DATASET ALIGNMENT OVER DIVERSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling pre-training data volume has become the industry standard for improving language model performance, yet this approach is computationally wasteful when the goal is strong downstream performance on a specific task, especially when models are often pre-trained on diverse datasets rather than specialized training on datasets with overlapping domain coverage. We challenge this paradigm and propose Dataset Alignment Over Diversity (DAOD), a less compute dependent framework that predicts how well a pre-training dataset will transfer to a target downstream task without requiring any model training. DAOD scores candidates using a closed-form function of three geometric properties extracted from a randomly initialized model: centroid alignment between candidate and target embeddings, alignment friction (variance), and intra-dataset diversity. Across a benchmark of 8 candidate datasets evaluated against a mathematical reasoning target (OpenWebMath) and a second target validated across three architectures (Gemma-3 1B, LLaMA-3.2 1B, LFM-1.2B), DAOD reliably sorts candidates into the correct pre-training tier, capturing 95–98% of the achievable downstream performance gain while using only 25–38% of the total pre-training compute. The underlying score also shows strong rank correlation with ground-truth fine-tuning loss (maximum Spearman $\rho = 0.93$ on OpenWebMath), though we note this correlation is computed over a modest number of candidates and is best interpreted alongside the tiering results, which are more robust to individual rank swaps. Together these results establish that data alignment, not volume, is the primary driver of efficient transfer.

1 INTRODUCTION

Scaling pre-training data volume has become the default strategy for improving language model performance. This approach, while effective in aggregate, is often computationally wasteful when the actual goal is strong performance on a specific downstream task, since models are typically pre-trained on broad, diverse corpora rather than on data selected for its relevance to that task. Deciding which pre-training data to use for a given downstream target is therefore a critical, and expensive, decision. In practice this decision is usually made heuristically, guided by intuition about domain similarity or by trial and error, since evaluating a candidate dataset's true usefulness normally requires actually pre-training on it and measuring downstream performance, an approach that does not scale to the large pools of candidate datasets available today.

Existing techniques for improving training efficiency, such as transfer learning and curriculum learning, help reduce total training time once a data source has been chosen, but they do not address the upstream question of which data to choose in the first place. Prior transferability estimation methods typically require either a trained or partially trained model, or handcrafted features tied to a specific task setting, both of which reintroduce the compute cost that dataset selection should ideally avoid.

In this work, we ask a more targeted question: can a candidate pre-training dataset's usefulness for a given downstream task be estimated cheaply, without any model training at all, in a way that is good enough to guide practical selection decisions? We do not aim to produce a perfect predictor of fine-tuning loss. Instead, our goal is a lightweight scoring framework that reliably separates useful candidate datasets from unhelpful ones, so that practitioners can allocate their pre-training compute to the data most likely to help, before committing to expensive training runs.

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.

We propose Dataset Alignment Over Diversity (DAOD), a near-zero-compute framework that scores candidate pre-training datasets using three geometric properties extracted from a single forward pass through a randomly initialized model: centroid alignment between candidate and target embeddings, alignment friction, and intra-dataset diversity. Because DAOD requires no gradient computation or parameter updates, scoring a candidate dataset costs a small fraction of what pre-training on it would cost, making it practical to evaluate many candidates before selecting any of them for training.

We evaluate DAOD on a benchmark of 8 candidate pre-training datasets against two downstream targets, mathematical reasoning (OpenWebMath) and news summarization (XSum), across three model architectures (Gemma-3 1B, LLaMA-3.2 1B, and LFM-1.2B). We show that a DAOD-guided tiered selection strategy captures 95 to 98 percent of the achievable downstream performance improvement while using only 25 to 38 percent of the pre-training compute that would be required to train on all candidates, and that the underlying score also correlates strongly with ground-truth fine-tuning loss rankings. We further show that including low-scoring, poorly aligned datasets in the pre-training mixture can actively degrade downstream performance, reinforcing that dataset alignment, not raw data volume, is the primary driver of efficient transfer.

Our contributions are as follows. First, we introduce a closed-form, zero-compute scoring function that combines alignment, friction, and diversity to rank candidate pre-training datasets. Second, we propose a practical tiered selection strategy built on this score and show that it recovers most of the achievable pre-training benefit at a fraction of the compute. Third, we validate this approach across two target domains and three model architectures, and provide a preliminary cross-modal extension to vision models. Finally, we characterize the limitations of the approach, including its sensitivity to format mismatch between candidate and target data and its lack of a principled rule for setting the diversity weighting term, to give practitioners a clear picture of where the method can and cannot be trusted.

2 METHOD

We estimate the usefulness of a pre-training dataset using three quantities: structural-semantic alignment, optimization friction, and intra-dataset diversity.

Structural-Semantic Alignment (c): This measures the geometric coherence between the pre-training manifold and the downstream task. It quantifies both the syntactical topology and the conceptual proximity of latent structural vectors, effectively measuring whether a candidate dataset points toward the downstream task in representation space. For each dataset, we embed sampled sequences using the model's final hidden states and mean-pool over non-padding tokens to obtain one vector per sequence. We then compute a centroid for the target downstream dataset and for each candidate pre-training dataset. The alignment term is defined as:

$$c = \cos(\mathbf{h}_D, \mathbf{h}_T) \tag{1}$$

where $\mathbf{h}_D$ is the candidate dataset centroid and $\mathbf{h}_T$ is the target dataset centroid.

Optimization Friction (μ): Alignment alone can overestimate datasets that share surface-level tokens with the target while lacking coherent task structure. We therefore define optimization friction as the resistance generated by high-entropy, contradictory, or stochastic structural manifolds within the pre-training data. Unlike alignment, which depends on proximity to the target, friction depends on the quality and structure of the pre-training data itself; syntactic but semantically false data generates less friction than pure entropic noise. We define this friction term as:

$$\mu = \mathrm{Var}\left(\cos(\mathbf{h}_i^D, \mathbf{h}_T)\right) \tag{2}$$

where $\mathbf{h}_i^D$ is the embedding of an individual sample from the candidate dataset. This term penalizes datasets whose examples align inconsistently with the target centroid.

Intra-Dataset Diversity (σ_D): While alignment captures how well a candidate dataset points toward the target, it does not capture how broadly the dataset covers the target distribution. A dataset may have high alignment but consist of near-duplicate samples, offering limited coverage of the downstream task's structural variety. We define the diversity term as:

$$\sigma_D = \mathrm{Std}\left(\left\{\cos(\hat{\mathbf{h}}_i^D, \hat{\mathbf{h}}_j^D)\right\}_{i<j}\right) \tag{3}$$

where $\hat{\mathbf{h}}_i^D$ are ℓ_2-normalized sample embeddings from the candidate dataset. This measures the spread of pairwise similarities within the dataset; higher diversity indicates greater structural coverage. A dataset of varied problem types or writing styles will exhibit higher σ_D than a dataset of repetitive templates.

Final Score: The dataset score combines all three terms:

$$S = (c - \mu) \cdot (1 + \alpha \cdot \sigma_D) \tag{4}$$

The base score $(c - \mu)$ favors datasets that are close to the target task and internally consistent in how their samples point toward that task. The diversity multiplier $(1 + \alpha \cdot \sigma_D)$ provides a bonus proportional to structural coverage, rewarding datasets that align with the target across a wider variety of samples. In our primary evaluation, we set α=1. However, we note that for certain broad downstream tasks, ablating the diversity term (setting α=0) can yield more stable rankings.

Gaussian Probe: Critically, all embeddings are extracted from a *randomly initialized* model: no pre-training or fine-tuning is required. The model's weights are drawn from the standard initialization distribution $\theta \sim \mathcal{N}(0, \sigma_{\text{init}}^2)$. This makes the scoring procedure truly zero-compute: the only cost is a single forward pass per dataset, with no gradient computation or parameter updates. We show empirically that geometric properties of the random embedding space are sufficient to predict downstream transfer quality. We strongly recommend averaging scores of multiple seed varying and model varying results to obtain more consistent results.

The number of batches used to score, increases the stability of results proportionally, till it plateaus at around 250 samples

2.1 Practical Usage: Tiered Dataset Selection

Rather than relying on exact rank orderings, we propose a tiered selection strategy that partitions candidate datasets into three priority groups based on their DAOD scores. Given a set of n candidate datasets scored against a target, we define:

- **Tier 1 (High Alignment):** Datasets whose scores fall within the top cluster. These datasets are strongly recommended for pre-training and capture the majority of transferable structure.
- **Tier 2 (Moderate Alignment):** Datasets with intermediate scores. These provide diminishing marginal returns when added to a Tier 1 mixture and may be included if additional training volume is desired.
- **Tier 3 (Low Alignment):** Datasets whose scores fall significantly below the population mean. These datasets are structurally misaligned with the target and should be *excluded* from the pre-training mixture. Our experiments show that including Tier 3 data actively even degrades downstream performance relative to training on Tiers 1 and 1+2 alone.

The recommended workflow is as follows:

1. **Score:** Initialize the target model architecture with random weights. Extract embeddings for the target dataset and each candidate. Compute S for every candidate.
2. **Partition:** Rank candidates by S and assign tiers. We find that natural score gaps typically emerge between tiers, making the partition intuitive. When gaps are ambiguous, we recommend assigning the top 35–40% of candidates to Tier 1, the next 35–40% to Tier 2, and the rest to Tier 3.
3. **Select:** Pre-train on Tier 1 datasets. If Tier 1 data volume is insufficient, extend to Tier 2. Including Tier 3 turns out to be a mixed bag as per our research providing very minimal help and sometimes even hurting the model.

This tiered approach is motivated by a key empirical finding: in our ablation studies, pre-training on Tier 1 alone captures 95% of the total performance improvement achievable with all candidate data, while using only 38% of the compute. Furthermore, including Tier 3 data consistently degrades performance, the loss after training on all three tiers is higher than training on Tiers 1+2, confirming that more data is not unconditionally better and that alignment-guided selection eliminates harmful noise from the pre-training mixture.

3 EXPERIMENTAL SETUP

We evaluate our scoring framework on 1B-scale language models across multiple candidate pre-training datasets and two distinct downstream targets: **OpenWebMath** (mathematical reasoning) and **XSum** (news summarization). For OpenWebMath, the candidate pre-training datasets include domain-aligned sets (NuminaMath-CoT, GSM8K (Cobbe et al., 2021), MATH-lighteval), code (CodeParrot), structured data (Text-to-SQL), and off-distribution text (WikiText, CC-News, TinyStories). For XSum, candidates include CC-News, Wikipedia, WikiText, ag_news, NuminaMath-CoT, MATH-lighteval, CodeParrot, and SciQ.

We test across three model families: Gemma-3-1B-IT (Gemma Team, 2025), LLaMA-3.2-1B (Grattafiori et al., 2024), and LFM-1.2B (Liquid AI, 2024). For each model and candidate dataset, we sample text sequences via streaming, tokenize them to a maximum length of 512, and extract the final hidden-state embeddings. We mean-pool these embeddings over the attention mask to prevent padding tokens from corrupting the sequence representations.

For each candidate dataset, we compute the alignment term (c), the optimization friction (μ), the intra-dataset diversity (σ_D), and the final score (S). We evaluate predictive validity by computing the Spearman rank correlation between our zero-compute dataset scores and the empirical rankings derived from actual downstream fine-tuning loss.

Gaussian Probe (Zero-Compute): All representations are extracted from a randomly initialized model using the target architecture's configuration. The model's weights are drawn from the standard initialization distribution, and no training of any kind is performed during scoring. This constitutes a truly zero-compute evaluation: each candidate dataset requires only a single forward pass per sequence with no gradient updates.

Training Validation Protocol: To obtain ground-truth rankings, we pre-train the randomly initialized model on each candidate dataset individually (500 steps, learning rate 1×10^{-4}) and then fine-tune on the target dataset (500 steps, learning rate 5×10^{-5}). The resulting fine-tuning loss serves as the empirical measure of transfer quality. Lower fine-tuning loss indicates better pre-training transfer.

Baseline Comparisons: We compare DAOD against three established transferability estimation methods adapted for the generative setting: CKA (Kornblith et al., 2019), LEEP (Nguyen et al., 2020), and simple vocabulary overlap. Since LEEP was originally designed for classification transfer, we adapt it to the language modeling setting, LEEP computes the KL divergence between candidate and target unigram token distributions.

Tiered Ablation: To evaluate the practical utility of DAOD-guided dataset selection, we conduct a tiered ablation study. Datasets are partitioned into three tiers based on their DAOD scores, and we progressively expand the pre-training mixture from Tier 1 (highest-scoring) through Tier 3 (lowest-scoring). This tests the hypothesis that alignment-guided selection achieves most of the pre-training benefit while using a fraction of the total data.

All experiments were conducted on a single NVIDIA L4 (24GB VRAM). By utilizing streaming datasets and computing metrics over a bounded sample limit (500 sequences), our score computation remains extremely lightweight.

Reproducibility: We provide the complete evaluation framework and downstream fine-tuning loss curves for all candidate datasets across all tested architectures in the supplementary material.

4 RESULTS

4.1 DATASET SCORE PREDICTS DOWNSTREAM LOSS RANKING

Across two target domains and three 1B-scale language models, the DAOD score reliably identifies the top-performing tier of pre-training datasets for downstream fine-tuning. On the OpenWebMath target, DAOD achieves a Spearman rank correlation of $\rho = 0.93$ ($p < 0.001$) with ground-truth loss rankings, correctly identifying NuminaMath-CoT as the optimal pre-training dataset; we note this correlation is computed over a modest number of candidates (n=8) and is best interpreted alongside the tiering results in §4.4, which are more robust to individual rank swaps. On the XSum

target, DAOD achieves 100% top-1 accuracy across all three architectures tested: Gemma-3-1B-IT, LLaMA-3.2-1B, and LFM-1.2B, each correctly identifying CC-News as the best pre-training source for news summarization.

This cross-architecture consistency demonstrates that the geometric signal captured by DAOD is not an artifact of any particular model family but reflects genuine structural properties of the data.

4.2 Alignment Is the Dominant Signal

We find that the alignment term c explains most of the ranking structure. Datasets that are structurally closer to the downstream task, such as mathematical reasoning datasets for mathematical targets or news corpora for summarization targets, consistently receive higher scores and lower downstream loss. The friction term μ acts as a correction rather than the dominant predictor: it penalizes datasets whose apparent alignment is unstable across samples. This is useful for filtering high-entropy datasets that may share surface-level tokens with the target but do not provide coherent task structure.

The diversity term σ_D acts as a domain-sensitive amplifier. On the specialized OpenWebMath target, diversity boosts ρ from 0.81 to 0.93, correctly promoting NuminaMath above CC-News: a distinction the alignment term alone cannot make, since both score similarly on c. However, on the broader XSum target, diversity provides no additional benefit ($\rho = 0.62$ vs. 0.79 without diversity), as broad-domain targets already span diverse structural modes. We set $\alpha = 1$ for all reported results; practitioners targeting broad domains may prefer $\alpha = 0$.

Table 1: Dataset ranking under Gemma Gaussian representations. Lower average loss is better. The proposed score S ranks NuminaMath highest, matching the best empirical downstream performance.

Dataset	Avg. Loss	Actual Rank	c	μ	σ_D	S	Score Rank
NuminaMath	5.0652	1	0.8750	0.00528	0.0894	0.9475	1
MATH-lighteval	5.1559	2	0.8672	0.00412	0.0942	0.9444	3
CodeParrot	5.2618	3	0.8320	0.00836	0.1296	0.9305	4
CC-News	5.3112	4	0.8750	0.00380	0.0849	0.9452	2
GSM8K	5.5352	5	0.8594	0.00317	0.0724	0.9182	5
WikiText	5.6062	6	0.7891	0.00275	0.1312	0.8895	6
TinyStories	5.6246	7	0.6680	0.00160	0.1760	0.7836	7
Text2SQL	5.7021	8	0.6758	0.00224	0.1347	0.7643	8

Figure 1 visualizes the fine-tuning loss trajectories for all eight candidate datasets on Gemma-3-1B-IT, showing that the effect of dataset choice persists throughout training rather than only at the final step.

4.3 Comparison to Existing Methods

We compare DAOD against three established transferability estimation methods adapted for the generative setting (Table 2). On OpenWebMath:

Table 2: Baseline comparison on OpenWebMath target (Gemma-3-1B-IT). ρ: Spearman rank correlation with ground-truth fine-tuning loss. Top-1: whether the method correctly identifies the best pre-training dataset.

Method	ρ	p-value	Top-1
DAOD (ours)	**+0.929**	0.0009	✓
LEEP (token KL)	+0.905	0.0020	✓
Cosine Only (c)	+0.810	0.0149	×
Vocab Overlap	+0.643	0.0856	×
CKA	+0.619	0.1017	×

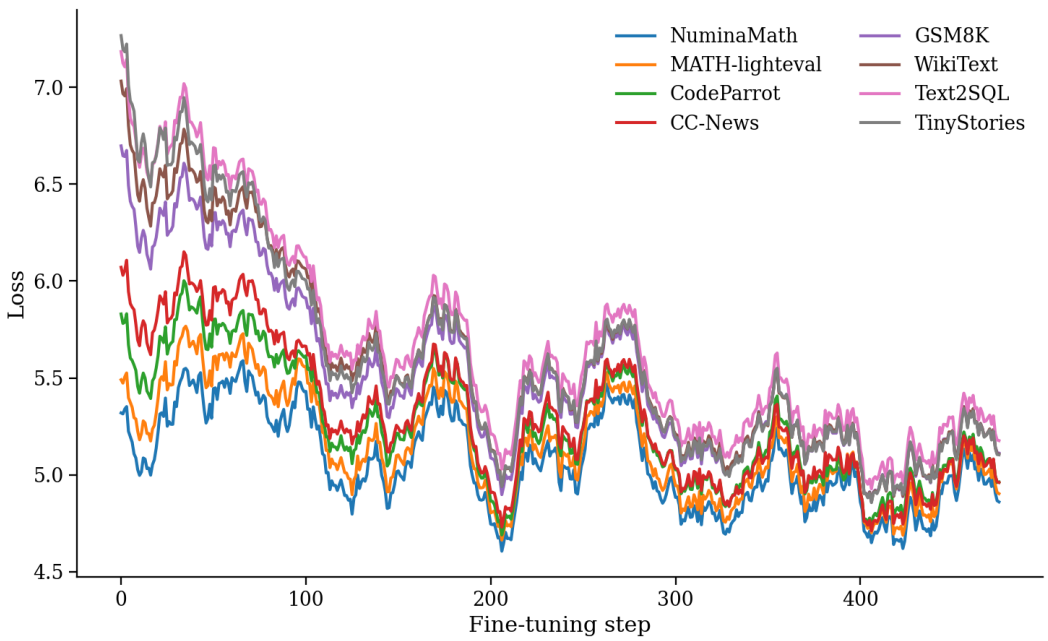


Figure 1: Gemma fine-tuning loss curves across all eight candidate pre-training datasets on OpenWebMath. Curves are smoothed with a 25-step moving average for visualization only. High-alignment datasets such as NuminaMath and MATH-lighteval maintain lower loss trajectories throughout training, indicating that dataset choice affects the full optimization path rather than only the final loss.

DAOD achieves the highest positive correlation ($\rho = 0.93$, $p < 0.001$) and is the only method to both rank datasets accurately and identify the correct top-1 candidate. LEEP (adapted to unigram token KL divergence) achieves comparable ρ but requires computing token frequency distributions, whereas DAOD operates directly on embeddings. CKA is statistically non-significant for dataset ranking at this scale.

4.4 Tiered Selection Captures Most of the Benefit at a Fraction of the Compute

We evaluate the practical utility of DAOD-guided dataset selection through a tiered ablation (Table 3). Datasets are partitioned into three tiers by DAOD score, and we progressively expand the pre-training mixture.

On OpenWebMath, Tier 1 alone (3 of 8 datasets, 38% of compute) captures 95% of the total achievable improvement. Adding Tier 3 data *degrades* performance, confirming that low-alignment data introduces harmful noise. On XSum, Tier 1 (2 of 8 datasets, 25% of compute) captures 98% of gains. These results demonstrate that DAOD enables practitioners to save 62–75% of pre-training compute while retaining nearly all downstream benefit.

Notably, on XSum the full Tier 1+2+3 mixture achieves the lowest loss overall (100% gain), unlike OpenWebMath where adding Tier 3 hurts - consistent with the diversity term's domain-dependence discussed in §5: broad targets tolerate, and here even benefit from, wider data mixtures that specialized targets do not.

4.5 Preliminary Cross-Modal Validation on ViT-B/16

To test whether the alignment-friction score is limited to language models, we also run a preliminary vision experiment using ViT-B/16. Candidate priming datasets are constructed from visually distinct subsets: aligned vehicle classes, misaligned flower classes, orthogonal people classes, and entropic noise. For this preliminary cross-modal check we report the base score $c - \mu$ only, since the diversity

Table 3: Tiered ablation on OpenWebMath and XSum (Gemma-3-1B-IT). "% Gain" is the fraction of total achievable improvement captured relative to the no-pretrain baseline; "% Compute" is the fraction of total pre-training compute used, proportional to the number of datasets in the mixture since each dataset is pre-trained for an equal number of steps.

Stage	OpenWebMath			XSum		
	Avg Loss	% Gain	% Compute	Avg Loss	% Gain	% Compute
No pretrain	6.473	-	0%	5.933	-	0%
Tier 1 only	4.671	95%	38%	4.261	98%	25%
Tier 1 + 2	4.524	100%	75%	4.275	97%	75%
Tier 1 + 2 + 3	4.570	98% ↓	100%	4.230	100%	100%

OpenWebMath tiers: Tier 1 = {NuminaMath, CC-News, MATH-lighteval}; Tier 2 additions = {CodeParrot, GSM8K, WikiText}; Tier 3 additions = {TinyStories, Text2SQL}.
XSum tiers: Tier 1 = {CC-News, Wikipedia}; Tier 2 additions = {NuminaMath, MATH-lighteval, WikiText, SciQ}; Tier 3 additions = {CodeParrot, ag_news}.

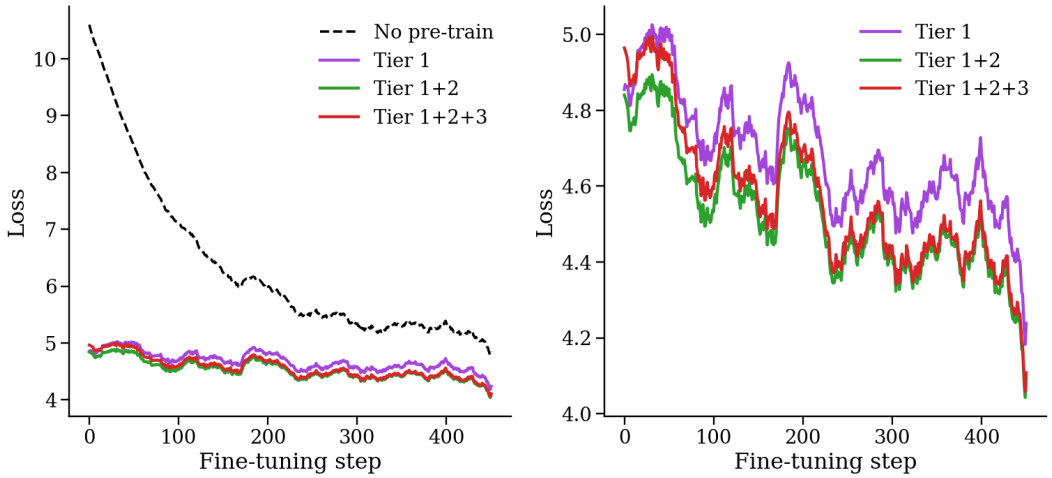


Figure 2: Fine-tuning loss curves on OpenWebMath across tiered pre-training mixtures. Left: full range including the no-pretrain baseline. Right: zoomed comparison of the three tiers. Tier 1+2 and Tier 1+2+3 closely track each other and both outperform Tier 1 alone throughout training however requiring way more compute, consistent with Table 3.

term σ_D was not evaluated in the visual domain. The score $c - \mu$ ranks the aligned vehicle subset highest, followed by people, flowers, and entropic noise. This ranking matches the empirical fine-tuning behavior: the score achieves a Spearman correlation of 0.80 with final loss ranking, and a correlation of 1.00 when using average loss or last-50-step loss. The aligned vehicle subset also achieves substantially lower loss than the non-primed baseline, reducing final loss by 59.0% and average loss by 39.7%. This preliminary result suggests that the proposed score may capture a broader transfer signal beyond language, although larger vision experiments (including with the full S formula) are needed.

5 Limitations

While our method is an effective zero-compute dataset selection system, it does present a few drawbacks. Users must be aware of these while using our method.

One of the major issues with this approach is that it is often biased against non-standard formats of datasets. Non-standard here refers to the comparison of format to the target dataset. For example, a dataset using a Human: Assistant format will be at a disadvantage against a web-scraped target dataset, even if both are intra-domain. This puts datasets that behave like Tier 1 into Tier 2. On

Table 4: Preliminary ViT-B/16 cross-modal validation using the base score $c - \mu$ (diversity term not evaluated in this domain). Lower average loss is better. The score ranks the aligned vehicle subset highest, matching the best empirical performance.

Dataset	c	μ	$c - \mu$	Avg. Loss
Vehicles	0.9879	0.0141	0.9738	0.4814
People	0.9053	0.0163	0.8889	0.6223
Flowers	0.7352	0.0170	0.7182	0.6885
Entropic Noise	0.6806	0.0014	0.6791	0.8146

the other hand, it gives a slight advantage to datasets that have a similar format to the target dataset, potentially pushing a Tier 2 dataset into Tier 1 when both the target and pre-training data are web-scraped.

We saw this directly in our Alpaca experiment. HH-RLHF, a dialogue-formatted dataset structurally close to Alpaca, actually achieved the best real fine-tuning loss of all eight candidates (0.842 avg), beating every other dataset we tested. But DAOD scored it second-to-last (0.589), well below datasets like SciQ and WikiText that trained worse in practice. Since our score is built on content-level embedding alignment, it missed the format-level transfer benefit entirely: this is essentially the same bias described above, just showing up as underrating instead of overrating, because HH-RLHF's format match did not come with the kind of content alignment our score looks for.

Another flaw is that our score has no real notion of dataset size baked in. c, μ, and σ_D are all computed from a fixed sample of embeddings, so a dataset with only a handful of examples can score exactly the same as a much bigger dataset with similar distributional properties. So a dataset can score very high on the board and still only have a few hundred examples total: the formula has no way of knowing the difference. In practice we'd expect such small, high-scoring datasets should be complemented with additional data to avoid overfitting during pre-training, though we did not run experiments specifically targeting this failure mode.

A third limitation is that the diversity term σ_D itself is domain-dependent, and we do not have a principled way to set α ahead of time. Our formula ablation (Table 5) shows that adding σ_D helps substantially on OpenWebMath (ρ: 0.810 $\rightarrow$ 0.929) but actually hurts on XSum (ρ: 0.786 $\rightarrow$ 0.619). This makes sense in hindsight: diversity is a useful signal when the target domain is narrow and specialized, but becomes noise when the target is broad and heterogeneous. However, it means a user has to already have some intuition about how specialized their target is before deciding whether to use $\alpha = 1$ or $\alpha = 0$. We use $\alpha = 1$ in our primary results since OpenWebMath is the specialized case, but we cannot yet offer a rule for picking α automatically, and flag $\alpha = 0$ as the safer default when the breadth of the target domain is unclear.

Table 5: Ablation over scoring formula variants on two target domains. ρ: Spearman rank correlation with ground-truth fine-tuning loss. Top-1: correct identification of the best pre-training dataset. All models are Gemma-3-1B-IT with Gaussian initialization.

	OpenWebMath			XSum		
Variant	ρ	p	Top-1	ρ	p	Top-1
c (alignment only)	+0.810	0.015	$\times$	+0.786	0.021	✓
$c - \mu$	+0.810	0.015	$\times$	+0.786	0.021	✓
$(c - \mu)(1 + \sigma_D)$	**+0.929**	**0.001**	✓	+0.619	0.102	✓

6 Related Work

Transfer Learning and Compute Efficiency: Transfer learning often relies on heuristic dataset selection or broad scaling of parameters and data (Pan & Yang, 2010; Kaplan et al., 2020). While upstream pre-training can improve downstream performance, the relationship between the source

dataset and target task is often evaluated only after training. Our work studies this dataset-selection problem directly by defining three simple geometric quantities: structural-semantic alignment c, optimization friction μ, and intra-dataset diversity σ_D. Rather than treating transfer as a fully explained mechanism, we use these quantities as an empirical proxy for estimating dataset usefulness before full optimization.

Curriculum Learning: Our framework also provides a geometric explanation for the efficacy of curriculum learning (Bengio et al., 2009). Curriculum learning posits that models converge faster when trained on structurally simpler concepts before complex ones. Under our view, this occurs because starting with high-alignment, low-friction datasets establishes a favorable initial manifold. By prioritizing datasets with a high S score, we are effectively automating a curriculum that minimizes structural resistance (friction) while preserving useful structural coverage (diversity), allowing gradient updates to progress toward the target manifold without being derailed by entropic noise.

Catastrophic Forgetting: Recent work demonstrates that language models develop internalized world representations to solve complex reasoning tasks (Li et al., 2023). We propose that the formation of these internalized representations is analogous to constructing a building: pre-training on structurally isomorphic data lays the foundation and builds the structural framing, while fine-tuning on the downstream task is akin to decorating the completed structure.

If a model is instead pre-trained on a dataset that is structurally and semantically misaligned with the final task, the model cannot simply decorate the existing structure. It may need to overwrite representations acquired during pre-training before adapting to the downstream task, a destructive optimization process that is the mechanism driving catastrophic forgetting (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017). Pre-training on aligned data helps prevent this by ensuring the initial foundation naturally supports the final downstream task, which is consistent with why the alignment term c is the dominant signal in our score (§4).

7 Conclusion

We presented Dataset Alignment Over Diversity (DAOD), a near-zero-compute framework for ranking candidate pre-training datasets by downstream usefulness using only a randomly initialized model. By measuring three geometric properties of embedding distributions (centroid alignment, optimization friction, and intra-dataset diversity), DAOD produces a single score that turns an otherwise expensive dataset-selection problem into a single forward pass, yielding rankings that are practically usable for pre-training selection.

A tiered selection strategy guided by DAOD scores captures 95–98% of achievable pre-training benefit while using only 25–38% of the total compute, and we show that including low-scoring datasets can actively degrade performance, underscoring that data alignment, not volume, is the primary driver of efficient transfer. The underlying score also correlates strongly with ground-truth fine-tuning loss (Spearman $\rho = 0.93$ on OpenWebMath) and achieves 100% top-1 accuracy across three model architectures on XSum, though we view the tiering results as the more robust and practically relevant signal given the scale of our candidate pool.

These findings challenge the prevailing assumption that more data is unconditionally better, and offer practitioners a lightweight, architecture-agnostic tool for making informed dataset decisions before committing to expensive pre-training runs. We believe this work opens a broader line of inquiry into what geometric structure randomly initialized networks already encode, and how that structure can be exploited for efficient model development.

AI Use Statement

In this work, we used generative AI tools (large language models) for two purposes: (1) refining English prose in the manuscript for clarity and readability, and (2) improving code documentation and formatting in our evaluation scripts to support reproducibility. All experimental design, methodology, mathematical formulations, result interpretation, and scientific claims were produced entirely by the authors. All AI-assisted text and code were reviewed, edited, and verified by the human authors prior to inclusion. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

Ethics Statement

This work presents a method for evaluating pre-training dataset compatibility using publicly available datasets and model architectures. Our framework does not involve human subjects, generate synthetic content, or introduce new data collection practices. All datasets used are publicly released under permissive licenses. We do not foresee direct negative societal impacts from this work; however, we note that any dataset selection method inherits the biases present in the candidate and target corpora. Practitioners should ensure that selected pre-training data complies with applicable data governance and fairness standards.

Reproducibility Statement

All experiments in this paper use publicly available datasets from the Hugging Face Hub and publicly released model architectures (Gemma-3-1B-IT, LLaMA-3.2-1B, LFM-1.2B). Our scoring procedure requires only a single forward pass through a randomly initialized model with no training, making it straightforward to reproduce. The complete scoring and training scripts, including exact hyperparameters, dataset loading configurations, and evaluation pipelines, are provided in the supplementary material. Section 4 details all experimental settings: sequence length, learning rates, step counts, sample limits, and hardware used (single NVIDIA L4, 24GB). Ground-truth fine-tuning loss curves for all candidate–target pairs across all architectures are included in the supplementary material to enable independent verification of our reported rankings.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48, 2009.

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.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 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, 2019.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023.

Liquid AI. Lfm: Liquid foundation models. Technical report, Liquid AI, 2024.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pp. 7294–7305, 2020.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.