Φ-ALIGN: ASSESSING LANGUAGE MODEL REPRESENTATIONS THROUGH PREDICTION-HEAD ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Representations with identical spectra can lead to very different logits, depending on their relationship with the model's prediction head. Existing spectral metrics describe how representational variance is distributed, but overlook how the model transforms the variations into predictions. We introduce Φ-Align, a label-free metric that quantifies the alignment between representational variance and the directional sensitivity of the language modeling head. By combining the complete representation spectrum with the head's response along each principal direction, Φ-Align captures a property that representation-only metrics ignore. Experimental results show that Φ-Align provides useful insights about Large Language Model (LLM) development. Across Pythia (1B-12B) checkpoints Φ-Align reveals a non-monotonic three-phase trajectory in representation-head alignment that correlates with downstream capability emergence during pre-training. A similar three-phase trajectory appears in OLMo-2 7B checkpoints, extending the observation beyond Pythia. Across twelve benchmarks, Φ-Align correlates negatively with downstream performance with Spearman correlation -0.923 on Gemma 3 (270M-27B), -0.946 on Qwen2.5 (0.5B-72B) and -0.908 on Qwen3(0.6B-14B). Unlike spectrum-only metrics such as RankMe and α-ReQ, Φ-Align maintains a stable direction of association with downstream performance and remains stable across profiling corpora. Finally, in size-matched Qwen2.5 and Qwen3 comparison, pre-training loss favors the weaker model in every pair while width-normalized Φ-Align favors the stronger model in four of five pairs.

1 INTRODUCTION

Autoregressive language models are trained to predict the next token. Given a sequence of tokens, the model transforms each token into a contextual hidden-state representation and uses that representation at the final position to predict what comes next. These representations are proven to support a wide range of tasks through fine-tuning or in-context learning (Devlin et al., 2019; Brown et al., 2020). However, the representations vary considerably among themselves and so does their impact in next token generation. Changes along some directions may substantially affect the next-token distribution, whereas changes along others may have little effect. Therefore, understanding representations requires understanding of their internal structure and how that structure interacts with the mechanism that turns hidden-states into predictions.

Several works have studied representations through their geometry. Probing methods test whether specific information can be decoded from hidden representations (Alain & Bengio, 2016; Conneau et al., 2018; Belinkov & Glass, 2019) although their conclusions depend on probe capacity and task design (Hewitt & Liang, 2019). In contrast, similarity measures compare representations across layers, models or training conditions (Raghu et al., 2017; Kornblith et al., 2019). More recently, label-free spectral metrics have gained attention to characterize representation quality. RankMe (Garrido et al., 2023) estimates the effective dimensionality of a representation from its covariance spectrum. Another spectral metric α-ReQ (Agrawal et al., 2022) relates the representation quality to the decay of that spectrum. When applied to language models, the spectral analysis have revealed non-monotonic changes in representation geometry across pre-training that correlate with model's capability emergence that are not directly visible from cross-entropy loss (Li et al., 2025). These findings highlight that the distribution of variance across representational directions contains meaningful information about model development and performance.

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

Although these spectral metrics are informative, they characterize the representation space independently of the model's prediction mechanism. Graziani et al. (2023) noted that dominant directions of representational variance are not necessarily the ones that strongly influence the model's predictions. The NC3 property of neural collapse theory (Papyan et al., 2020) further motivates the importance of connecting representations with their prediction mechanism by showing that in a well-trained classifier the class-mean representations align with the classifier weights. Although autoregressive language models differ substantially from the supervised classification setting of neural collapse, the principle of the theory suggests that learned representations should be studied together with the prediction heads that act on them. In particular, two representations can have identical covariance eigenvalues but differ in orientation. And depending on orientation, the language model-head (LM-head) can respond differently (See details in section 2).

These observations motivate a complementary measure that explicitly connects the representation space with its corresponding LM-head. We introduce Φ-Align, a label-free scalar that measures the normalized overlap between hidden-state covariance and LM-head sensitivity. Each covariance direction is weighted by its share of representational variance and by the normalized strength with which the head maps that direction into logit space. Since softmax ignores uniform shifts of the logits, we measure this strength through a vocabulary-centered head, so that only logit changes that alter the next-token distribution contribute. Φ-Align measures neither dimensionality nor accuracy. Instead, it measures how a representation's variation is coupled to the model's prediction.

We evaluate Φ-Align in two complementary settings. First one is on pretraining dynamics and second is on fully trained models across scales. On Pythia checkpoints, Φ-Align exhibits non-monotonic trajectories that coincide with model's short and long-context capability emergence. The same trajectory appears on a second profiling corpus and on OLMo-2. Across twelve benchmarks and three model families (Gemma 3, Qwen2.5 and Qwen3), Φ-Align shows strong negative association with downstream performance. Unlike RankMe and α-ReQ, Φ-Align's sign and magnitude of association stay stable across model families and profiling corpora respectively.

Our contributions are as follows:

- We introduce Φ-Align, a label-free metric of how representational variance aligns with the directional sensitivity of the LM-head. By using a vocabulary-centered head, it counts only the sensitivity that affects the next-token prediction.
- We show that Φ-Align is bounded, scale-invariant and unaffected by uniform shifts of logits that are ignored by softmax.
- We show that Φ-Align reveals a three-phase trajectory during pre-training that coincides with model's short and long-context capability emergence and is consistent across model families and profiling corpora.
- Finally, we show that across fully trained models, Φ-Align maintains a consistent direction of association with downstream performance in three model families while spectrum-only metrics change directions. We also show that in a size-matched comparison between Qwen2.5 and Qwen3, Φ-Align favors the stronger model in four out of five pairs while loss favors the weaker model.

These contributions position Φ-Align as a complementary tool for studying how learned representation geometry relates to next token prediction and model capability.

2 Representation Orientation and Logit Variance

For a fixed linear prediction $z = Wh + b$, representation covariance Σ induces logit covariance $W\Sigma W^\top$. Consequently, total logit variance is $\mathrm{tr}(\Sigma W^\top W)$, which generally depends on the orientation of the representation covariance relative to the head, as well as their spectra. Here, W is a fixed prediction head and h is a random hidden representation.

For illustration, let

$$\Sigma_A = \mathrm{diag}(0.9, 0.1), \qquad \Sigma_B = \mathrm{diag}(0.1, 0.9), \qquad W^\top W = \mathrm{diag}(0.9, 0.1).$$

where Σ_A and Σ_B are two representation covariances and $W^\top W$ is the head-sensitivity matrix. The two representation covariances have identical eigenvalues, so any statistic depending on their spectra

takes the same value. However, their logit variances are 0.82 and 0.18, respectively. Details of this illustration are given in Appendix Section A.

This illustration highlights that representation spectra alone do not determine the total variance of the logits produced by a fixed prediction head. It motivates a head-dependent descriptor that establishes the relationship between representation and prediction head.

3 PROPOSED METRIC

The illustration in section 2 motivates the necessity of a new scalar that connects representational variance with the sensitivity of the prediction head. We derive such a scalar named Φ-Align by weighting each principal representation direction by its share of the total variance and the normalized response of the LM-head along that direction. In this section, we first describe these two components and then provide definition, interpretation, and properties of the newly proposed metric.

3.1 REPRESENTATIONAL VARIANCE DISTRIBUTION

First step to measure Φ-Align is to obtain the distribution of representational variance across hidden-state directions through eigendecomposition of the empirical covariance matrix. Given a pretrained language model f_θ and an evaluation corpus $\mathcal{D} = \{x_1, x_2, ..., x_M\}$ of M input sequences, we extract the last-layer final-token hidden-state $h_i \in \mathbb{R}^d$ of each sequence. Then, these hidden states are assembled into a centered feature matrix $F \in \mathbb{R}^{M\times d}$. Each i-th row of this matrix is equal to $(h_i - \bar{h})$, where $\bar{h}$ is the mean representation. The empirical feature covariance matrix is,

$$\Sigma = \frac{1}{M} F^\top F \in \mathbb{R}^{d\times d} \tag{1}$$

which is symmetric and positive semidefinite by construction and hence the spectral theorem guarantees an eigendecomposition $\Sigma = Q\Lambda Q^\top$ where Q is an orthogonal matrix. It yields orthonormal eigenvectors $q_1, q_2, \ldots, q_d$ that represent the principal directions of representational variance and the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d \geq 0$ that measure the variance along each direction. The detailed construction process is described in Appendix B.1.

3.2 LM-HEAD DECOMPOSITION

The LM-head $W \in \mathbb{R}^{|\mathcal{V}|\times d}$ maps a hidden representation $h \in \mathbb{R}^d$ to logits over a vocabulary $\mathcal{V}$. These logits are normalized by a softmax which is invariant to uniform shifts. So, $\mathrm{softmax}(z + c\mathbf{1}) = \mathrm{softmax}(z)$ for any $c \in \mathbb{R}$, where $\mathbf{1} \in \mathbb{R}^{|\mathcal{V}|}$ is the all-ones vector. The component of h along the mean unembedding vector $\bar{w}$ shifts all logits by the same amount, hence does not affect the next-token distribution. Since the goal of our metric is establish a relationship between representation variance and prediction mechanism and this part does not contribute to next token distribution, we discard it by subtracting $\bar{w}$ from every row of W. This gives us the centered head $W_c = W - \mathbf{1}\bar{w}^\top$.

Since modern LLMs typically have $|\mathcal{V}| \gg d$, the SVD of W_c can be written as $W_c = U_W S_W V_W^\top$, where $U_W \in \mathbb{R}^{|\mathcal{V}|\times d}$ has orthonormal columns, $S_W = \mathrm{diag}(s_1, \ldots, s_d)$ with $s_1 \geq \cdots \geq s_d \geq 0$, and the columns v_j^W of $V_W \in \mathbb{R}^{d\times d}$ are the principal sensitivity directions of the head in hidden-state space, with s_j^2 measuring the sensitivity strength along v_j^W. For any covariance eigenvector q_i,

$$\|W_c q_i\|_2^2 = \sum_{j=1}^{d} s_j^2 \, \langle q_i, v_j^W \rangle^2, \tag{2}$$

so the head responds strongly to q_i when it overlaps with high-sensitivity directions. The overlap $\langle q_i, v_j^W \rangle^2$ measures directional alignment, and s_j^2 measures response strength. The derivation is given in Appendix B.2.

3.3 SPECTRAL ALIGNMENT

Having obtained the principal variance directions (q_i) of the hidden representations and their corresponding eigenvalues (λ_i) from the covariance eigendecomposition (Section 3.1), and the principal

sensitivity directions(v_j^W) of the LM head and their corresponding squared singular values (s_j^2) from the LM-head decomposition (Section 3.2), we investigate whether directions carrying large representational variance are also those to which the LM-head is sensitive. We define Φ_W as our scalar by weighting each representation eigenvector by its share of total variance and its normalized LM-head sensitivity.

$$\Phi_W = \sum_{i=1}^{d} \underbrace{\frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}}_{\text{variance weight}} \cdot \underbrace{\frac{||W_c q_i||_2^2}{||W_c||_F^2}}_{\text{normalized head sensitivity}} \tag{3}$$

The variance weight term $\lambda_i / \sum_{j=1}^{d} \lambda_j$ is the fraction of total representational variance along the direction q_i. The normalized head sensitivity term $||W_c q_i||_2^2 / ||W_c||_F^2$ quantifies how strongly the LM head responds to movement along that direction. Their product therefore weights the head's sensitivity by the amount of representational variance occurring in that direction.

Taking all d principal directions of the representation into consideration, Φ_W provides a variance-weighted average of the LM-head's normalized directional sensitivity. For a fixed head, a higher Φ_W means that more representational variance lies along directions to which the head is strongly sensitive. In contrast, a lower value indicates that variance lies along less sensitive directions. Hence, Φ_W characterizes both variance distribution and how that distribution is oriented relative to the prediction head. Representations with identical covariance eigenvalues can therefore receive different scores when their principal directions differ relative to the head. Since Φ_W represents the alignment between the distribution of representational variance and the directional sensitivity of the LM-head, we define this scalar as Φ-Align. Since $\sum_i \lambda_i \|W_c q_i\|_2^2 = \mathrm{tr}(\Sigma W_c^\top W_c)$,

$$\Phi_W = \frac{\mathrm{tr}\big(W_c \Sigma W_c^\top\big)}{\mathrm{tr}(\Sigma)\, \|W_c\|_F^2}. \tag{4}$$

The numerator is the total variance of the vocabulary-centered logits. It is the logit variance of Section 2 that the softmax does not ignore. Φ-Align is therefore this variance, normalized by the total representation variance and the size of the head. Details are in Appendix B.3 and the properties of Φ-Align are given in B.4.

4 Experimental Setup

Φ-Align is a label-free descriptor of representation and LM-head organization. A useful descriptor should characterize meaningful changes in that organization and relate them to model behavior consistently across settings. We therefore evaluate Φ-Align along two complementary axes. First we test whether Φ-Align provides an interpretable view of representational dynamics during pre-training (Section 5) and second, we compare fully trained models of different sizes to test whether Φ-Align is associated with downstream performance (Section 6).

Models. To study the pre-training dynamics, we use Pythia (Biderman et al., 2023) (1B, 1.4B, 6.9B, and 12B) intermediate checkpoints. We omit the Pythia 2.8B model from this experiment as early stage checkpoints of this model share identical weights. To verify that the observed training dynamics generalize beyond a single model family, we additionally evaluate the intermediate checkpoints of OLMo-2 7B (OLMo et al., 2024). For downstream association analysis, we evaluate Φ-Align on five Gemma 3 models spanning 270M to 27B parameters (Team et al., 2025), seven Qwen2.5 models spanning 0.5B to 72B parameters (Qwen et al., 2025) and five Qwen3 models spanning from 0.6B to 14B parameters (Yang et al., 2025).

Φ-Align setup. For each intermediate checkpoints (Pythia and OLMo-2) and models (Gemma 3, Qwen2.5 and Qwen3), we run a forward pass over $M = 15,000$ FineWeb (Penedo et al., 2024) sequences truncated to 512 tokens. We extract the final-layer hidden-state of the last-token in each sequence. We use this representation because, according to Li et al. (2025) (i) under causal attention, the final token is attended by every other preceding token and hence aggregates information from the entire context, (ii) it directly parametrizes the model's next-token predictive distribution through the

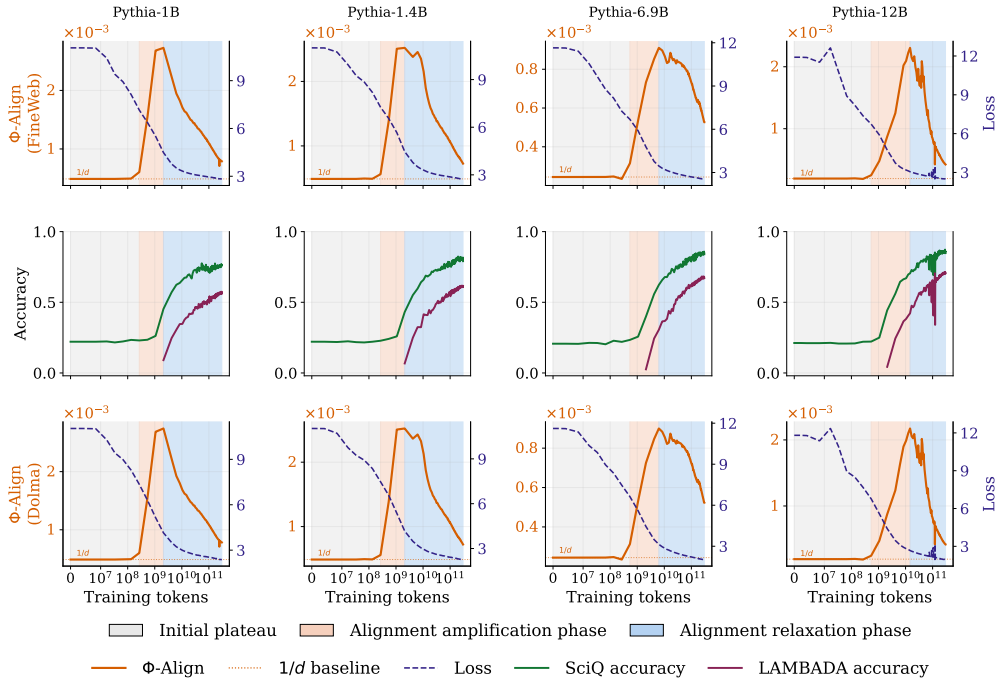

Figure 1: Φ-Align reveals three training phases that align with capability emergence. **Top and bottom row**: Φ-Align across Pythia (1B, 1.4B, 6.9B and 12B) checkpoints, profiled on FineWeb and Dolma datasets respectively. During **Initial plateau** phase, Φ-Align stays close to initial value. During **alignment amplification** phase, it rises to its peak and in **alignment relaxation** phase, it declines gradually till the end of training. **Middle row**: SciQ and LAMBADA accuracy. SciQ starts improving during alignment amplification phase while LAMBADA gains are primarily observed during alignment relaxation phase.

prediction head, (iii) this representation is typically gets passed to task-specific layers in downstream applications. Li et al. (2025) further showed that spectral properties of these representations follow the same training-phase trajectory across intermediate layers, and hence the final layer is sufficient for tracking global geometric dynamics. Since Φ-Align measures different quantity than the metrics used by Li et al. (2025), we run separate experiment by measuring Φ-Align for all layers on Pythia models, as shown in Figure 3. Φ-Align shows similar trend for all layers so we take the final layer as the representative layer.

The extracted representations and model's prediction head are then used to measure Φ-Align following the procedure described in Section 3. To assess the robustness of Φ-Align to profiling corpus, we run exact same experiments on $M = 15,000$ Dolma (Soldaini et al., 2024) sequences truncated to 512 tokens on Pythia checkpoints and Gemma 3 models.

Downstream benchmarks and evaluation: For pre-training analysis, we use SciQ (Welbl et al., 2017) and LAMBADA (Paperno et al., 2016) benchmarks. The downstream performance association was evaluated on twelve benchmarks spanning reasoning (MATH (Hendrycks et al., 2021), GSM8K(Cobbe et al., 2021), and GPQA Diamond (Rein et al., 2023)), knowledge (MMLU-Pro (Wang et al., 2024) and MMLU-STEM (Hendrycks et al., 2020)), language understanding (HellaSwag (Zellers et al., 2019), ARC-Challenge (Clark et al., 2018), and LAMBADA (Paperno et al., 2016)), question answering (NQ (Kwiatkowski et al., 2019; Lee et al., 2019), TriviaQA (Joshi et al., 2017)), and comprehensive (AGIEval (Zhong et al., 2024), DROP (Dua et al., 2019)) tasks. All benchmarks use `lm-evaluation-harness` (Gao et al., 2024) across models. A detailed experimental setup is given in Appendix C.

5 ALIGNMENT DYNAMICS DURING PRE-TRAINING

The primary motivation of Φ-Align is that representation geometry matters through both representation variance and its relationship with prediction head. Pre-training provides a longitudinal test of this perspective. Does the relationship simply strengthen or reorganize as the model learns? Our experimental results show that Φ-Align follows a non-monotonic three phase trajectory during pre-training that coincides with the emergence of model capabilities. We track short and long context capability development. SciQ is a four-way multiple choice science-exam benchmark that can often be answered through short-context distributional alignment between the question and correct answer. LAMBADA on the other hand requires the model to predict the final word of a passage from a broader context. Thus, SciQ probes short-context distributional capability, while LAMBADA probes long-context dependency.

5.1 TRAINING TRAJECTORY AND CAPABILITY EMERGENCE

Figure 1 illustrates the trajectories of Φ-Align across pre-training for Pythia 1B, 1.4B, 6.9B, and 12B checkpoints computed from representations of 15,000 samples collected from FineWeb (top row) and Dolma (bottom row) datasets. The trajectory is non-monotonic and can be divided into three phases.

Initial Plateau. During this phase, Φ-Align remains close to its initial value (Figure 1), indicating the relationship between representational variance with the LM-head is mostly unchanged. Likewise, the accuracy of SciQ and LAMBADA remains near to their initial level meaning the model didn't develop any capability to perform those tasks. Pre-training loss in contrast, keeps decreasing steadily, suggesting early loss reduction occurs without any visible gain in task capability.

Alignment amplification. Φ-Align rises sharply in this phase. Based on Equation 3 this indicates representational variance increasingly concentrates along the directions to which model's head is most sensitive to (Equation 3). Since Φ-Align is invariant to uniform rescaling of the representations or LM-head (Section B.4) this increase therefore reflects a redistribution of representational variance, a change in the head's relative directional sensitivity, or a change in their relative orientation. SciQ accuracy increases rapidly during this phase indicating model's short-context capability emergence. Pre-training loss keeps decreasing with no sign of this capability emergence.

Alignment relaxation. After the peak, Φ-Align starts decreasing gradually as training progresses, indicating a relaxation in the overlap between representation covariance and LM-head sensitivity. The decline coincides with the emergence of long-context capability. Most of the LAMBADA accuracy gain for all the models occurs in this phase highlighting model's long-context capability emergence. One possible interpretation of this relaxation is that, as the model encodes finer task-relevant structure, representational variance spreads across more directions including those to which the LM head is less sensitive.

Li et al. (2025) observed similar three phase trajectory for RankMe and α-ReQ that can also track the capability development of the model. However, these metrics characterize changes in representation spectrum while Φ-Align measures how representational directions are organized relative to the LM-head. Therefore, as shown in Figure 4, the correspondence between trajectories is not one-to-one. The spectrum can reorganize while Φ-Align remains stable and Φ-Align continue to evolve when spectrum only metrics flatten or change direction. A detailed comparison between Φ-Align and spectrum only metrics is given in Appendix D.

5.2 ROBUSTNESS ACROSS CORPORA AND FAMILIES

A useful diagnostic should not depend on its profiling corpus or the particular model family that is under study. We evaluate the robustness of Φ-Align for both cases.

Profiling Corpus. We repeat the trajectory experiment of Section 5.1 with sequences from Dolma (Soldaini et al., 2024). Φ-Align exhibits the same three phases in the same order (Figure 1 bottom row), indicating the trajectory reflects a property of the training rather than of the profiling corpus.

Table 1: Mean signed Spearman correlation with downstream benchmark performance across twelve benchmarks. Gemma 3 includes five models (270M–27B), Qwen2.5 includes seven models (0.5B–72B) and Qwen3(0.6B-14B) includes five models. The largest absolute mean correlation in each column is in bold, including ties. Φ-Align is profiled with FineWeb and Dolma datasets to demonstrate stability across corpora. Loss is used for reference.

Metric	**Gemma 3 (FineWeb)**	**Gemma 3 (Dolma)**	**Qwen2.5 (FineWeb)**	**Qwen3 (FineWeb)**
α-ReQ	-0.371	-0.647	+0.798	**+0.908**
RankMe	+0.743	+0.743	-0.949	**-0.908**
Φ-Align (ours)	**-0.923**	**-0.923**	-0.946	**-0.908**
Loss (reference)	**-0.923**	**-0.923**	**-0.985**	**-0.908**

Table 2: Final checkpoints of matched width models across families, ordered by hidden dimension d and profiled on FineWeb. The baseline $1/d$ is the value of Φ_W for isotropic representations, and $d\,\Phi_W = 1$ corresponds to it. Models with same width differs in width normalized $d\,\Phi_W$.

Model	d	$10^3/d$	$\Phi_W \times 10^3$	$d\,\Phi_W$
Pythia-1B	2048	0.488	0.768	1.57
Pythia-1.4B	2048	0.488	0.677	1.39
Qwen2.5-3B	2048	0.488	0.364	0.75
Qwen3-1.7B	2048	0.488	0.441	0.90
Gemma 3-4B	2560	0.391	0.675	1.73
Qwen3-4B	2560	0.391	0.337	0.86
OLMo-2-7B	4096	0.244	0.517	2.12
Pythia-6.9B	4096	0.244	0.431	1.77
Qwen3-8B	4096	0.244	0.195	0.80
Pythia-12B	5120	0.195	0.361	1.85
Qwen2.5-14B	5120	0.195	0.123	0.63
Qwen2.5-32B	5120	0.195	0.074	0.38
Qwen3-14B	5120	0.195	0.072	0.37

Model Family. To test whether the captured trajectory is model family specific or not, we repeat the analysis on OLMo-2 7B checkpoints profiled on FineWeb (Penedo et al., 2024). For OLMo-2 7B pre-training, Φ-Align shows similar sharp rise and then gradual decline reproducing the alignment amplification and alignment relaxation phases observed in Pythia (Bottom-left of Figure 4). The released OLMo-2 checkpoints begin after the initial training steps, so the warmup phase is only partially observed. Overall the shape of the trajectory is preserved for OLMo-2 7B checkpoints.

6 Alignment Across Fully Trained Models

Section 5 showed that Φ-Align follows a consistent three-phase trajectory on Pythia and OLMo-2 that coincides with capability emergence during pre-training. We now ask whether Φ-Align is associated with the downstream performance across fully trained models of different sizes using models from three additional families: Gemma 3 (270M-27B), Qwen2.5(0.5B-72B) and Qwen3(0.6B-14B). Our aim is not to build a performance predictor. Instead these experiments provide additional evidence that Φ-Align describes a property of the representation that is consistent across model families. To this end, we examine three aspects of association: (i) the strength of association of Φ-Align and downstream performance within a model family, (ii) whether the direction of association is stable across model families and (iii) whether the association is stable under a change in profiling corpora.

Association Strength. We measure the Spearman correlation of Φ-Align, RankMe, and α-ReQ with downstream accuracy using pre-training loss as a reference of attainable strength. Table 1 lists the mean signed Spearman correlation of the metrics across twelve benchmarks. Φ-Align shows strong negative association for all three model families with mean correlation of -0.923 for Gemma 3

Table 3: Size-matched comparison of Qwen2.5 and Qwen3 base models. The two families share essentially the same tokenizer, so their FineWeb losses are directly comparable. Qwen3 has higher mean accuracy in every pair while Qwen2.5 has lower loss. Φ_W is reported as $\Phi_W \times 10^3$. $d\,\Phi_W$ normalizes Φ_W by the isotropic baseline $1/d$. Lower loss and higher $d\Phi_W$ and mean accuracy are marked bold.

	Loss		Φ_W ($\times 10^3$)		$d\,\Phi_W$		Mean acc.	
Size (Qwen2.5 / Qwen3)	Qwen2.5	Qwen3	Qwen2.5	Qwen3	Qwen2.5	Qwen3	Qwen2.5	Qwen3
0.5B / 0.6B	**2.891**	2.899	0.670	0.961	0.60	**0.98**	26.6	**30.5**
1.5B / 1.7B	**2.571**	2.630	0.434	0.441	0.67	**0.90**	38.9	**41.6**
3B / 4B	**2.448**	2.475	0.364	0.337	0.75	**0.86**	45.6	**51.3**
7B / 8B	**2.329**	2.356	0.109	0.195	0.39	**0.80**	52.9	**55.4**
14B / 14B	**2.167**	2.271	0.123	0.072	**0.63**	0.37	57.7	**59.1**
Agrees with accuracy	0/5		3/5		4/5		–	

under both FineWeb and Dolma profiling, -0.946 for Qwen2.5 and -0.908 for Qwen3. For Gemma 3 and Qwen3, the mean correlation is equal and for Qwen2.5 is it competitive to that of pre-training loss. Spearman correlation for each benchmark and model families are listed in Table 14- 17.

Directional Stability Across Model Families. Table 1 also shows that RankMe correlates positively with performance on Gemma 3 (+0.743) but negatively on Qwen2.5 (-0.949) and Qwen3 (-0.908). α-ReQ correlates negatively with Gemma 3 (-0.371 on FineWeb and -0.647 on Dolma profiling) and positively with Qwen2.5 models (+0.798) and Qwen3(+0.908). This reversal holds for each of the twelve benchmarks with non-zero correlation, not only on average (Table 14- 17). Φ-Align in contrast achieves strong negative correlation with all three families.

Stability Across Profiling Corpora Spectral metrics require a profiling corpus and a stable diagnostic should come to the same conclusion when that corpus changes. To evaluate the corpus stability, we compare FineWeb and Dolma profiles for the Gemma 3 models. Table 1 shows that both Φ-Align and RankMe preserve their mean correlation with downstream performance while α-ReQ does not. The raw metric values are listed in Table 7 and the effect of corpus change is listed in Table 8 (Appendix E). Φ-Align is almost unaffected by the change of corpus with an average change of 0.4% in its value. Small change alone does not establish stability. The order of the models should also be preserved. α-ReQ changes by 1.4% on average in value but reorders the models between FineWeb and Dolma and its mean Spearman correlation changes from -0.371 to -0.647. Φ-Align maintains the raw values and the order, supporting its stability across profiling corpora.

7 Discussion

Effect of hidden-state dimensionality. If the representation covariance is oriented uniformly at random relative to the LM-head, the expected Φ-Align value $\mathbb{E}[\Phi_W] = 1/d$ where d is the hidden dimension (d_{model}). Since d varies across models, the observed Φ_W values could in principle reflect the baseline $1/d$ value. However, three observations prove that hidden dimension does not alone determine Φ_W. The first observation is that d is fixed within training run, but Φ_W departs from the baseline $1/d$ for every Pythia model (Figure 1). The second observation is that models with identical hidden dimension, hence identical baseline differ substantially in raw Φ_W values. Pythia-6.9B and OLMo-2-7B share both width ($d = 4096$) and depth ($L = 32$), yet their values are $\Phi_W = 0.431 \times 10^{-3}$ and 0.517×10^{-3} while their width normalized $d\,\Phi_W$ is 1.77 and 2.12 respectively, as shown in Table 2. Other models with similar width also differ when normalized by width ($d\Phi_W$). The third observation is that the ordering of Φ_W values can run against the baseline $1/d$. Fore example OLMo-2 7B ($d = 4096, \Phi_W = 0.517 \times 10^{-3}$) has higher dimension than Qwne2.5 3B ($d = 2048, \Phi_W = 0.364 \times 10^{-3}$) but also higher Φ-Align. Table 2 therefore proves that hidden dimension sets the reference level for Φ_W, but the value reflects how representational variance is organized with respect to the head's directional sensitivity.

Loss Across Model Families. Comparing loss across model families is usually confounded by tokenization since different tokenizer defines different tasks. Qwen2.5 and Qwen3 share the same tokenizer. We evaluate both on FineWeb sequences so their losses are directly comparable. Paring the models by their sizes removes the scale confound. So in this setting loss should be most informative. But at every matched size in Table 3, Qwen2.5 has lower loss than Qwen3 while Qwen3 has higher mean accuracy across twelve benchmarks. So, loss ranks the two families in wrong order for all five pairs. We examine whether Φ_W carries complementary information that can distinguish the two families. Since the paired models vary in width, we used width normalized $d\Phi_W$ which measures alignment relative to its isotropic baseline $1/d$. Qwen3 has higher $d\Phi_W$ in four out of five pairs. It should be noted that this comparison is restricted to families where loss is directly comparable.

8 Related Work

Representation Geometry for LLMs. The geometry of hidden representation exposed by spectral metrics help to describe how representational variance is distributed. Metrics like RankMe (Garrido et al., 2023) measure effective rank and α-ReQ measures eigenspectrum decay. Li et al. (2025) showed that these metrics evolve non-monotonically during pre-training, revealing phases that loss does not expose. These metrics use representational variance alone and do not consider which direction is used by the prediction head. However, contextual representations are known to be anisotropic (Ethayarajh, 2019) with embeddings concentrated in a narrow cone (Gao et al., 2019) and a few outlier dimensions dominating the variance (Timkey & Van Schijndel, 2021; Sun et al., 2024). Such dimensions can distort variance based measures. This motivates a new metric that measures variance through the head's sensitivity instead measuring it in isolation.

Neural Collapse. Neural collapse is a phenomenon observed near the end phase of supervised classification training (Papyan et al., 2020). It describes four geometric properties that emerge collectively and the most relevant one to our work is NC3. This property formalizes the alignment between learned features and the final classifier weights. Language Models (LMs) differ substantially from the settings of neural collapse because they predict over a very large and highly imbalanced vocabulary rather than a small set of balanced classes. Therefore, the exact weight-feature alignment of NC3 does not directly apply to these models, but the main principle remains useful. A well-trained prediction head should rely on directions that are meaningful in the representation space. SliceGPT (Ashkboos et al., 2024) supports this intuition by showing that LLM performance can be largely preserved by removing selected hidden-space directions. This suggests that different hidden directions have unequal functional importance.

Downstream Performance Prediction. Several works have been done to predict downstream performance from compute, loss, or benchmark observations across models (Ruan et al., 2024; Chen et al., 2024; Xu et al., 2026). The goal of this paper is not to build a performance predictor. Instead, we examine whether a label-free geometric property relates to performance consistently across model families and profiling corpora.

9 Conclusion

We introduce Φ-Align, a label-free metric that measures how representational variance aligns with the directional sensitivity of the LM-head, while ignoring the uniform logit shifts that does not impact the next token distribution. During pre-training, Φ-Align captures a three-phase trajectory that is stable across profiling corpora and model families and coincides with models capability emergence. Across fully trained models, Φ-Align maintains a negative correlation with downstream performance that is stable under changes in model families and profiling corpora. Φ-Align is not determined by model width alone as matched-width analysis shows that models with identical width have substantially different values indicating Φ-Align adds information beyond model width. Finally, Φ-Align is complementary to loss. In a size-matched comparison between Qwen2.5 and Qwen3 models, width-normalized Φ-Align favors the stronger model in four out of five pairs while loss favors the weaker model in all five. These results suggest that along with pre-training loss and representation spectrum, the relationship between representational variance and prediction head is an informative and complementary lens on model development.

REFERENCES

Kumar Krishna Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake A. Richards. α-req : Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/70596d70542c51c8d9b4e423f4bf2736-Abstract-Conference.html.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Saleh Ashkboos, Maximilian Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. Slicegpt: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, volume 2024, pp. 11682–11701, 2024.

Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. Scaling laws for predicting downstream performance in llms. *arXiv preprint arXiv:2410.08527*, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, 2019.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pp. 10929–10974. PMLR, 2023.

Mara Graziani, Laura O' Mahony, An-phi Nguyen, Henning Müller, and Vincent Andrearczyk. Uncovering unique concept vectors through latent space decomposition. *arXiv preprint arXiv:2307.06913*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6086–6096, 2019.

Melody Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Teru, Adam Santoro, Guillaume Lajoie, and Blake Richards. Tracing the representation geometry of language models from pretraining to post-training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 54691–54724. Curran Associates, Inc., 2025. doi: 10.52202/085713-1829. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4eac651e6c06d6b374332cd14ab50f55-Paper-Conference.pdf.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1525–1534, 2016.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=n6SCkn2QaG.

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, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Yangjun Ruan, Chris J Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of langauge model performance. *Advances in Neural Information Processing Systems*, 37:15841–15892, 2024.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

William Timkey and Marten Van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, 2021.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Chengyin Xu, Kaiyuan Chen, Xiao Li, Ke Shen, and Chenggang Li. Unveiling downstream performance scaling of llms: A clustering-based perspective. In *International Conference on Learning Representations*, volume 2026, pp. 80188–80212, 2026.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 2299–2314, 2024.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing, including improving grammar, clarity, and readability. We did not use generative AI tools to formulate, develop Φ-Align or design experiments or interpret the results. Generating synthetic data, translation, and qualitative data analysis are not applicable to this work. All AI assisted texts are reviewed and edited by the authors. We take full responsibility for the final content of this paper.

ETHICS STATEMENT

The research conforms with the ICLR Code of Ethics. The work uses publicly available pre-trained models and datasets and does not involve human subjects and the proposed metric does not introduce new risks.

REPRODUCIBILITY STATEMENT

The metric formulation is clearly mentioned in Section 3. The paper uses publicly available models, datasets and benchmarks and the full experimental setups for reproducibility are given in Appendix C.

A COVARIANCE AND TOTAL VARIANCE OF LINEAR LOGITS

Let $h_1, \ldots, h_M \in \mathbb{R}^d$ be hidden representations with empirical mean and covariance

$$\bar{h} = \frac{1}{M}\sum_{i=1}^{M} h_i, \qquad \Sigma = \frac{1}{M}\sum_{i=1}^{M}(h_i - \bar{h})(h_i - \bar{h})^\top.$$

For a fixed prediction head, $W \in \mathbb{R}^{|V|\times d}$ and bias $b \in \mathbb{R}^{|V|}$, the corresponding logits are $z_i = Wh_i + b$. Their empirical mean is,

$$\bar{z} = \frac{1}{M}\sum_{i=1}^{M} z_i = W\bar{h} + b.$$

Thus, $z_i - \bar{z} = W(h_i - \bar{h})$, and the empirical logit covariance is

$$\Sigma_z = \frac{1}{M}\sum_{i=1}^{M}(z_i - \bar{z})(z_i - \bar{z})^\top \quad = \frac{1}{M}\sum_{i=1}^{M} W(h_i - \bar{h})(h_i - \bar{h})^\top W^\top \; = W\Sigma W^\top.$$

Each diagonal entry of Σ_z is the empirical variance of one vocabulary logit across the M representations. Adding these entries gives the total logit variance,

$$\mathrm{tr}(\Sigma_z) = \mathrm{tr}(W\Sigma W^\top) = \mathrm{tr}(\Sigma W^\top W),$$

We derived the final form from the cyclic property of the trace. This quantity measures variation across input representations.

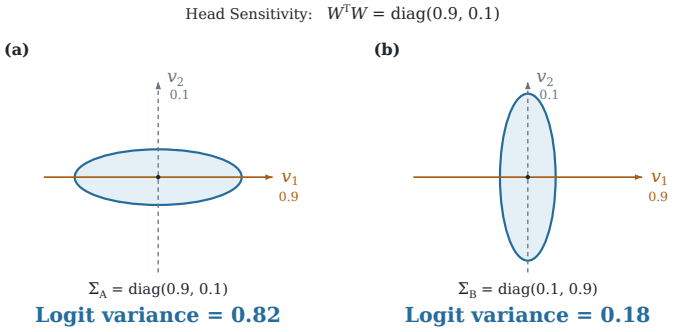

Figure 2: Identical eigenspectra can yield different total logit variance under a fixed prediction head.

Figure 2 visualizes the Section 2 example. The example involves two covariance matrices $\Sigma_A = \mathrm{diag}(0.9, 0.1)$ and $\Sigma_B = \mathrm{diag}(0.1, 0.9)$ a fixed prediction head with sensitivity matrix $W^\top W = \mathrm{diag}(0.9, 0.1)$).

So for first case, we get:

$$\mathrm{tr}(\Sigma_A W^\top W) = 0.9(0.9) + 0.1(0.1) = 0.82,$$

and for second case, we get:

$$\mathrm{tr}(\Sigma_B W^\top W) = 0.1(0.9) + 0.9(0.1) = 0.18.$$

This example illustrates, two representations with identical eigenspectra result in different logit variance under the same prediction head.

B Derivations of covariance matrix and LM-head decomposition

B.1 Covariance matrix construction

Given a pretrained language model f_θ and an evaluation corpus $\mathcal{D} = \{x_1, x_2, ..., x_M\}$ of M input sequences, the model produces final-layer hidden states $H_i^{(L)} = f_\theta^{(L)}(x_i) \in \mathbb{R}^{T_i \times d}$ for each sequence x_i, where T_i is the sequence length, L denotes the final transformer layer, and d is the hidden dimension. We only use the final-token hidden state, $h_i = H_i^{(L)}[T_i] \in \mathbb{R}^d$. Then these representations are assembled into a centered feature matrix:

$$F = \begin{bmatrix} (h_1 - \bar{h})^\top \\ \vdots \\ (h_M - \bar{h})^\top \end{bmatrix} \in \mathbb{R}^{M \times d}.$$

where $\bar{h} = \frac{1}{M}\sum_{i=1}^{M} h_i$ is the mean representation. So the empirical feature covariance is:

$$\Sigma = \frac{1}{M} F^\top F \in \mathbb{R}^{d \times d}$$

Since Σ is constructed as $F^\top F$ (scaled by $1/M$), it is symmetric and positive semidefinite by construction. By the spectral theorem, its eigendecomposition is, $\Sigma = Q\Lambda Q^\top$ where $Q = [q_1, q_2, ..., q_d] \in \mathbb{R}^{d \times d}$. Q is an orthogonal matrix whose columns $\{q_1, q_2, ..., q_d\}$ represent the eigenvectors and $\Lambda = \text{diag}\,(\lambda_1, \lambda_2, ..., \lambda_d)$ is a diagonal matrix of eigenvalues $\{\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d \geq 0\}$. Eigenvectors ($q_i$) are the principal directions along which the representations vary and the eigenvalues (λ_i) are the measure of how much variance the representations exhibit along each corresponding direction.

B.2 Sensitivity directions of the LM-head

The LM head W maps the hidden representation h to a vector of vocabulary logits $z = Wh$. Each row w_v of W is the unembedding vector of token v, so the logit of token v is $z_v = w_v^\top h$. To determine how strongly logits respond to different hidden-state directions, consider changing h by a small amount εu, where u is a unit direction. The corresponding change in the logits is,

$$\Delta z = W(h + \varepsilon u) - Wh = \varepsilon W u$$

Since $\text{softmax}(z + c\mathbf{1}) = \text{softmax}(z)$, the part of Δz that shifts all logits equally does not affect the next token distribution. Let $\bar{w} = \frac{1}{|\mathcal{V}|}\sum_v w_v$ be the mean unembedding vector or the average row of W. The logit of token v can then be split as,

$$z_v = \bar{w}^\top h + (w_v - \bar{w})^\top h$$

The first part is the same for every token and only shifts all logits together, while the second part is the logit relative to the average and determines the prediction. As described in Section 3.2, we remove this shift term by centering the prediction head. The centered head can be written as,

$$W_c = W - \mathbf{1}\bar{w}^\top$$

The logit change can be split as, $Wu = W_c u + (\bar{w}^\top u)\mathbf{1}$. Since the rows of W_c sum to zero, the two parts are orthogonal. So we can write,

$$\|Wu\|_2^2 = \|W_c u\|_2^2 + |\mathcal{V}|\,(\bar{w}^\top u)^2, \tag{5}$$

where the second term, the shift term, measures a change that the softmax ignores. Centering removes only this term, since $W_c^\top W_c = W^\top W - |\mathcal{V}|\bar{w}\bar{w}^\top$ differs from $W^\top W$ only along $\bar{w}$.

For a unit change, the squared centered logit response is therefore $\|W_c u\|_2^2$:

$$\|W_c u\|_2^2 = (W_c u)^\top (W_c u) = u^\top W_c^\top W_c u \tag{6}$$

As mentioned in section 3.2, the SVD of centered LM-head is, $W_c = U_W S_W V_W^\top$. Since the columns of U_W and V_W are orthonormal which means each column has unit length and that every pair of distinct columns are orthogonal. Therefore, $U_W^\top U_W = I_d$. For diagonal matrix S_W, we can write:

$$W_c^\top W_c = V_W S_W^\top U_W^\top U_W S_W V_W^\top = V_W S_W^2 V_W^\top \tag{7}$$

For any unit direction $u \in \mathbb{R}^d$, using Equation 6 and 7 we get,

$$||W_c u||_2^2 = u^\top W_c^\top W_c u = u^\top V_W S_W^2 V_W^\top u \tag{8}$$

Let, q_i be the i-th eigenvector of the representation matrix. So from, Equation 8, we get,

$$||W_c q_i||_2^2 = q_i^\top V_W S_W^2 V_W^\top q_i \tag{9}$$

Let's assume, $a = V_W^\top q_i = [\langle q_i, v_1^W \rangle, \ldots, \langle q_i, v_d^W \rangle]$. Here, each entry $a_j = \langle q_i, v_j^W \rangle$ is the projection or coordinate of q_i along the right singular vector v_j^W.

From Equation 9 we can write,

$$q_i^\top V_W S_W^2 V_W^\top q_i = a^\top S_W^2 a$$

Since, $S_W^2 = \text{diag}(s_1^2, \ldots, s_d^2)$ is diagonal, we can write the above quadratic form as,

$$a^\top S_W^2 a = s_1^2 a_1^2 + \ldots + s_d^2 a_d^2 = \sum_{j=1}^{d} s_j^2 a_j^2$$

From Equation 9, we get,

$$||W_c q_i||_2^2 = q_i^\top V_W S_W^2 V_W^\top q_i = a^\top S_W^2 a = \sum_{j=1}^{d} s_j^2 \langle q_i, v_j^W \rangle^2$$

This is the Equation 2 of Section 3.2. v_j^W of V_W are eigenvectors of $W_c^\top W_c$ and the corresponding eigenvalues are s_j^2.

B.3 Derivation from Equation 3 to 4

In this section, we derive the trace form of Φ-Align. We start from Equation 3,

$$\Phi_W = \sum_{i=1}^{d} \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j} \cdot \frac{\|W_c q_i\|_2^2}{\|W_c\|_F^2}$$

The covariance has the eigendecomposition $\Sigma = Q \Lambda Q^\top$, where $Q = [q_1, \ldots, q_d]$ has orthonormal columns. So we can write $Q^\top Q = I_d$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$. Using the cyclic property of the trace:

$$\text{tr}(\Sigma) = \text{tr}(Q \Lambda Q^\top) = \text{tr}(\Lambda Q^\top Q) = \text{tr}(\Lambda) = \sum_{j=1}^{d} \lambda_j. \tag{10}$$

The denominator of the variance weight is therefore $\text{tr}(\Sigma)$. Since neither $\text{tr}(\Sigma)$ nor $\|W_c\|_F^2$ depends on i, both can be taken outside the sum:

$$\Phi_W = \frac{1}{\text{tr}(\Sigma)\, \|W_c\|_F^2} \sum_{i=1}^{d} \lambda_i \|W_c q_i\|_2^2. \tag{11}$$

From Equation 8, we can write:

$$\|W_c q_i\|_2^2 = (W_c q_i)^\top (W_c q_i) = q_i^\top W_c^\top W_c\, q_i. \tag{12}$$

Let $A = W_c^\top W_c$. The quantity $q_i^\top A q_i$ is a scalar and a scalar equals its own trace. Using the cyclic property of the trace:

$$q_i^\top A q_i = \mathrm{tr}(q_i^\top A q_i) = \mathrm{tr}(A\, q_i q_i^\top). \tag{13}$$

Substituting Equation 13 into the Equation 11 and using the linearity of the trace,

$$\sum_{i=1}^{d} \lambda_i \|W_c q_i\|_2^2 = \sum_{i=1}^{d} \lambda_i \,\mathrm{tr}(A\, q_i q_i^\top) = \mathrm{tr}\Big(A \sum_{i=1}^{d} \lambda_i\, q_i q_i^\top\Big). \tag{14}$$

The eigendecomposition can also be written as a sum of rank-one terms, $\Sigma = Q\Lambda Q^\top = \sum_{i=1}^{d} \lambda_i\, q_i q_i^\top$. Therefore,

$$\sum_{i=1}^{d} \lambda_i \|W_c q_i\|_2^2 = \mathrm{tr}(A\,\Sigma) = \mathrm{tr}(W_c^\top W_c\,\Sigma) = \mathrm{tr}(W_c\,\Sigma\, W_c^\top), \tag{15}$$

where the last equality again uses the cyclic property of the trace.

Substituting Equation 15 into Equation 11 gives

$$\Phi_W = \frac{\mathrm{tr}(W_c\,\Sigma\, W_c^\top)}{\mathrm{tr}(\Sigma)\,\|W_c\|_F^2}. \tag{16}$$

B.4 Properties of Φ-Align

Boundedness. The division by Frobenius norm $||W_c||_F^2$ normalizes Φ-Align between $[0, 1]$ for non-zero representation variance and LM-head weights. A value of zero means the observed representation variance makes no change in the linear logit. A higher value indicates greater sensitivity of the LM-head along the directions of the representation.

Invariance to Scale. Φ-Align is indifferent to scaling of the representations or LM-head weights. Multiplying either all representations or all LM-head by a non-zero scalar will not change the value of Φ-Align because corresponding scale factors get canceled through normalization. In this way, the score captures the relative distribution of variance and head sensitivity, not their overall magnitudes. Rescaling individual directions differently can change the score.

Invariance to Uniform Logit Shifts. Because Φ-Align uses the centered head, it is unaffected by change in LM-head that raises or decreases all the logits uniformly. Such changes adds or subtracts the same vector to every row of W. It gets removed when the mean unembedding vector is subtracted. This subtraction makes Φ-Align invariant to uniform logit shift.

Independence of Downstream Labels. Measuring Φ-Align requires only unlabeled data and their resulting representations. No downstream annotations or probe training is required. However, the score is dependent on the profiling corpus which should be kept consistent across comparisons.

C Detailed experimental setup for reproducibility

C.1 Models

Table 4: Models used for the experiments of this paper. d denotes the hidden dimension. Pythia and OLMo-2 intermediate and final checkpoints were used for pre-training dynamics analysis. Gemma 3, Qwen2.5 and Qwen3 models were used for downstream performance analysis.

Family	**Size**	d	**Checkpoints**	**HuggingFace ID**
Pythia	1B	2048	Intermediate + final	`EleutherAI/pythia-1b`
Pythia	1.4B	2048	Intermediate + final	`EleutherAI/pythia-1.4b`
Pythia	6.9B	4096	Intermediate + final	`EleutherAI/pythia-6.9b`
Pythia	12B	5120	Intermediate + final	`EleutherAI/pythia-12b`
OLMo-2	7B	4096	Intermediate + final	`allenai/OLMo-2-1124-7B`
Gemma 3	270M	640	Final	`google/gemma-3-270m`
Gemma 3	1B	1152	Final	`google/gemma-3-1b-pt`
Gemma 3	4B	2560	Final	`google/gemma-3-4b-pt`
Gemma 3	12B	3840	Final	`google/gemma-3-12b-pt`
Gemma 3	27B	5376	Final	`google/gemma-3-27b-pt`
Qwen2.5	0.5B	896	Final	`Qwen/Qwen2.5-0.5B`
Qwen2.5	1.5B	1536	Final	`Qwen/Qwen2.5-1.5B`
Qwen2.5	3B	2048	Final	`Qwen/Qwen2.5-3B`
Qwen2.5	7B	3584	Final	`Qwen/Qwen2.5-7B`
Qwen2.5	14B	5120	Final	`Qwen/Qwen2.5-14B`
Qwen2.5	32B	5120	Final	`Qwen/Qwen2.5-32B`
Qwen2.5	72B	8192	Final	`Qwen/Qwen2.5-72B`
Qwen3	0.6B	1024	Final	`Qwen/Qwen3-0.6B-Base`
Qwen3	1.7B	2048	Final	`Qwen/Qwen3-1.7B-Base`
Qwen3	4B	2560	Final	`Qwen/Qwen3-4B-Base`
Qwen3	8B	4096	Final	`Qwen/Qwen3-8B-Base`
Qwen3	14B	5120	Final	`Qwen/Qwen3-14B-Base`

C.2 Geometric metric calculation setup

Table 5 lists loss, RankMe, α-ReQ, and Φ-Align computation parameters. Identical across all models and intermediate checkpoints.

Table 5: Φ-Align, RankMe, α-ReQ and loss computation parameters. Identical across all models and families.

Parameter	**Value**
Dataset	FineWeb and Dolma
Number of sequences (M)	15,000
Sequence length	512 tokens (truncated)
Source corpus	FineWeb (Penedo et al., 2024) (License: `odc-by`)
FineWeb subset	`sample-10BT`
Hidden state extraction	Last-layer, final-token
Forward-pass precision	float16
Covariance / decomposition	`svd` (default), `torch.linalg.eigh` (fallback) [float32]

C.3 Compute resources

Pythia (1B-12B), OLMo-2 7B, Qwen3 (0.6B-14B), Qwen2.5(0.5B-14B), Gemma 3 (270M-12B) checkpoints and models used 1$\times$ A100 GPU. Qwen2.5 32B and Gemma 3 27B used 2$\times$ A100 80GB GPUs and Qwen2.5 72B used 3$\times$ A100 80GB GPUs.

C.4 BENCHMARK SETUP

All evaluations use `lm-evaluation-harness` (Gao et al. (2024)) `v0.4.7`. A detailed benchmark setting is shared in Table 6. Proper citations for each benchmark is added in the main paper.

Table 6: Benchmark configurations. The table lists task names, metrics used to evaluate respective benchmark and number of shots used for evaluation.

Benchmark	Task Name	Metric	Shots
MATH	`hendrycks_math`	exact_match	4
GSM8K	`gsm8k`	exact_match (flex.)	4
GPQA Diamond	`gpqa_diamond_n_shot`	acc_norm	5
MMLU-Pro	`mmlu_pro`	exact_match	5
MMLU-STEM	`mmlu_stem`	acc	5
HellaSwag	`hellaswag`	acc_norm	10
ARC-Challenge	`arc_challenge`	acc_norm	25
LAMBADA	`lambada_openai`	acc	0
NQ	`nq_open`	exact_match	0
TriviaQA	`triviaqa`	exact_match	0
AGIEval	`agieval`	acc_norm	0
DROP	`drop`	f1	0

C.5 ALL LAYER Φ-ALIGN DYNAMICS DURING PRE-TRAINING

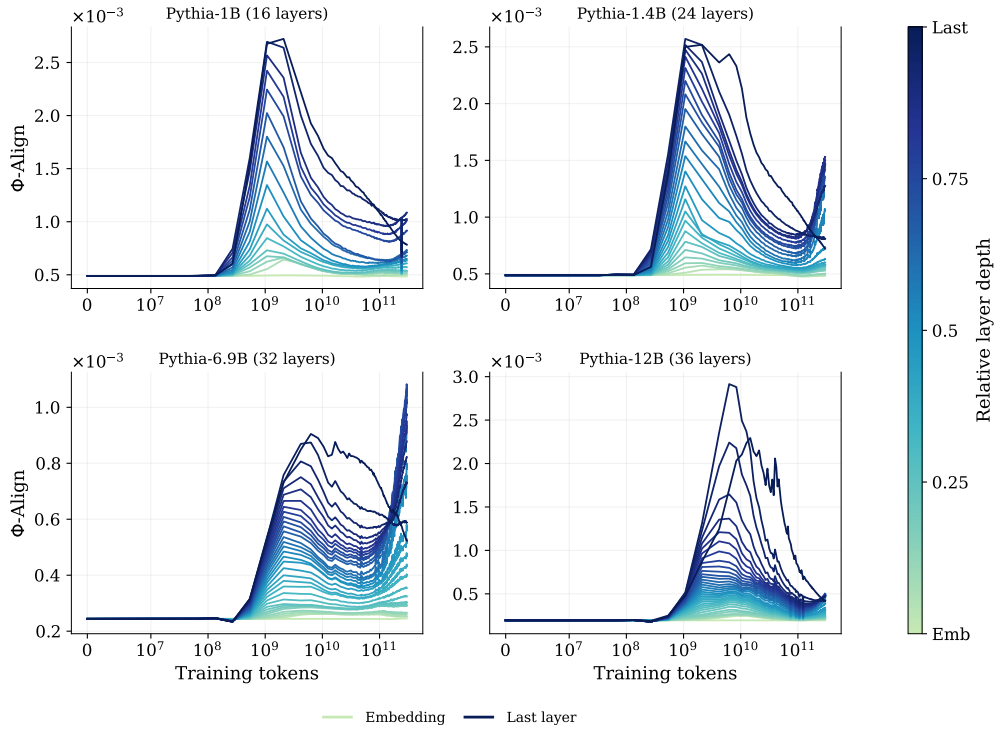

Figure 3: Training dynamics of all layers of Pythia (1B, 1.4B, 6.9B and 12B) models.

D COMPARISON WITH SPECTRUM ONLY METRICS

Li et al. (2025) showed a three phase trajectory using RankMe and α-ReQ naming these phases. The authors named those phases as warmup, entropy-seeking and compression-seeking. Figure 4

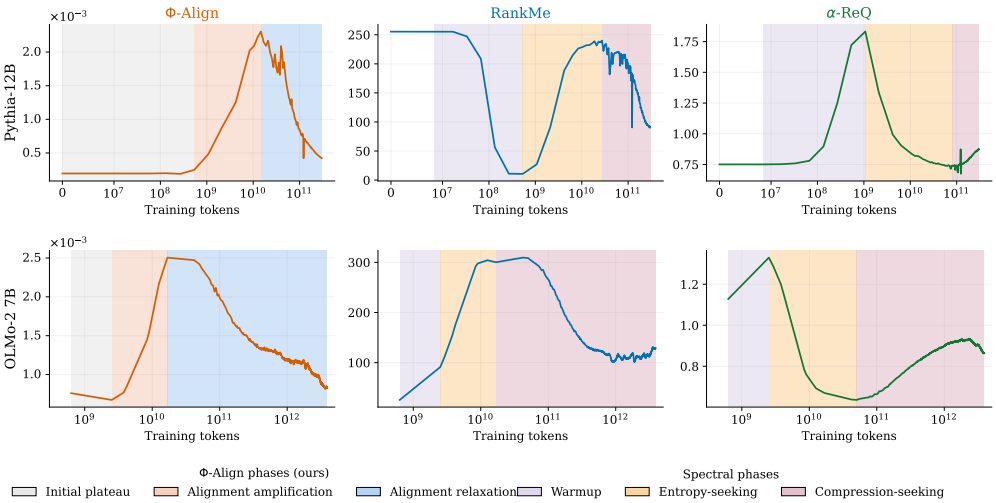

Figure 4: Φ-Align, RankMe and α-ReQ during pre-training of Pythia 12B (top) and OLMo-2 7B (bottom). Φ-Align phases are initial plateau, alignment amplification and alignment relaxation while RankMe and α-ReQ phases are warmup, entropy-seeking and compression-seeking (Li et al., 2025). Φ-Align shows similar trends in trajectory for both models in each phases. Excluding the initial training steps, RankMe and α-ReQ also demonstrate three phase trajectory.

compares Φ-Align, RankMe, and α-ReQ on Pythia 12B and OLMo-2 7B. OLMo-2 7b reproduces the initial plateau (partially since initial checkpoints are unavailable), alignment amplification and alignment relaxation phases (Bottom-left of Figure 4).

Our analysis differs from Li et al. (2025) in checkpoint coverage. Li et al. (2025) started reporting the RankMe and α-ReQ values from 10^7 tokens while we started reporting from step 0. Over these additional checkpoints, the spectral metrics mostly stay close to their initial values till the warmup collapse occurs. Subsequently the entropy-seeking and compression-seeking phase show up. The plateau simply precedes these phases. Φ-Align on the other hand, stays at its initial values during this plateau and warmup collapse. This indicates, the spectrum can reorganize substantially without changing how variance aligns with the head. Late in OLMo-2 7b training, RankMe plateaus and then rises slightly and α-ReQ turns downward in the same phase while Φ-Align consistently decreases till the end of training.

E Raw metric values for Gemma 3, Qwen2.5 and Qwen3 families

In this section we list the raw loss, α-ReQ, RankMe and Φ-Align values in table 7, 9 and 10 using FineWeb as profiling corpus. For stability test, we recompute all four metrics for Gemma 3 models on 15,000 Dolma sequences of 512 tokens each. Table 8 lists the resulting change in each metric and whether the ordering of the models is preserved.

Table 7: Raw metric values for the Gemma 3 models, computed on 15,000 FineWeb (FW) and 15,000 Dolma (DM) sequences. Φ-Align is computed with the vocabulary-centered head and reported as $\Phi_W \times 10^3$.

	Loss		**α-ReQ**		**RankMe**		**Φ-Align**	
Model	FW	DM	FW	DM	FW	DM	FW	DM
270M	3.002	3.045	0.9033	0.9177	174.9	178.0	1.243	1.241
1B	2.617	2.597	0.7439	0.7560	162.0	167.2	1.074	1.081
4B	2.357	2.169	0.7407	0.7506	196.4	196.4	0.675	0.677
12B	2.197	1.942	0.7175	0.7261	246.9	244.8	0.551	0.551
27B	2.130	1.820	0.7444	0.7534	224.5	225.1	0.418	0.415

Table 8: Sensitivity of each metric to the profiling corpus, measured on the five Gemma 3 models (270M–27B) between the FineWeb and Dolma profiles. Changes are relative to the FineWeb value. The model in parentheses is where the maximum change occurs. The last column reports the Spearman correlation between the FineWeb and Dolma values. It measures whether the ordering of the five models is preserved.

Metric	**Mean $\lvert\Delta\rvert$ (%)**	**Max $\lvert\Delta\rvert$ (%)** (Model)	**Mean $\lvert\Delta\rvert$**	**Rank ρ**
Loss	7.27	14.55 (27B)	0.163	1.000
α-ReQ	1.39	1.63 (1B)	1.08×10^{-2}	0.900
RankMe	1.21	3.19 (1B)	2.18	1.000
Φ-Align (ours)	0.40	0.72 (27B)	3.02×10^{-6}	1.000

Table 9: Loss, α-ReQ, RankMe, and Φ-Align values for the Qwen2.5 models, computed on 15,000 FineWeb sequences. Φ-Align is computed with the vocabulary-centered head and reported as $\Phi_W \times 10^3$.

Model	**Loss**	**α-ReQ**	**RankMe**	**Φ-Align**
0.5B	2.891	1.001	78.7	0.670
1.5B	2.571	0.950	46.9	0.434
3B	2.448	0.945	46.3	0.364
7B	2.329	1.228	41.3	0.109
14B	2.167	1.222	40.8	0.123
32B	2.146	1.333	28.7	0.074
72B	2.041	1.374	29.3	0.029

Table 10: Loss, α-ReQ, RankMe, and Φ-Align values for the Qwen3 base models, computed on 15,000 FineWeb sequences. Φ-Align is computed with the vocabulary-centered head and reported as $\Phi_W \times 10^3$.

Model	**Loss**	**α-ReQ**	**RankMe**	**Φ-Align**
0.6B	2.899	0.912	80.7	0.961
1.7B	2.630	0.960	38.2	0.441
4B	2.475	0.980	37.6	0.337
8B	2.356	1.051	18.0	0.195
14B	2.271	1.244	9.5	0.072

F DOWNSTREAM PERFORMANCE

This section lists performance score of Gemma 3, Qwen2.5 and Qwen3 models. Each benchmark used different settings shots of evaluation as listed in Section Table 6.

Table 11: Downstream performance of the Gemma 3 models on the twelve benchmarks (%). All benchmarks report accuracy except DROP, which reports F1. The best score in each row is in bold.

Category	Benchmark	270M	1B	4B	12B	27B
Reasoning	MATH	3.2	3.7	15.0	25.0	**31.1**
	GSM8K	1.6	2.0	38.1	72.1	**81.4**
	GPQA Diamond	29.3	22.2	33.3	33.3	**40.9**
Knowledge	MMLU-Pro	7.0	9.4	29.7	44.5	**52.4**
	MMLU-STEM	27.2	26.6	51.7	68.7	**77.1**
Lang. understanding	HellaSwag	42.1	62.9	77.7	84.1	**86.1**
	ARC-Challenge	29.7	39.2	58.3	67.7	**70.4**
	LAMBADA	44.5	57.3	72.2	75.8	**78.2**
QA	NQ	2.4	9.6	17.6	28.1	**31.5**
	TriviaQA	9.7	32.6	51.9	70.3	**72.3**
Comprehensive	AGIEval	25.4	26.0	34.8	45.8	**49.8**
	DROP (F1)	1.6	**3.0**	2.1	2.9	2.3
Mean		18.6	24.5	40.2	51.5	**56.1**

Table 12: Downstream performance of the Qwen2.5 models on the twelve benchmarks (%). All benchmarks report accuracy except DROP, which reports F1. The best score in each row is in bold.

Category	Benchmark	0.5B	1.5B	3B	7B	14B	32B	72B
Reasoning	MATH	8.6	13.3	13.9	25.0	29.2	35.8	**40.0**
	GSM8K	35.0	62.0	74.0	83.6	87.3	88.1	**91.0**
	GPQA Diamond	26.8	34.3	37.9	35.4	42.9	40.9	**46.0**
Knowledge	MMLU-Pro	15.3	29.1	38.0	48.9	56.5	61.4	**63.4**
	MMLU-STEM	43.5	55.0	62.0	71.1	77.7	82.9	**84.8**
Lang. understanding	HellaSwag	51.7	68.0	74.6	80.2	84.3	85.2	**87.6**
	ARC-Challenge	34.7	54.4	56.2	63.8	67.2	70.6	**72.4**
	LAMBADA	52.6	63.1	68.1	72.1	74.8	77.0	**78.1**
QA	NQ	4.8	7.8	16.4	23.2	26.0	25.7	**34.2**
	TriviaQA	12.7	29.4	42.2	50.4	59.6	61.7	**71.6**
Comprehensive	AGIEval	28.0	40.2	48.9	56.2	63.9	69.5	**72.7**
	DROP (F1)	5.1	9.8	14.8	24.4	23.3	**46.8**	30.9
Mean		26.6	38.9	45.6	52.9	57.7	62.1	**64.4**

Table 13: Downstream performance of the Qwen3 models on the twelve benchmarks (%). All benchmarks report accuracy except DROP, which reports F1. The best score in each row is in bold.

Category	Benchmark	0.6B	1.7B	4B	8B	14B
Reasoning	MATH	11.5	17.7	27.8	31.1	**36.4**
	GSM8K	48.5	71.6	84.5	84.9	**87.6**
	GPQA Diamond	29.8	33.3	37.9	**40.9**	26.3
Knowledge	MMLU-Pro	25.8	37.4	51.3	56.3	**61.9**
	MMLU-STEM	49.7	61.1	75.3	78.6	**82.5**
Lang. understanding	HellaSwag	53.6	67.1	75.7	79.6	**82.5**
	ARC-Challenge	45.4	55.6	64.3	67.5	**70.1**
	LAMBADA	55.3	65.2	70.6	73.7	**75.7**
QA	NQ	3.4	8.7	11.7	22.9	**25.0**
	TriviaQA	2.3	28.8	40.6	52.7	**59.3**
Comprehensive	AGIEval	32.4	40.6	54.7	60.1	**65.1**
	DROP (F1)	7.8	11.6	21.3	15.7	**36.5**
Mean		30.5	41.6	51.3	55.4	**59.1**

G DOWNSTREAM CORRELATION

In this section we share the detailed per benchmark Spearman correlation of each benchmark for each model family.

Table 14: Per-benchmark Spearman correlation of each metric with downstream accuracy across Gemma 3 model sizes (270M, 1B, 4B, 12B, 27B) profiled on FineWeb.

Category	Benchmark	Loss	α-ReQ	RankMe	Φ-Align
Reasoning	MATH	-1.000	-0.400	+0.800	-1.000
	GSM8K	-1.000	-0.400	+0.800	-1.000
	GPQA Diamond	-0.872	-0.154	+0.821	-0.872
Knowledge	MMLU-Pro	-1.000	-0.400	+0.800	-1.000
	MMLU-STEM	-0.900	-0.200	+0.900	-0.900
Lang. understanding	HellaSwag	-1.000	-0.400	+0.800	-1.000
	ARC-Challenge	-1.000	-0.400	+0.800	-1.000
	LAMBADA	-1.000	-0.400	+0.800	-1.000
QA	NQ	-1.000	-0.400	+0.800	-1.000
	TriviaQA	-1.000	-0.400	+0.800	-1.000
Comprehensive	AGIEval	-1.000	-0.400	+0.800	-1.000
	DROP	-0.300	-0.500	+0.000	-0.300

Table 15: Per-benchmark Spearman correlation of each metric with downstream accuracy across Gemma 3 model sizes (270M, 1B, 4B, 12B, 27B) profiled on Dolma.

Category	Benchmark	Loss	α-ReQ	RankMe	Φ-Align
Reasoning	MATH	-1.000	-0.700	+0.800	-1.000
	GSM8K	-1.000	-0.700	+0.800	-1.000
	GPQA Diamond	-0.872	-0.564	+0.821	-0.872
Knowledge	MMLU-Pro	-1.000	-0.700	+0.800	-1.000
	MMLU-STEM	-0.900	-0.600	+0.900	-0.900
Lang. understanding	HellaSwag	-1.000	-0.700	+0.800	-1.000
	ARC-Challenge	-1.000	-0.700	+0.800	-1.000
	LAMBADA	-1.000	-0.700	+0.800	-1.000
QA	NQ	-1.000	-0.700	+0.800	-1.000
	TriviaQA	-1.000	-0.700	+0.800	-1.000
Comprehensive	AGIEval	-1.000	-0.700	+0.800	-1.000
	DROP	-0.300	-0.300	+0.000	-0.300

Table 16: Per-benchmark Spearman correlation of each metric with downstream accuracy across Qwen2.5 model sizes (0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B) profiled on FineWeb.

Category	Benchmark	Loss	α-ReQ	RankMe	Φ-Align
Reasoning	MATH	-1.000	+0.821	-0.964	-0.964
	GSM8K	-1.000	+0.821	-0.964	-0.964
	GPQA Diamond	-0.929	+0.607	-0.857	-0.821
Knowledge	MMLU-Pro	-1.000	+0.821	-0.964	-0.964
	MMLU-STEM	-1.000	+0.821	-0.964	-0.964
Lang. understanding	HellaSwag	-1.000	+0.821	-0.964	-0.964
	ARC-Challenge	-1.000	+0.821	-0.964	-0.964
	LAMBADA	-1.000	+0.821	-0.964	-0.964
QA	NQ	-0.964	+0.750	-0.893	-0.893
	TriviaQA	-1.000	+0.821	-0.964	-0.964
Comprehensive	AGIEval	-1.000	+0.821	-0.964	-0.964
	DROP	-0.929	+0.821	-0.964	-0.964

Table 17: Per-benchmark Spearman correlation of each metric with downstream accuracy across Qwen3 model sizes (0.6B, 1.7B, 4B, 8B, 14B) profiled on FineWeb.

Category	Benchmark	Loss	α-ReQ	RankMe	Φ-Align
Reasoning	MATH	-1.000	+1.000	-1.000	-1.000
	GSM8K	-1.000	+1.000	-1.000	-1.000
	GPQA Diamond	+0.000	+0.000	+0.000	+0.000
Knowledge	MMLU-Pro	-1.000	+1.000	-1.000	-1.000
	MMLU-STEM	-1.000	+1.000	-1.000	-1.000
Lang. understanding	HellaSwag	-1.000	+1.000	-1.000	-1.000
	ARC-Challenge	-1.000	+1.000	-1.000	-1.000
	LAMBADA	-1.000	+1.000	-1.000	-1.000
QA	NQ	-1.000	+1.000	-1.000	-1.000
	TriviaQA	-1.000	+1.000	-1.000	-1.000
Comprehensive	AGIEval	-1.000	+1.000	-1.000	-1.000
	DROP	-0.900	+0.900	-0.900	-0.900