v	89.3	24.8	64.5	o	117.3	24.2	93.1
	InsTag	k	1.4	2.8	-1.4	q	2.7	4.9	-2.2	k	109.1	31.1	78.0	q	113.1	28.7	84.4
		v	7.3	15.4	-8.1	o	7.0	11.8	-4.8	v	114.6	28.8	85.8	o	153.9	33.4	120.5
	Reward	k	1.1	5.8	-4.7	q	2.2	9.2	-7.0	k	109.3	37.7	71.6	q	114.6	33.2	81.4
		v	5.7	32.2	-26.5	o	5.8	23.5	-17.7	v	115.9	33.6	82.3	o	157.0	38.2	118.9

Table 30: Llama-3.1-8B - OpenHermes 2.5 - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
OpenHermes 2.5	Difficulty	k - v	-3.8e-04	4.2e-04	q - o	3.1e-06	1.4e-06	k	-8.7e-05	3.1e-04	q	-3.3e-04	-1.1e-04
		v			o			v	2.5e-04	4.1e-04	o	4.3e-04	5.9e-04
	IFD (GPT2)	k - v	4.2e-04	1.1e-04	q - o	-1.3e-07	-2.7e-06	k	2.6e-04	-5.0e-04	q	5.1e-05	-2.2e-04
		v			o			v	1.7e-05	1.8e-04	o	2.9e-04	2.0e-03
InsTag	k - v	4.7e-05	1.6e-04	q - o	1.8e-06	1.1e-06	k	-3.8e-04	3.3e-04	q	-1.7e-04	-6.4e-05	
	v			o			v	3.0e-04	-1.4e-04	o	4.3e-05	6.8e-04	
Reward	k - v	-6.5e-04	7.0e-05	q - o	1.2e-06	4.3e-06	k	-2.0e-04	8.0e-05	q	-2.4e-04	-4.8e-06	
	v			o			v	3.8e-04	3.1e-04	o	9.2e-04	5.6e-04	

Table 31: Llama-3.1-8B - OpenHermes 2.5 - Similarity-based Metrics

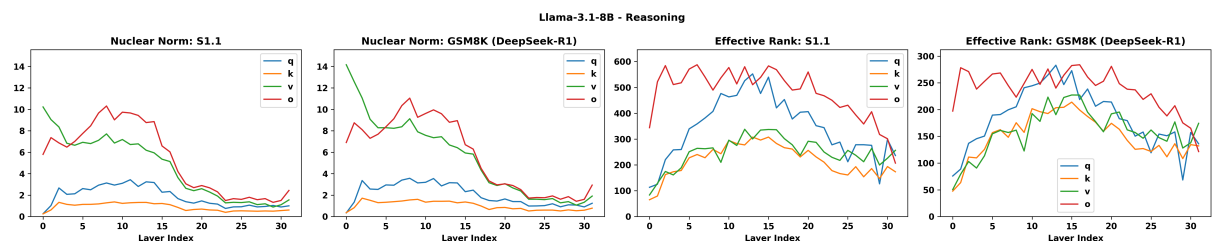


Figure 45: Llama 3.1 8B - Reasoning Data

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Magpie	Difficulty	k	1.3	1.2	0.0	q	2.4	2.4	0.0	k	123.6	96.7	26.9	q	132.5	100.7	31.9
		v	6.6	6.4	0.2	o	6.6	6.4	0.2	v	129.6	100.1	29.6	o	178.5	132.6	45.9
	IFD (GPT2)	k	1.2	1.1	0.1	q	2.3	2.0	0.4	k	119.5	111.5	8.0	q	128.7	108.6	20.1
		v	6.1	5.6	0.4	o	6.3	5.2	1.0	v	128.1	114.0	14.1	o	176.6	147.8	28.8
	InsTag	k	1.3	1.2	0.1	q	2.5	2.4	0.0	k	131.1	104.4	26.7	q	136.3	112.1	24.3
		v	6.6	6.4	0.2	o	6.5	6.4	0.1	v	138.7	109.0	29.7	o	190.2	147.6	42.5
	Reward	k	1.2	0.8	0.4	q	2.3	1.5	0.8	k	119.3	127.4	-8.1	q	130.7	149.4	-18.7
		v	6.0	4.1	1.9	o	6.2	4.1	2.2	v	130.4	135.6	-5.2	o	184.8	204.8	-20.1

Table 32: Llama-3.1-8B - Magpie - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Magpie	Difficulty	k - v	-2.2e-04	-2.2e-04	q - o	6.1e-06	1.5e-06	k	5.9e-04	5.0e-04	q	-3.3e-04	2.6e-05
		v	-3.6e-04	-5.9e-05	o	1.3e-03	5.5e-04						
	IFD (GPT2)	k - v	-4.9e-04	-2.8e-04	q - o	-2.1e-06	3.6e-06	k	2.7e-04	2.5e-04	q	3.1e-05	1.6e-05
		v	2.5e-04	2.6e-04	o	2.4e-04	1.0e-03						
	InsTag	k - v	-7.3e-04	-1.8e-04	q - o	-3.1e-06	2.0e-06	k	5.2e-04	4.6e-04	q	-2.4e-04	-8.9e-06
		v	2.3e-04	-5.4e-05	o	6.3e-04	1.0e-03						
	Reward	k - v	-5.5e-04	-9.9e-04	q - o	-2.8e-06	4.3e-06	k	4.3e-04	9.0e-04	q	-9.8e-06	-4.0e-04
		v	2.0e-04	-1.6e-04	o	2.7e-04	1.7e-03						

Table 33: Llama-3.1-8B - Magpie - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Reasoning	Reasoning	k	0.9	1.0	-0.2	q	1.8	2.0	-0.2	k	216.0	150.1	65.9	q	343.5	184.2	159.4
		v	4.6	5.5	-0.9	o	5.3	5.7	-0.4	v	248.4	159.6	88.8	o	480.5	238.5	241.9

Table 34: Llama-3.1-8B - Reasoning - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Reasoning	Reasoning	k - v	8.4e-05	-4.7e-05	q - o	-1.2e-05	1.8e-05	k	-3.8e-05	-2.1e-04	q	5.7e-04	2.8e-06
		v	-5.2e-05	-9.6e-04	o	2.0e-03	1.6e-04						

Table 35: Llama-3.1-8B - Reasoning - Similarity-based Metrics

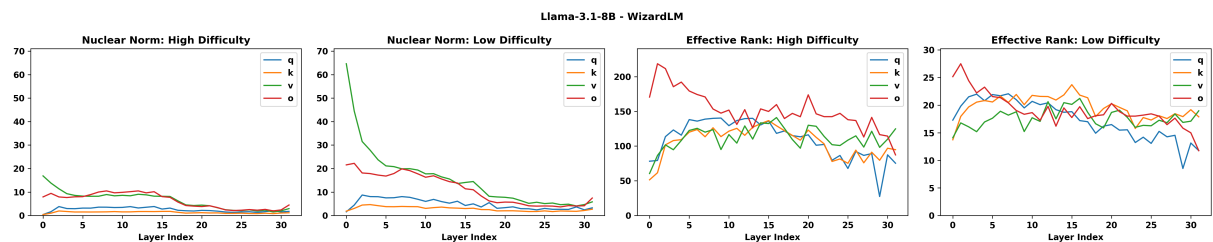


Figure 46: Llama 3.1 8B - WizardLM with Difficulty Metric

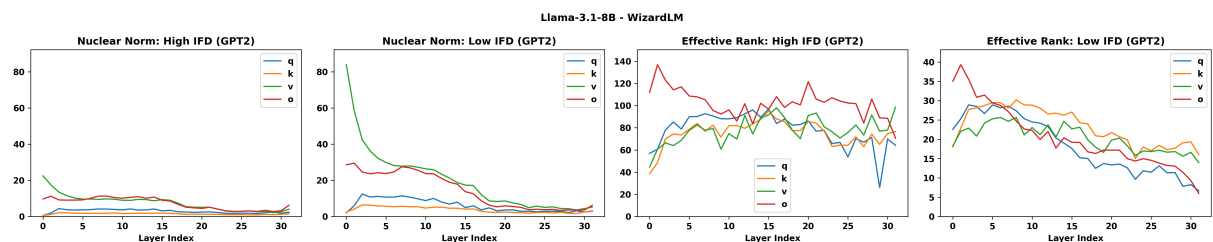


Figure 47: Llama 3.1 8B - WizardLM with IFD (GPT-2) Metric

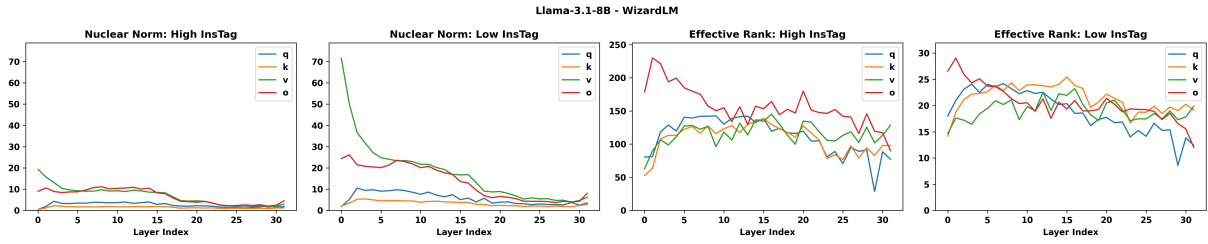


Figure 48: Llama 3.1 8B - WizardLM with InsTag Metric

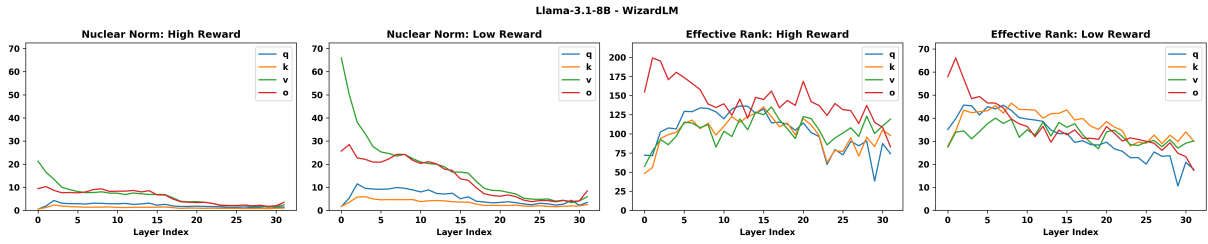


Figure 49: Llama 3.1 8B - WizardLM with Reward Model Metric

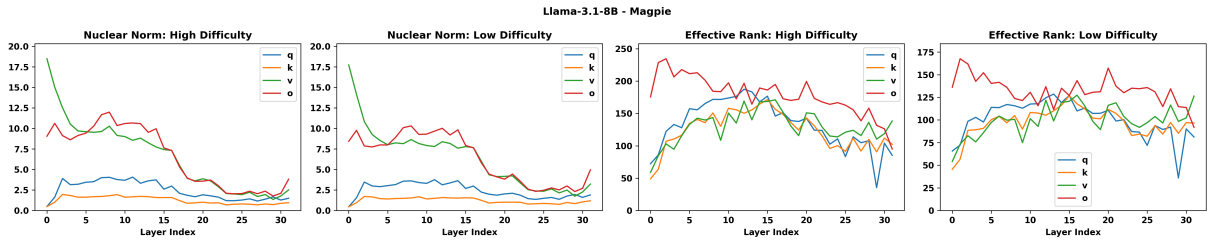


Figure 50: Llama 3.1 8B - Magpie with Difficulty Metric

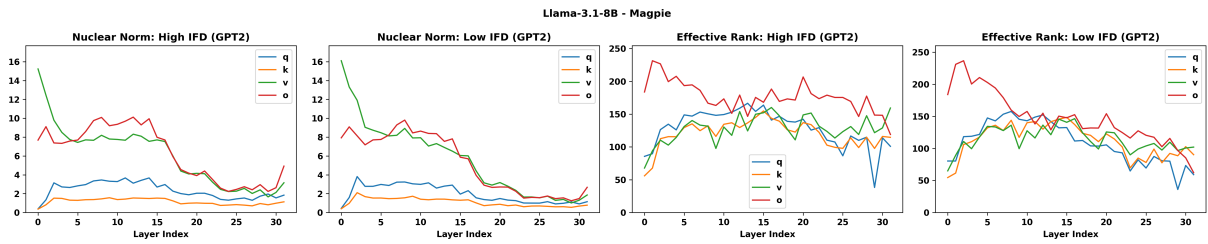


Figure 51: Llama 3.1 8B - Magpie with IFD (GPT-2) Metric

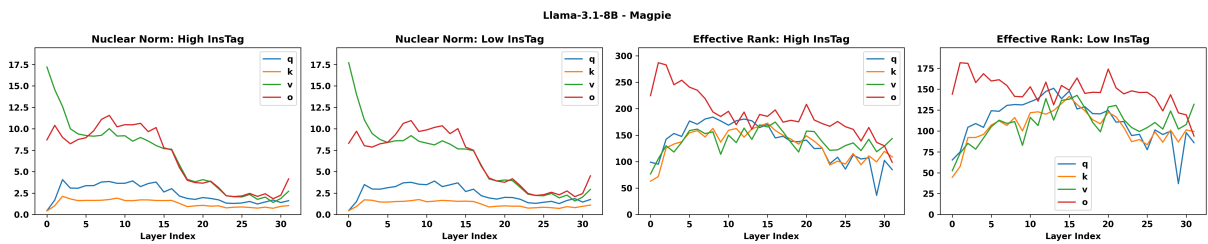


Figure 52: Llama 3.1 8B - Magpie with InsTag Metric

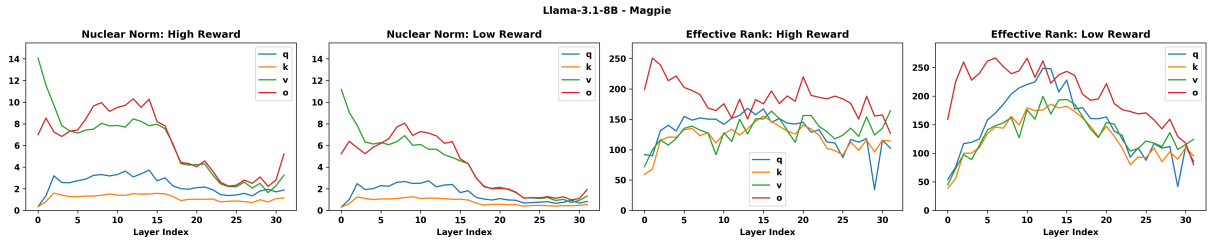


Figure 53: Llama 3.1 8B - Magpie with Reward Model Metric

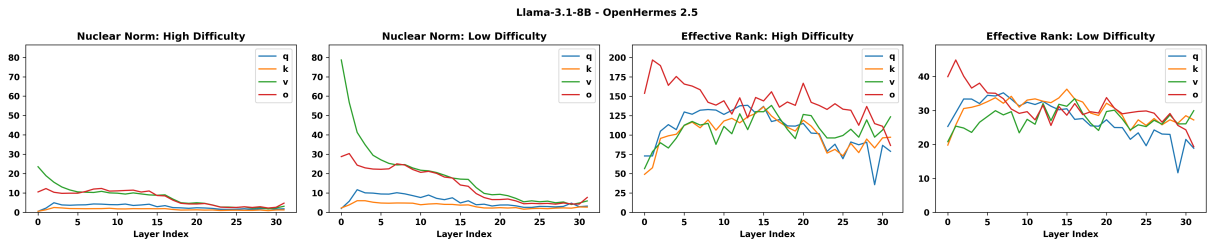


Figure 54: Llama 3.1 8B - OpenHermes with Difficulty Metric

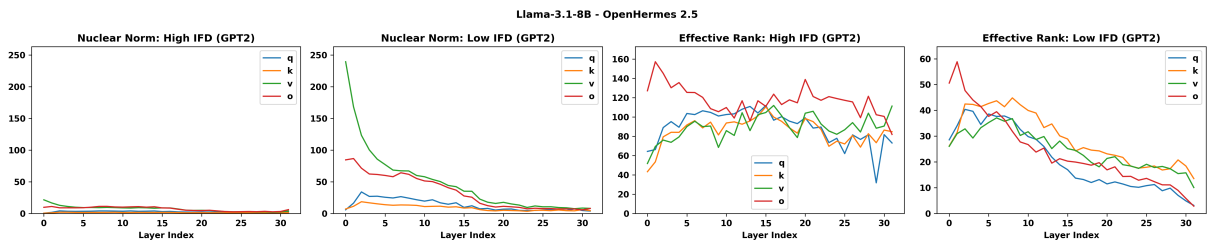


Figure 55: Llama 3.1 8B - OpenHermes with IFD (GPT-2) Metric

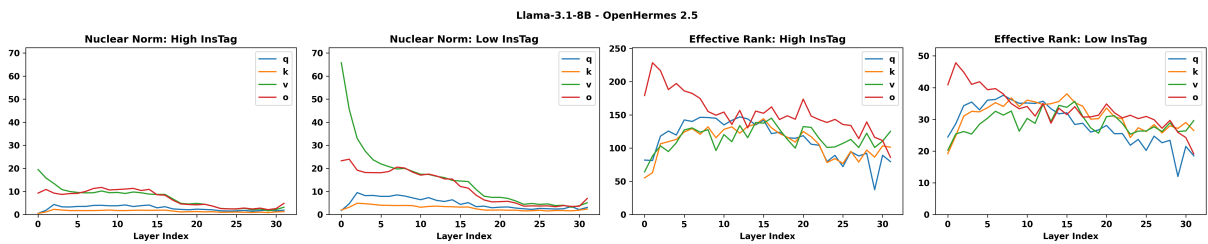


Figure 56: Llama 3.1 8B - OpenHermes with InsTag Metric

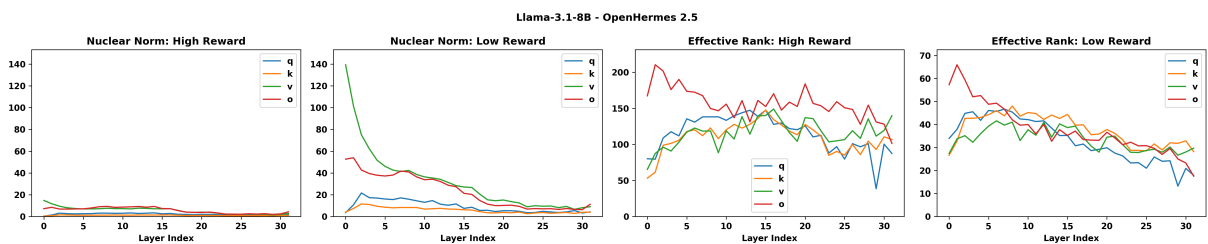


Figure 57: Llama 3.1 8B - OpenHermes with Reward Model Metric

C.5 Llama 3.2 1B

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
WizardLM	Difficulty	k	2.2	5.3	-3.1	q	4.1	8.4	-4.3	k	85.8	18.9	67.0	q	110.6	18.3	92.4
		v	8.9	22.0	-13.0	o	10.6	19.6	-9.1	v	110.6	18.7	91.9	o	157.4	20.1	137.3
	IFD (GPT2)	k	2.6	6.3	-3.7	q	4.8	9.2	-4.4	k	63.9	21.6	42.3	q	78.6	19.1	59.4
		v	10.8	23.9	-13.2	o	12.3	21.1	-8.9	v	79.5	21.0	58.4	o	105.5	21.6	83.9
	InsTag	k	2.2	6.2	-4.0	q	4.2	9.7	-5.5	k	88.3	20.5	67.8	q	114.0	20.0	94.0
		v	9.1	24.8	-15.8	o	10.7	22.3	-11.6	v	114.9	20.5	94.4	o	164.9	22.1	142.7
	Reward	k	1.7	5.4	-3.6	q	3.2	8.6	-5.3	k	86.9	33.8	53.0	q	110.0	32.8	77.2
		v	7.4	22.2	-14.9	o	8.6	20.7	-12.1	v	110.7	34.0	76.7	o	157.5	37.8	119.7

Table 36: Llama-3.2-1B - WizardLM - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
WizardLM	Difficulty	k - v	-4.9e-04	1.6e-04	q - o	9.9e-06	-2.5e-06	k	2.5e-03	2.0e-03	q	3.0e-04	3.0e-04
		v			o			v	-3.7e-04	1.5e-03	o	1.0e-03	-1.9e-03
	IFD (GPT2)	k - v	-1.1e-04	4.4e-04	q - o	1.2e-06	-9.2e-06	k	1.5e-03	1.8e-03	q	3.3e-04	1.6e-04
		v			o			v	8.9e-04	1.4e-03	o	-2.3e-05	-1.3e-03
InsTag	k - v	-9.1e-04	-7.3e-05	q - o	1.5e-05	-5.8e-06	k	2.7e-03	1.2e-03	q	3.9e-04	-7.3e-05	
	v			o			v	-1.1e-03	8.2e-04	o	2.7e-04	-1.3e-03	
Reward	k - v	-2.6e-04	-4.5e-04	q - o	1.1e-06	5.3e-06	k	1.3e-03	1.8e-03	q	-1.8e-04	1.1e-04	
	v			o			v	8.8e-04	-6.8e-04	o	2.3e-04	-5.2e-04	

Table 37: Llama-3.2-1B - WizardLM - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
OpenHermes 2.5	Difficulty	k	2.5	6.0	-3.5	q	4.5	9.1	-4.6	k	84.0	28.6	55.4	q	107.8	29.1	78.6
		v	10.3	23.9	-13.6	o	11.7	21.8	-10.1	v	107.3	29.8	77.4	o	150.1	33.3	116.8
	IFD (GPT2)	k	2.3	14.1	-11.8	q	4.4	19.3	-14.9	k	72.0	26.5	45.5	q	90.1	22.1	68.0
		v	10.0	53.7	-43.8	o	11.6	45.4	-33.8	v	90.6	26.3	64.3	o	121.4	26.3	95.1
	InsTag	k	2.3	5.1	-2.8	q	4.3	8.0	-3.7	k	89.0	28.7	60.3	q	114.0	29.7	84.3
		v	9.6	20.0	-10.5	o	11.2	18.9	-7.7	v	113.8	30.5	83.3	o	160.9	34.9	126.0
	Reward	k	1.7	9.5	-7.7	q	3.4	13.6	-10.2	k	90.2	32.9	57.3	q	116.0	33.0	83.1
		v	7.6	38.6	-31.0	o	9.3	33.7	-24.4	v	116.4	33.8	82.6	o	163.2	38.8	124.4

Table 38: Llama-3.2-1B - OpenHermes 2.5 - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
OpenHermes 2.5	Difficulty	k - v	-7.0e-04	-3.5e-05	q - o	-2.6e-06	1.1e-05	k	1.8e-03	1.9e-03	q	2.5e-04	1.1e-04
		v			o			v	3.8e-04	-2.3e-04	o	8.8e-04	-1.0e-03
	IFD (GPT2)	k - v	-6.6e-04	1.5e-03	q - o	-9.8e-06	3.4e-06	k	1.8e-03	2.0e-03	q	1.0e-04	-3.2e-04
		v			o			v	-9.4e-05	5.1e-04	o	-1.4e-04	2.0e-03
InsTag	k - v	-9.9e-04	-3.4e-04	q - o	-1.2e-05	-8.6e-07	k	2.1e-03	2.4e-03	q	-8.8e-05	1.8e-04	
	v			o			v	3.2e-04	-2.2e-04	o	4.3e-04	-1.1e-03	
Reward	k - v	-7.9e-04	-3.3e-04	q - o	-1.6e-05	2.8e-06	k	1.8e-03	2.1e-03	q	-3.6e-07	1.0e-04	
	v			o			v	-2.5e-04	-3.4e-05	o	8.4e-04	-1.8e-04	

Table 39: Llama-3.2-1B - OpenHermes 2.5 - Similarity-based Metrics

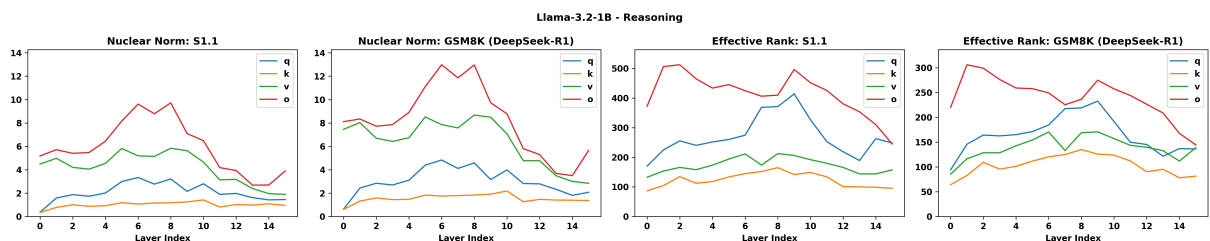


Figure 58: Llama 3.2 1B - Reasoning Data

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Magpie	Difficulty	k	1.9	2.0	-0.1	q	3.5	3.9	-0.3	k	91.6	80.2	11.4	q	123.7	100.3	23.4
		v	8.0	8.7	-0.7	o	9.5	10.4	-0.9	v	116.4	100.9	15.5	o	174.4	137.3	37.0
	IFD (GPT2)	k	1.9	1.7	0.2	q	3.8	3.0	0.8	k	94.2	87.9	6.4	q	126.6	107.3	19.3
		v	8.3	6.9	1.4	o	10.3	7.9	2.4	v	122.8	110.1	12.7	o	179.7	152.7	27.0
	InsTag	k	2.0	1.9	0.1	q	3.8	3.7	0.1	k	99.2	83.9	15.3	q	133.0	109.0	24.0
		v	8.4	8.2	0.1	o	10.1	10.0	0.1	v	128.6	106.5	22.1	o	192.5	150.8	41.7
	Reward	k	1.8	1.6	0.2	q	3.7	3.0	0.7	k	95.5	86.3	9.2	q	129.5	120.9	8.6
		v	7.8	6.6	1.2	o	10.0	7.9	2.0	v	126.6	110.9	15.7	o	188.9	174.1	14.8

Table 40: Llama-3.2-1B - Magpie - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Magpie	Difficulty	k - v	-5.1e-04	-8.4e-04	q - o	-9.5e-06	1.6e-06	k	1.6e-03	1.6e-03	q	2.5e-04	4.7e-04
		v	-6.2e-04	9.4e-05	o	1.0e-03	2.8e-04						
	IFD (GPT2)	k - v	-8.4e-04	-1.3e-03	q - o	-1.9e-06	-1.2e-05	k	1.9e-03	1.6e-03	q	4.9e-04	-1.6e-04
		v	5.6e-04	7.6e-04	o	7.9e-04	7.7e-04						
InsTag	k - v	-5.7e-04	-7.5e-04	q - o	-1.0e-05	-9.2e-06	k	1.6e-03	1.8e-03	q	6.9e-05	5.1e-04	
	v	4.0e-04	3.4e-04	o	6.6e-04	1.4e-03							
Reward	k - v	-9.7e-04	-7.8e-04	q - o	8.4e-06	1.3e-05	k	1.8e-03	1.4e-03	q	3.6e-04	4.5e-05	
	v	5.5e-04	-3.6e-04	o	2.8e-04	8.4e-04							

Table 41: Llama-3.2-1B - Magpie - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Reasoning	Reasoning	k	1.0	1.5	-0.5	q	2.1	3.0	-1.0	k	123.2	103.2	20.0	q	270.6	164.9	105.8
		v	4.2	6.4	-2.2	o	6.0	8.3	-2.3	v	172.7	138.9	33.8	o	414.9	240.8	174.1

Table 42: Llama-3.2-1B - Reasoning - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Reasoning	Reasoning	k - v	-1.2e-03	-3.9e-04	q - o	1.9e-05	7.5e-06	k	-1.0e-04	1.3e-03	q	-2.7e-04	7.8e-04
		v	-2.8e-03	-2.5e-03	o	-2.1e-03	1.2e-03						

Table 43: Llama-3.2-1B - Reasoning - Similarity-based Metrics

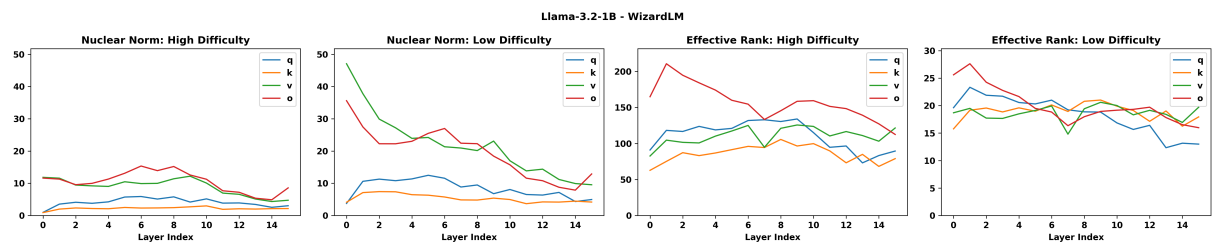


Figure 59: Llama 3.2 1B - WizardLM with Difficulty Metric

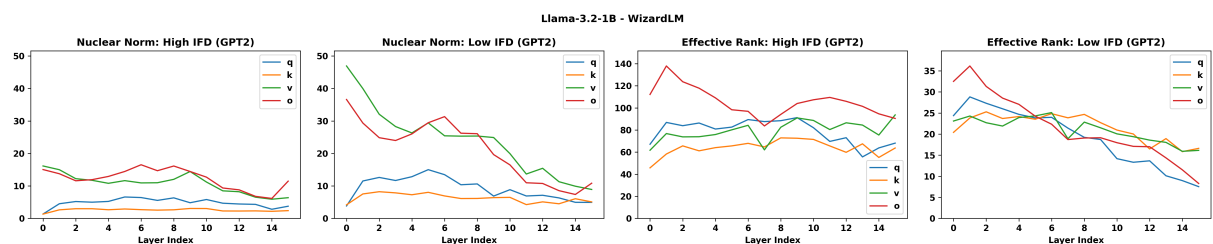


Figure 60: Llama 3.2 1B - WizardLM with IFD (GPT-2) Metric

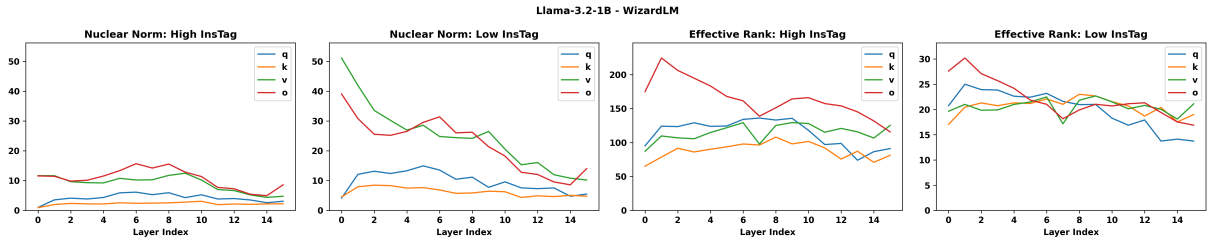


Figure 61: Llama 3.2 1B - WizardLM with InsTag Metric

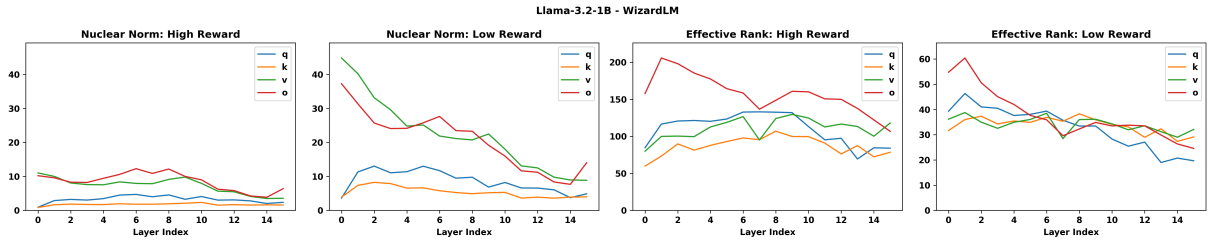


Figure 62: Llama 3.2 1B - WizardLM with Reward Model Metric

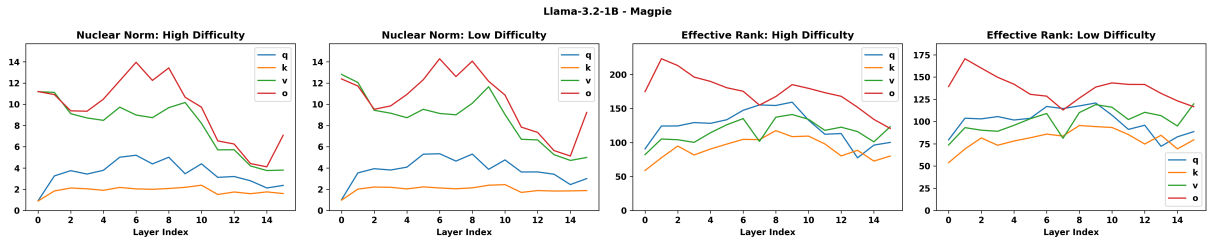


Figure 63: Llama 3.2 1B - Magpie with Difficulty Metric

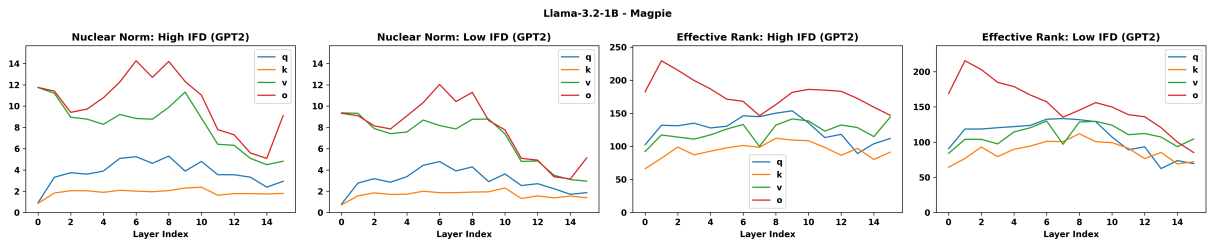


Figure 64: Llama 3.2 1B - Magpie with IFD (GPT-2) Metric

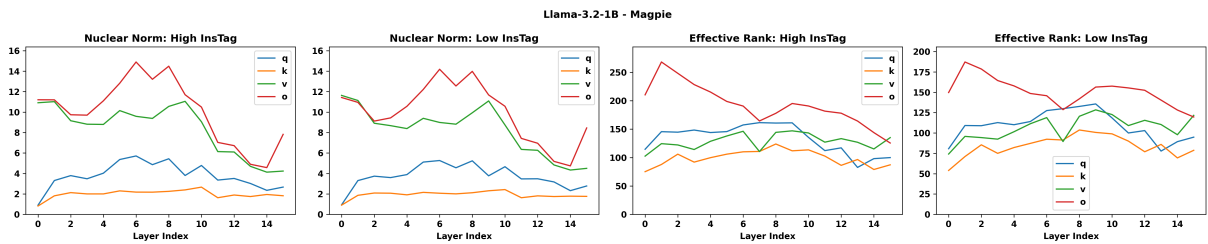


Figure 65: Llama 3.2 1B - Magpie with InsTag Metric

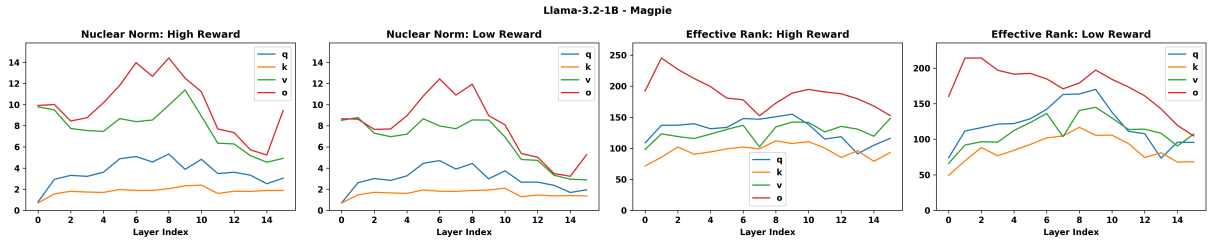


Figure 66: Llama 3.2 1B - Magpie with Reward Model Metric

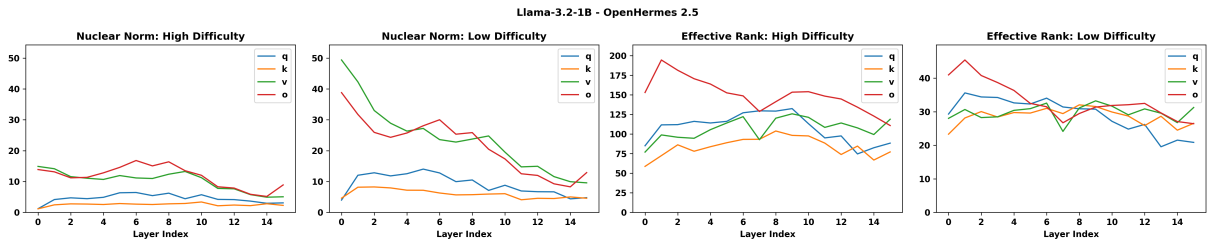


Figure 67: Llama 3.2 1B - OpenHerms with Difficulty Metric

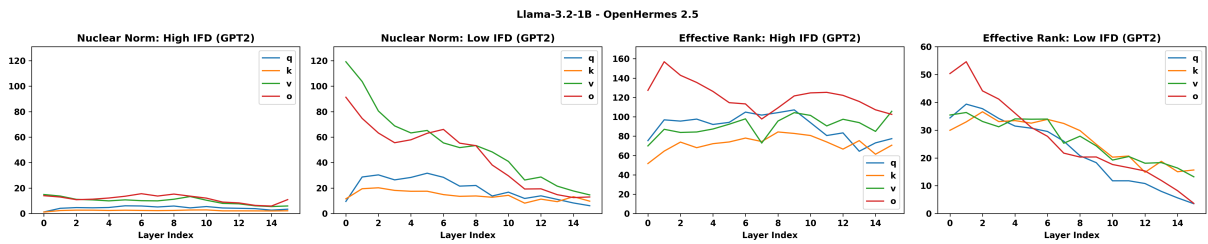


Figure 68: Llama 3.2 1B - OpenHerms with IFD (GPT-2) Metric

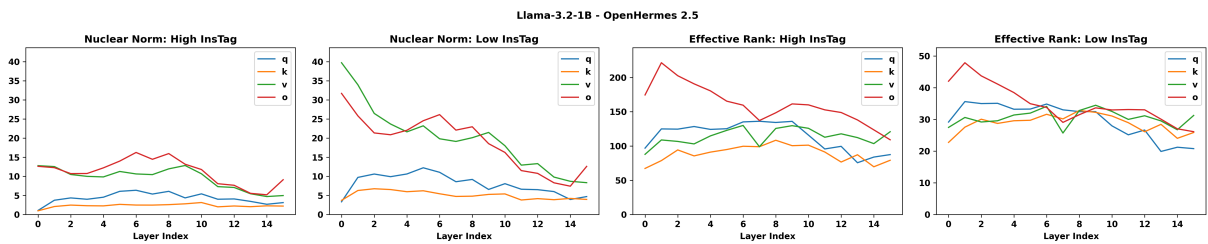


Figure 69: Llama 3.2 1B - OpenHerms with InsTag Metric

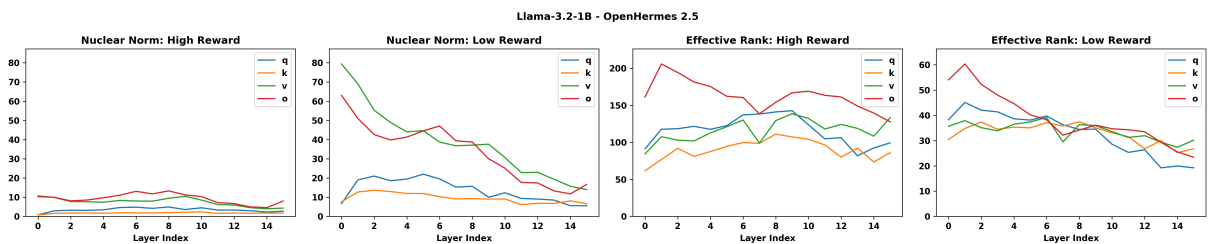


Figure 70: Llama 3.2 1B - OpenHerms with Reward Model Metric

C.6 Llama 3.2 3B

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
WizardLM	Difficulty	k	1.5	3.1	-1.6	q	2.3	4.4	-2.1	k	107.6	19.6	88.0	q	113.8	17.8	96.0
		v	5.5	12.0	-6.5	o	5.7	9.8	-4.2	v	124.5	18.7	105.8	o	159.1	19.9	139.2
	IFD (GPT2)	k	1.7	4.1	-2.5	q	2.6	5.4	-2.8	k	75.4	23.1	52.3	q	79.4	18.9	60.5
		v	6.3	15.2	-8.9	o	6.4	12.2	-5.9	v	85.4	21.3	64.1	o	106.2	21.3	84.9
	InsTag	k	1.5	3.6	-2.1	q	2.4	5.1	-2.7	k	110.5	21.4	89.1	q	117.0	19.4	97.7
		v	5.7	13.8	-8.2	o	5.8	11.4	-5.6	v	128.4	20.3	108.0	o	165.2	21.5	143.6
	Reward	k	1.3	3.4	-2.2	q	1.9	4.8	-2.9	k	106.6	37.0	69.6	q	111.2	32.9	78.3
		v	4.8	13.0	-8.2	o	4.8	11.2	-6.4	v	121.8	34.9	86.9	o	155.1	37.9	117.2

Table 44: Llama-3.2-3B - WizardLM - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
WizardLM	Difficulty	k - v	5.1e-04	-3.3e-04	q - o	-6.3e-06	-6.8e-06	k	-1.5e-03	-6.5e-04	q	-3.2e-04	-6.3e-04
		v	5.1e-04	-3.6e-04	o	5.6e-04	7.3e-04	v	5.1e-04	-3.6e-04	o	5.6e-04	7.3e-04
	IFD (GPT2)	k - v	4.7e-04	-5.0e-04	q - o	6.9e-06	-5.8e-06	k	-7.5e-04	-4.6e-04	q	-7.0e-05	-5.5e-04
		v	4.7e-04	-3.8e-05	o	8.8e-04	6.1e-04	v	-1.3e-04	-3.8e-05	o	8.8e-04	6.1e-04
InsTag	k - v	6.5e-04	-2.1e-04	q - o	-7.2e-06	1.5e-07	k	-1.7e-03	-3.1e-04	q	-3.6e-04	-3.3e-04	
	v	6.5e-04	-4.1e-04	o	5.0e-04	1.0e-03	v	5.2e-04	-4.1e-04	o	5.0e-04	1.0e-03	
Reward	k - v	4.1e-04	3.8e-05	q - o	3.1e-06	-3.8e-06	k	-1.0e-03	-4.1e-04	q	-4.7e-04	3.9e-05	
	v	4.1e-04	-3.6e-04	o	1.4e-04	6.5e-04	v	1.4e-04	-3.6e-04	o	1.4e-04	6.5e-04	

Table 45: Llama-3.2-3B - WizardLM - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
OpenHermes 2.5	Difficulty	k	1.8	3.8	-2.1	q	2.6	5.2	-2.6	k	104.5	30.1	74.4	q	110.8	28.2	82.6
		v	6.5	14.6	-8.1	o	6.5	12.1	-5.6	v	119.2	29.5	89.7	o	151.1	32.5	118.6
	IFD (GPT2)	k	1.6	9.8	-8.2	q	2.5	12.1	-9.6	k	86.1	29.0	57.0	q	91.2	21.9	69.3
		v	6.0	36.5	-30.5	o	6.2	28.6	-22.4	v	97.9	26.4	71.5	o	121.9	25.9	96.0
	InsTag	k	1.6	3.3	-1.6	q	2.5	4.6	-2.1	k	112.0	30.3	81.7	q	117.9	28.4	89.5
		v	6.0	12.2	-6.2	o	6.2	10.4	-4.2	v	128.0	30.0	98.0	o	162.8	33.3	129.4
	Reward	k	1.2	6.1	-4.9	q	1.9	7.8	-5.9	k	112.2	37.0	75.3	q	118.8	33.6	85.3
		v	4.6	23.4	-18.8	o	4.9	18.9	-13.9	v	129.8	35.5	94.3	o	164.5	39.3	125.2

Table 46: Llama-3.2-3B - OpenHermes 2.5 - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
OpenHermes 2.5	Difficulty	k - v	5.8e-04	3.5e-04	q - o	2.9e-06	-3.7e-06	k	-8.6e-04	-3.3e-04	q	-3.4e-04	-4.4e-04
		v	5.8e-04	-2.5e-04	o	1.1e-03	7.1e-04	v	1.4e-05	-2.5e-04	o	1.1e-03	7.1e-04
	IFD (GPT2)	k - v	5.2e-04	-4.3e-04	q - o	-3.7e-06	-4.0e-06	k	-4.7e-04	-5.5e-04	q	-3.2e-04	-2.1e-04
		v	5.2e-04	8.1e-05	o	7.6e-04	-6.6e-05	v	-1.2e-04	8.1e-05	o	7.6e-04	-6.6e-05
InsTag	k - v	5.4e-04	1.8e-04	q - o	5.5e-06	2.7e-06	k	-7.7e-04	-4.8e-04	q	-2.6e-05	-4.7e-04	
	v	5.4e-04	3.1e-04	o	5.0e-04	9.2e-04	v	2.0e-04	3.1e-04	o	5.0e-04	9.2e-04	
Reward	k - v	3.4e-04	8.5e-05	q - o	-1.9e-06	1.2e-06	k	-9.4e-04	-1.0e-04	q	-3.5e-04	-2.6e-04	
	v	3.4e-04	2.1e-05	o	1.5e-03	3.5e-04	v	2.6e-05	2.1e-05	o	1.5e-03	3.5e-04	

Table 47: Llama-3.2-3B - OpenHermes 2.5 - Similarity-based Metrics

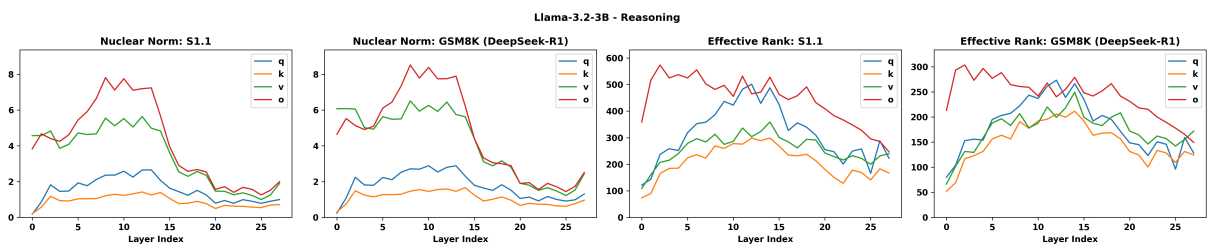


Figure 71: Llama 3.2 3B - Reasoning Data

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Magpie	Difficulty	k	1.4	1.4	0.0	q	2.1	2.2	-0.1	k	123.0	97.4	25.5	q	134.0	102.7	31.3
		v	5.2	5.2	-0.0	o	5.5	5.5	-0.0	v	139.7	110.0	29.7	o	181.8	138.1	43.7
	IFD (GPT2)	k	1.3	1.2	0.1	q	2.2	1.7	0.4	k	120.6	114.3	6.4	q	131.5	113.3	18.2
		v	5.1	4.3	0.8	o	5.5	4.4	1.2	v	140.7	127.1	13.5	o	182.5	156.5	25.9
	InsTag	k	1.4	1.3	0.1	q	2.2	2.2	0.1	k	132.5	105.1	27.4	q	140.4	113.7	26.6
		v	5.4	5.1	0.3	o	5.6	5.5	0.2	v	152.7	119.4	33.3	o	197.6	152.5	45.1
	Reward	k	1.3	1.2	0.0	q	2.1	1.9	0.2	k	122.0	118.7	3.3	q	134.4	134.7	-0.3
		v	4.8	4.6	0.2	o	5.3	4.9	0.4	v	144.6	134.3	10.3	o	191.4	179.3	12.0

Table 48: Llama-3.2-3B - Magpie - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Magpie	Difficulty	k - v	5.7e-04	7.9e-04	q - o	2.9e-06	-1.1e-05	k	-8.0e-04	-9.7e-04	q	-1.1e-04	-2.6e-04
		v	6.1e-04	2.9e-04	o	1.3e-03	1.4e-03	v	-6.1e-04	-2.9e-04	o	1.3e-03	1.4e-03
	IFD (GPT2)	k - v	4.8e-04	1.5e-04	q - o	-1.2e-06	1.4e-06	k	-1.1e-03	-5.9e-04	q	1.6e-04	-1.7e-05
		v	6.1e-06	-1.7e-04	o	1.0e-03	3.3e-04	v	6.1e-06	-1.7e-04	o	1.0e-03	3.3e-04
InsTag	k - v	6.1e-04	5.2e-04	q - o	8.1e-06	-3.6e-07	k	-1.1e-03	-8.4e-04	q	-2.0e-05	-2.1e-04	
	v	-1.2e-04	-6.7e-04	o	4.2e-04	1.5e-03	v	-1.2e-04	-6.7e-04	o	4.2e-04	1.5e-03	
Reward	k - v	1.1e-03	9.6e-05	q - o	-3.1e-06	1.3e-05	k	-1.0e-03	-6.3e-04	q	-1.9e-04	5.9e-04	
	v	3.4e-04	-1.0e-03	o	1.3e-03	1.3e-03	v	3.4e-04	-1.0e-03	o	1.3e-03	1.3e-03	

Table 49: Llama-3.2-3B - Magpie - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Reasoning	Reasoning	k	0.9	1.1	-0.2	q	1.5	1.8	-0.2	k	208.3	149.0	59.3	q	315.5	181.7	133.8
		v	3.5	4.2	-0.7	o	4.2	4.7	-0.5	v	258.3	173.7	84.6	o	445.5	243.2	202.3

Table 50: Llama-3.2-3B - Reasoning - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Reasoning	Reasoning	k - v	1.3e-04	4.0e-05	q - o	-2.6e-06	-2.5e-06	k	-6.8e-04	-1.2e-03	q	-3.2e-04	-2.9e-04
		v	1.6e-03	-1.1e-03	o	9.4e-04	-1.2e-03	v	-1.6e-03	-1.1e-03	o	9.4e-04	-1.2e-03

Table 51: Llama-3.2-3B - Reasoning - Similarity-based Metrics

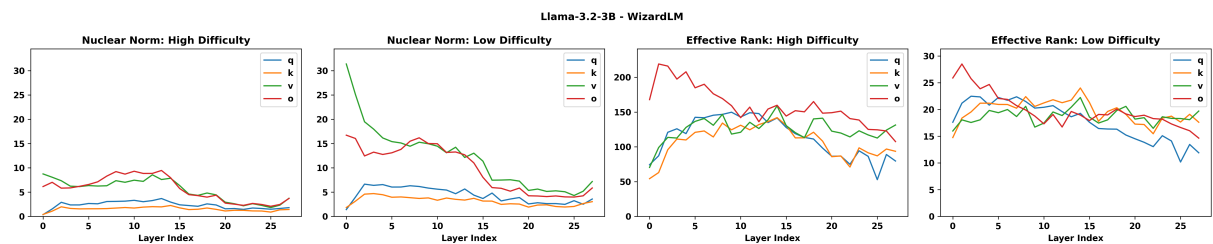


Figure 72: Llama 3.2 3B - WizardLM with Difficulty Metric

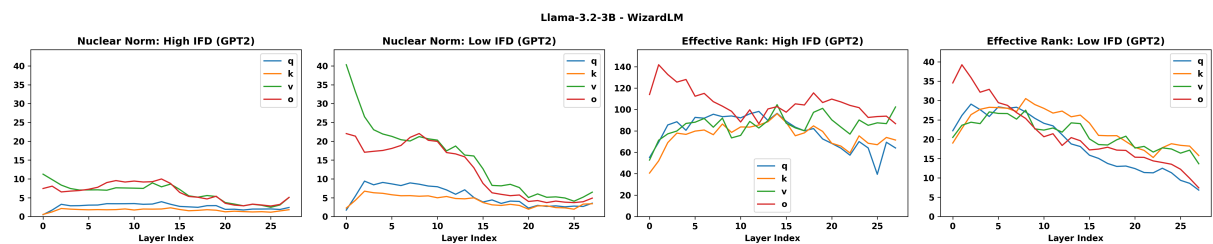


Figure 73: Llama 3.2 3B - WizardLM with IFD (GPT-2) Metric

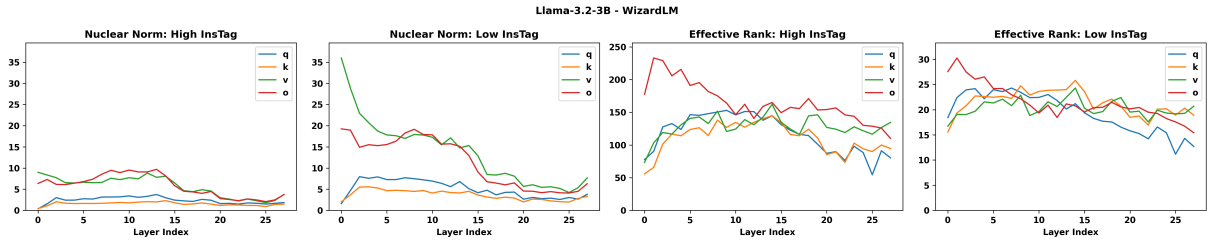


Figure 74: Llama 3.2 3B - WizardLM with InsTag Metric

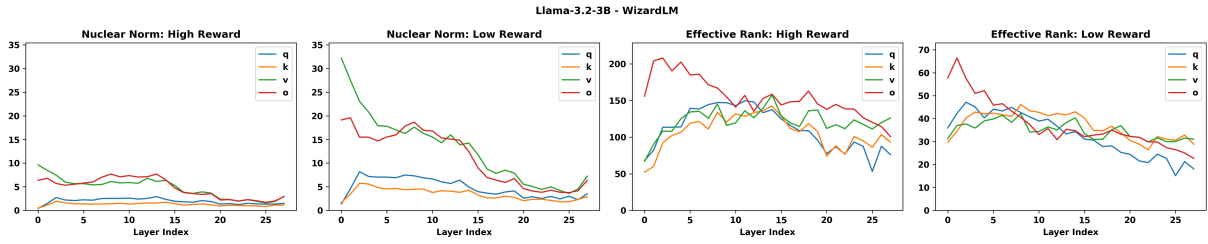


Figure 75: Llama 3.2 3B - WizardLM with Reward Model Metric

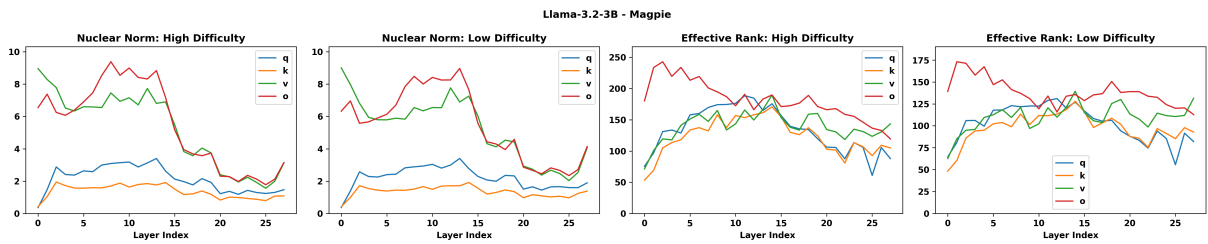


Figure 76: Llama 3.2 3B - Magpie with Difficulty Metric

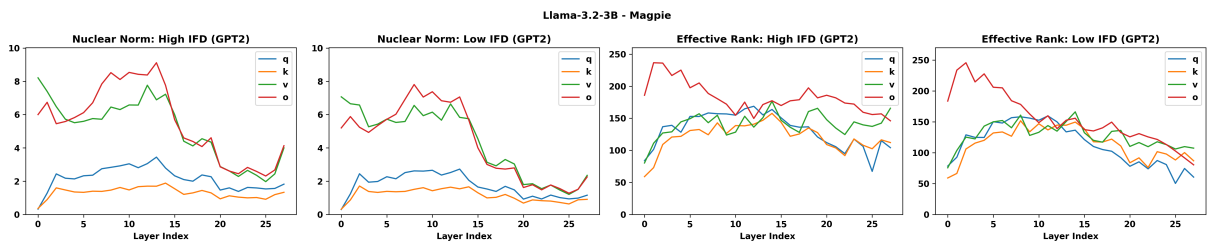


Figure 77: Llama 3.2 3B - Magpie with IFD (GPT-2) Metric

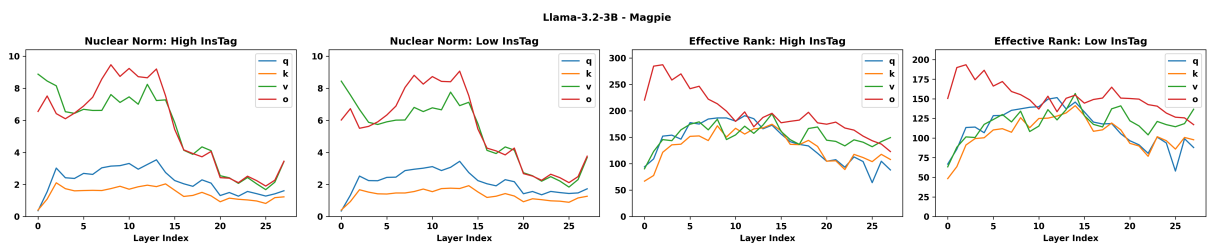


Figure 78: Llama 3.2 3B - Magpie with InsTag Metric

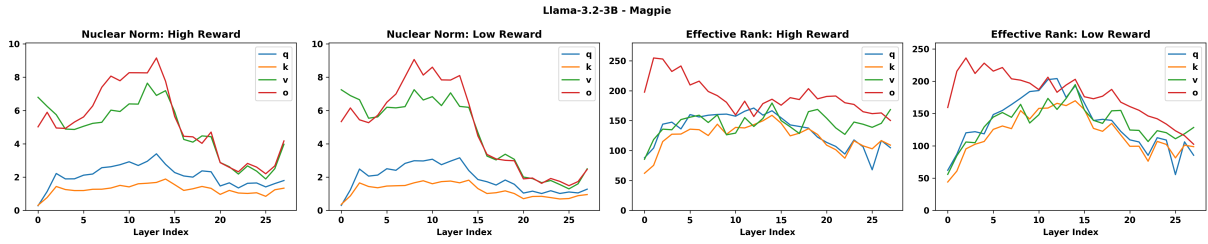


Figure 79: Llama 3.2 3B - Magpie with Reward Model Metric

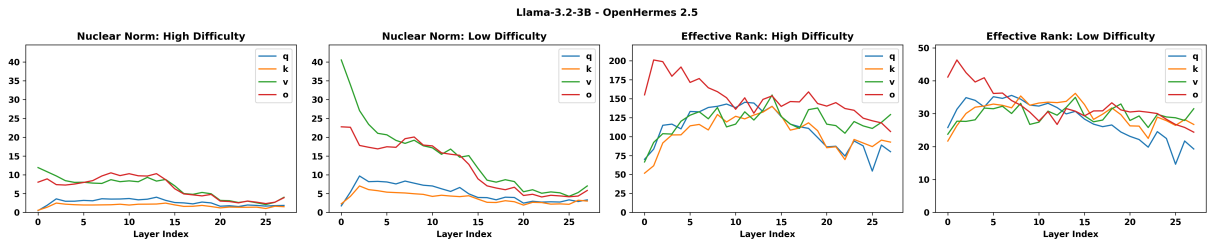


Figure 80: Llama 3.2 3B - OpenHermes with Difficulty Metric

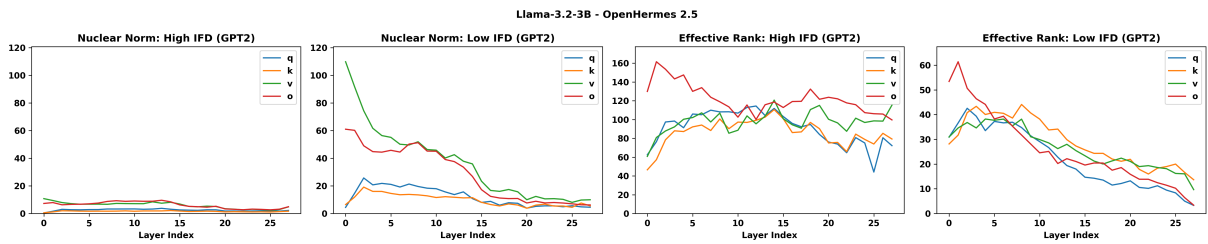


Figure 81: Llama 3.2 3B - OpenHermes with IFD (GPT-2) Metric

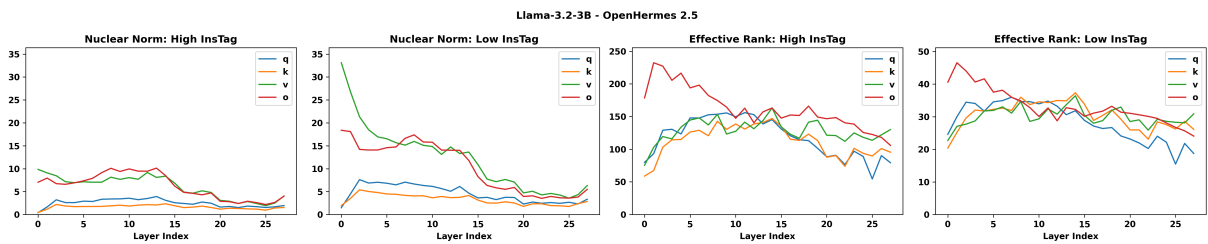


Figure 82: Llama 3.2 3B - OpenHermes with InsTag Metric

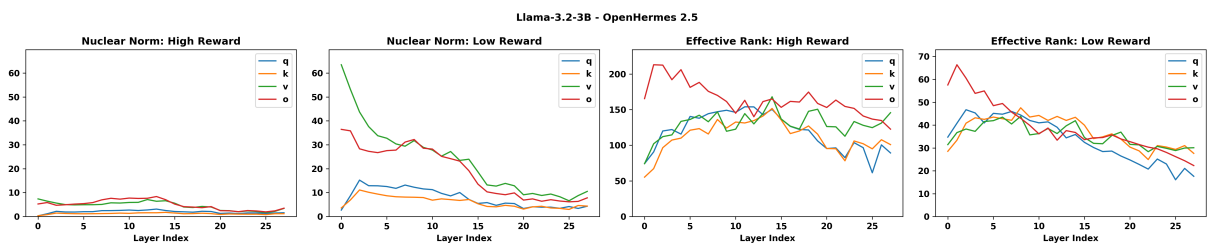


Figure 83: Llama 3.2 3B - OpenHermes with Reward Model Metric

C.7 Gemma 2 2B

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
WizardLM	Difficulty	k	3.6	10.4	-6.8	q	3.7	9.6	-5.8	k	105.4	19.2	86.1	q	103.4	16.1	87.3
		v	7.9	23.0	-15.1	o	7.6	17.6	-10.1	v	91.2	15.3	75.9	o	140.1	19.1	121.0
	IFD (GPT2)	k	4.0	13.8	-9.8	q	4.3	12.2	-7.9	k	74.3	23.0	51.3	q	72.5	17.9	54.6
		v	9.4	28.0	-18.6	o	8.7	21.8	-13.1	v	65.3	17.3	48.0	o	95.6	21.2	74.4
	InsTag	k	3.6	11.3	-7.7	q	3.7	10.4	-6.7	k	111.0	21.0	90.0	q	108.8	17.6	91.2
		v	7.8	24.3	-16.4	o	7.5	18.9	-11.4	v	96.2	16.7	79.4	o	147.5	21.0	126.5
	Reward	k	2.6	9.0	-6.4	q	2.7	8.5	-5.8	k	114.8	36.0	78.8	q	111.7	30.0	81.7
		v	5.7	20.2	-14.5	o	5.6	16.5	-10.8	v	99.6	26.8	72.8	o	149.1	35.4	113.7

Table 52: Gemma 2 2B - WizardLM - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
WizardLM	Difficulty	k - v	5.7e-04	3.0e-04	q - o	2.1e-07	2.9e-06	k	-1.2e-03	-8.6e-04	q	1.0e-04	-1.9e-04
		v			o			v	1.2e-04	2.0e-04	o	1.6e-03	8.9e-04
	IFD (GPT2)	k - v	5.2e-04	7.2e-05	q - o	-5.4e-06	7.4e-06	k	-5.0e-04	-8.9e-04	q	-5.4e-04	-5.7e-05
		v			o			v	-2.4e-04	3.0e-04	o	1.4e-04	-1.3e-04
InsTag	k - v	5.0e-04	3.8e-05	q - o	-1.2e-05	-3.6e-06	k	-1.2e-03	-8.3e-04	q	4.5e-05	-7.2e-05	
	v			o			v	3.9e-04	-6.5e-05	o	1.4e-03	6.1e-04	
Reward	k - v	1.7e-04	7.8e-05	q - o	-1.6e-06	2.2e-06	k	-1.4e-03	-7.8e-04	q	9.4e-05	-6.1e-05	
	v			o			v	-6.1e-04	1.6e-04	o	6.7e-05	1.1e-03	

Table 53: Gemma 2 2B - WizardLM - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
OpenHermes 2.5	Difficulty	k	4.1	11.9	-7.8	q	4.2	10.8	-6.6	k	104.7	30.0	74.7	q	102.6	26.1	76.4
		v	9.2	25.5	-16.3	o	8.5	20.3	-11.7	v	90.6	23.9	66.6	o	136.7	31.2	105.5
	IFD (GPT2)	k	3.7	28.5	-24.8	q	4.0	23.7	-19.7	k	86.5	27.2	59.2	q	84.8	19.4	65.5
		v	8.7	56.3	-47.6	o	8.3	43.6	-35.3	v	74.9	19.9	55.0	o	110.0	23.6	86.4
	InsTag	k	3.7	8.0	-4.3	q	3.8	7.6	-3.7	k	115.5	30.5	85.0	q	112.6	26.7	86.0
		v	8.2	17.5	-9.3	o	7.9	14.3	-6.4	v	98.0	24.6	73.5	o	148.6	32.4	116.2
	Reward	k	2.9	18.2	-15.3	q	3.1	16.0	-12.9	k	112.9	36.8	76.1	q	110.7	31.2	79.5
		v	6.6	41.6	-35.1	o	6.6	32.0	-25.4	v	100.0	27.6	72.4	o	148.3	37.0	111.4

Table 54: Gemma 2 2B - OpenHermes 2.5 - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
OpenHermes 2.5	Difficulty	k - v	3.2e-04	7.3e-05	q - o	9.7e-06	2.2e-06	k	-1.4e-03	-5.3e-05	q	-2.5e-04	3.6e-05
		v			o			v	3.0e-04	5.6e-05	o	1.0e-03	6.3e-05
	IFD (GPT2)	k - v	3.4e-04	-8.0e-05	q - o	2.6e-06	2.9e-06	k	-2.5e-05	-9.9e-04	q	-7.0e-05	4.2e-04
		v			o			v	-4.8e-04	5.4e-04	o	-2.6e-04	6.6e-04
InsTag	k - v	2.3e-04	3.9e-04	q - o	2.5e-06	-2.9e-07	k	-1.4e-03	-2.7e-04	q	1.5e-04	3.4e-04	
	v			o			v	7.8e-04	2.5e-04	o	5.5e-04	3.0e-04	
Reward	k - v	2.7e-04	4.5e-05	q - o	-3.9e-06	1.2e-05	k	-1.1e-03	-2.7e-04	q	-2.8e-04	-6.9e-05	
	v			o			v	-3.0e-04	1.8e-04	o	4.1e-04	3.8e-04	

Table 55: Gemma 2 2B - OpenHermes 2.5 - Similarity-based Metrics

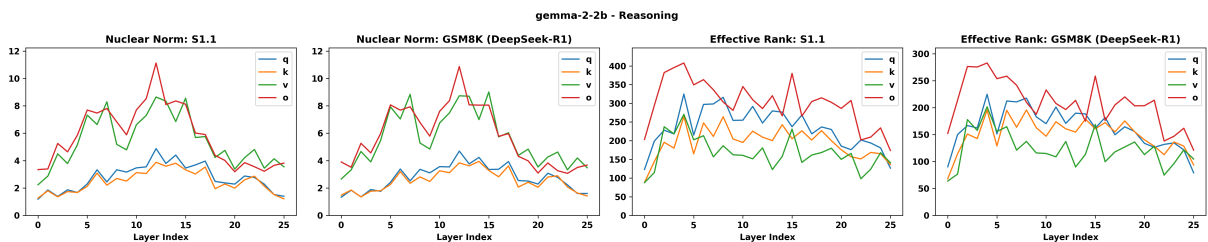


Figure 84: Gemma 2 2B - Reasoning Data

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Magpie	Difficulty	k	3.3	3.2	0.1	q	3.5	3.4	0.0	k	123.1	96.2	26.9	q	122.9	94.4	28.5
		v	7.3	7.3	-0.0	o	7.3	7.2	0.0	v	102.9	83.2	19.7	o	160.4	123.6	36.9
	IFD (GPT2)	k	3.0	2.7	0.4	q	3.4	2.7	0.6	k	120.8	120.6	0.3	q	121.5	113.2	8.3
		v	7.1	5.7	1.4	o	7.2	5.6	1.6	v	105.5	99.1	6.4	o	161.4	147.9	13.5
	InsTag	k	3.3	3.1	0.2	q	3.4	3.3	0.1	k	137.3	104.5	32.8	q	135.6	104.0	31.5
		v	7.2	7.1	0.2	o	7.2	7.0	0.1	v	115.5	89.3	26.2	o	178.6	136.0	42.6
	Reward	k	3.0	2.8	0.2	q	3.3	2.9	0.3	k	121.2	120.1	1.1	q	122.5	123.3	-0.8
		v	7.0	6.0	1.0	o	7.0	6.0	1.0	v	106.6	99.2	7.4	o	166.4	160.9	5.5

Table 56: Gemma 2 2B - Magpie - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Magpie	Difficulty	k - v	4.8e-04	5.9e-04	q - o	-5.4e-06	-5.1e-06	k	-1.1e-03	-6.1e-04	q	-2.2e-04	-4.7e-04
		v	-8.3e-05	-5.2e-05	o	4.3e-04	4.4e-04						
	IFD (GPT2)	k - v	3.8e-04	1.7e-04	q - o	-3.3e-06	-3.1e-06	k	-6.5e-04	-1.5e-03	q	-2.7e-04	-5.4e-04
		v	-5.9e-05	-2.4e-04	o	2.2e-04	-5.7e-04						
	InsTag	k - v	5.5e-04	2.1e-04	q - o	-3.3e-06	2.6e-07	k	-1.4e-03	-9.4e-05	q	-4.5e-06	-6.7e-04
		v	4.4e-04	2.6e-04	o	-4.5e-05	3.5e-04						
	Reward	k - v	6.4e-04	-2.5e-04	q - o	-1.1e-05	3.9e-06	k	-8.0e-04	-8.1e-04	q	-5.1e-04	6.7e-05
		v	9.5e-05	1.1e-04	o	4.6e-04	3.2e-04						

Table 57: Gemma 2 2B - Magpie - Similarity-based Metrics

Dataset	Metrics	Nuclear Norm								Effective Rank							
		Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap	Proj	High	Low	Gap
Reasoning	Reasoning	k	2.5	2.5	-0.1	q	2.8	2.8	-0.0	k	194.3	149.7	44.6	q	233.8	162.8	71.0
		v	5.4	5.6	-0.2	o	5.8	5.8	0.0	v	167.4	124.1	43.3	o	301.0	208.8	92.2

Table 58: Gemma 2 2B - Reasoning - SVD-based Metrics

Dataset	Metric	Same-layer Similarity						Adjacent-layer Similarity					
		Proj	High	Low	Proj	High	Low	Proj	High	Low	Proj	High	Low
Reasoning	Reasoning	k - v	7.5e-04	4.0e-04	q - o	-2.0e-05	-5.3e-06	k	-8.5e-04	-7.4e-04	q	7.1e-04	1.0e-04
		v	1.0e-03	5.0e-04	o	-9.7e-04	-8.6e-04						

Table 59: Gemma 2 2B - Reasoning - Similarity-based Metrics

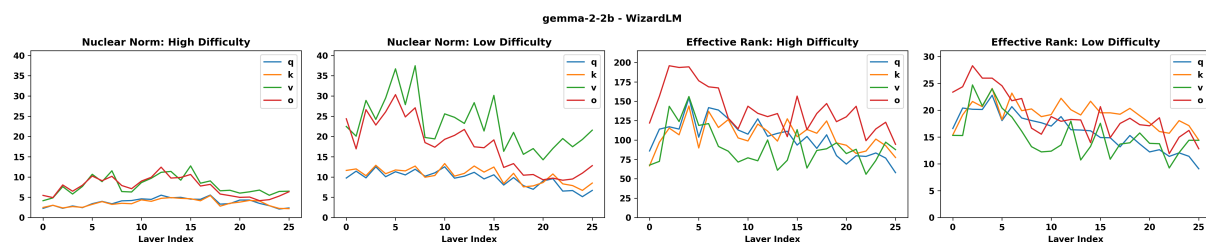


Figure 85: Gemma 2 2B - WizardLM with Difficulty Metric

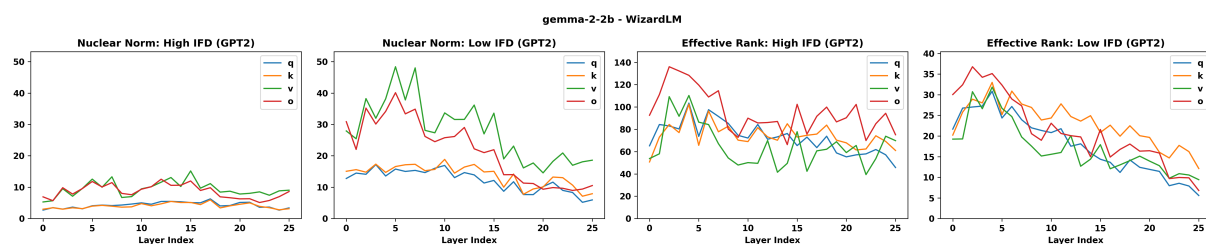


Figure 86: Gemma 2 2B - WizardLM with IFD (GPT-2) Metric

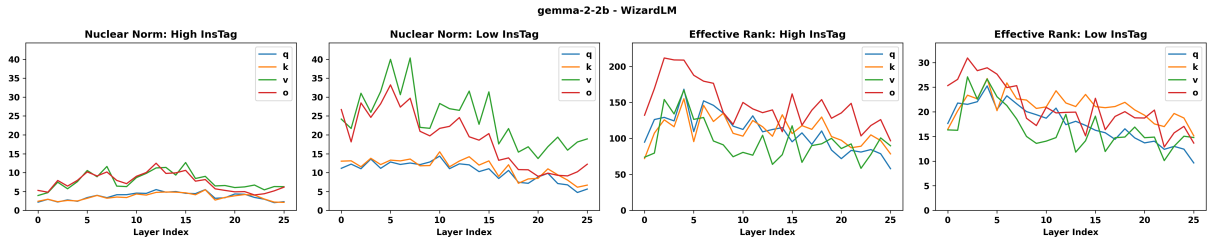


Figure 87: Gemma 2 2B - WizardLM with InsTag Metric

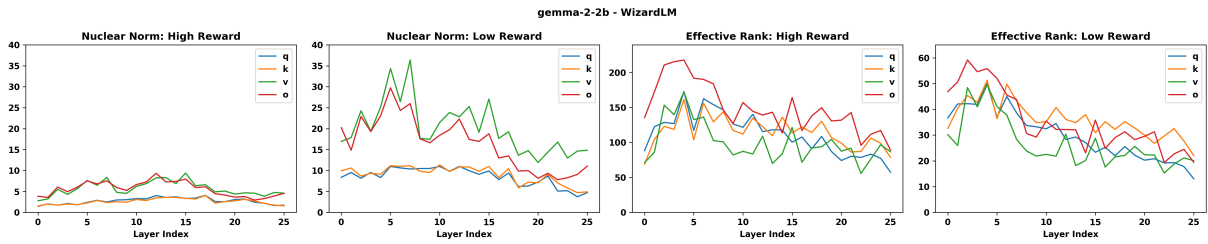


Figure 88: Gemma 2 2B - WizardLM with Reward Model Metric

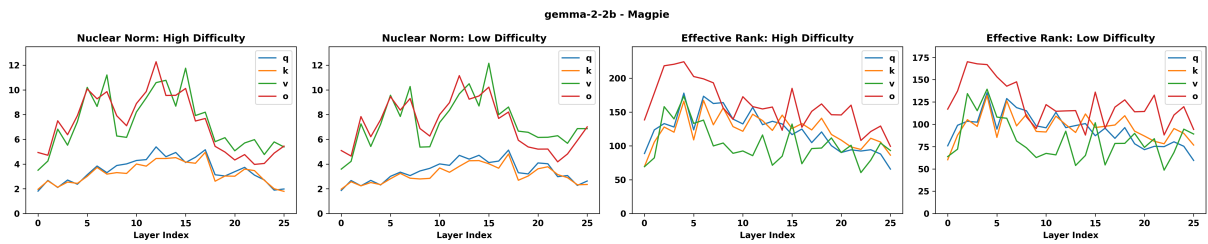


Figure 89: Gemma 2 2B - Magpie with Difficulty Metric

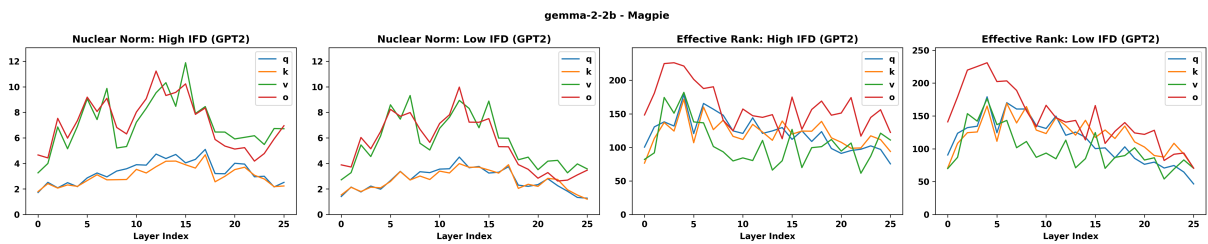


Figure 90: Gemma 2 2B - Magpie with IFD (GPT-2) Metric

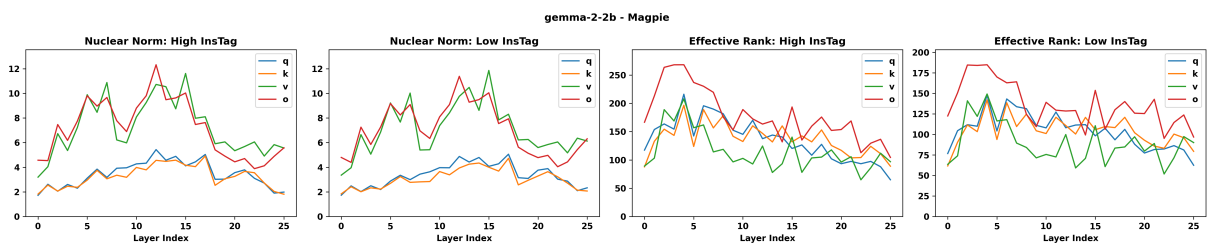


Figure 91: Gemma 2 2B - Magpie with InsTag Metric

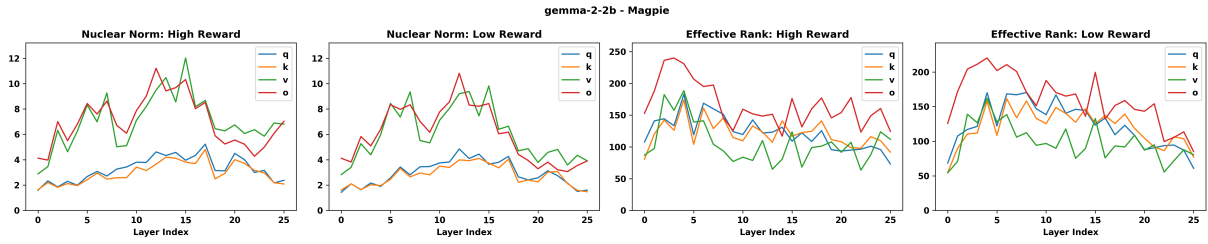


Figure 92: Gemma 2 2B - Magpie with Reward Model Metric

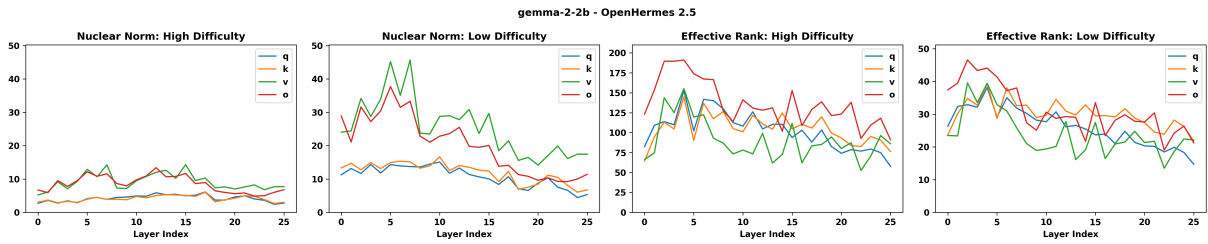


Figure 93: Gemma 2 2B - OpenHermes with Difficulty Metric

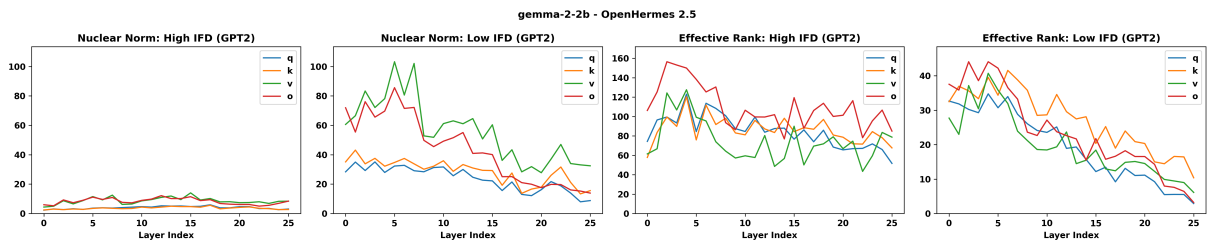


Figure 94: Gemma 2 2B - OpenHermes with IFD (GPT-2) Metric

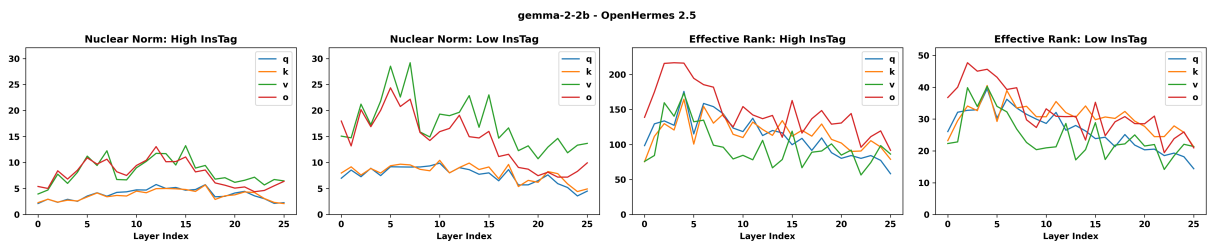


Figure 95: Gemma 2 2B - OpenHermes with InsTag Metric

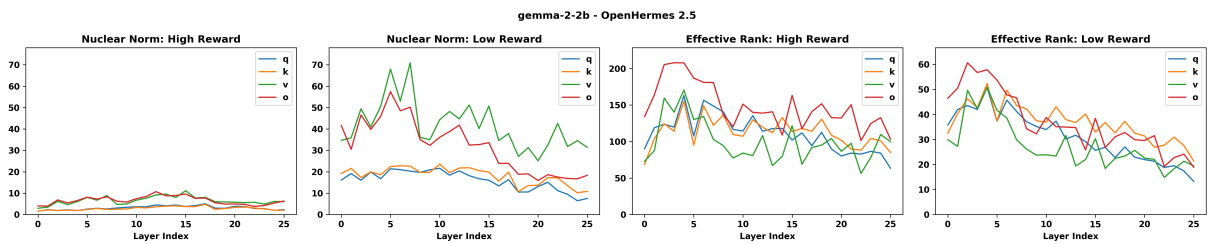


Figure 96: Gemma 2 2B - OpenHermes with Reward Model Metric