FLARE: Task-Agnostic Embedding Model Evaluation via Normalizing Flows

Jingzhou Jiang and Yixuan Tang and Yi Yang and Kar Yan Tam
The Hong Kong University of Science and Technology
{jjiang105,ytangch}@connect.ust.hk, {imyiyang,kytam}@ust.hk

Abstract

Selecting an embedding model for a specific target corpus is difficult when task-specific labels are unavailable. Existing label-free metrics based on kernel estimators or Gaussian mixtures fail in high-dimensional spaces and produce unstable rankings. We propose **FLARE** (**F**low-based **L**abel-free **A**ssessment of **R**epresentation **E**mbeddings), which uses normalizing flows to estimate information sufficiency directly from log-likelihoods, avoiding distance-based density estimates. We give finite-sample bounds showing that the estimation error depends on the intrinsic dimension of the data manifold rather than the raw embedding dimension. On 11 datasets and eight embedders, FLARE attains Spearman's ρ up to 0.90 with supervised benchmarks and remains stable for high-dimensional embeddings ($d \geq 3{,}584$), where existing label-free baselines collapse.

1 Introduction

Recent advances in text embeddings have produced powerful semantic representation models such as Qwen3 Embedding (Yang et al., 2025) and Gemini Embedding (Lee et al., 2025). However, as the number of available models grows, each with different architectures, training objectives, and pretraining corpora, selecting the most suitable model for a given corpus has become increasingly challenging. The standard approach relies on supervised benchmarks like MTEB (Muennighoff et al., 2023), ranking models by their performance on annotated tasks.

This approach requires labeled data, which is often unavailable in practice. Consider deploying a retrieval system over proprietary documents such as legal contracts, medical records, or financial reports. These collections have no existing labeled query-document pairs, and creating annotations requires significant time and domain expertise. Specialized corpora may also differ substantially from public benchmarks in vocabulary, style, and topic distribution (Tang and Yang, 2025). An embedding model that ranks highly on MTEB may perform poorly on domain-specific text, but without labels we cannot measure this gap. Benchmark contamination further undermines public leaderboards: as test sets appear in pretraining data, scores become inflated. This raises our central question: *how can we evaluate embedding models without labels?*

Recent work has explored task-agnostic evaluation using only unlabeled corpora. One approach analyzes geometric properties of embedding models like uniformity and alignment (Wang and Isola, 2020; Rudman et al., 2022). However, these metrics measure the embedding hypersphere structure rather than semantic content. A random projection can achieve perfect uniformity while preserving no information. A more principled alternative estimates mutual information between embeddings, quantifying how much information the embedding retains. Existing implementations use nonparametric estimators like Kernel Density Estimation (KDE) or Gaussian Mixtures (Darrin et al., 2024). These methods suffer from the curse of dimensionality (Beirlant et al., 1997): as embedding dimension grows (modern models often exceed $d = 3{,}000$), reliable density estimation requires exponentially more data. In high-dimensional space, these estimators become unstable and fail to predict downstream performance.

In this work, we propose **F**low-based **L**abel-free **A**ssessment of **R**epresentation **E**mbeddings (**FLARE**), a framework grounded in **information-theoretic sufficiency** (Darrin et al., 2024). Specifically, we quantify embedding quality by measuring the reduction in uncertainty about input data given the embedding. The core innovation lies in leveraging Normalizing Flows (Durkan et al., 2019), deep generative models that learn invertible transformations from complex distributions to simple base densities. Flows enable exact log-likelihood

Findings of the Association for Computational Linguistics: ACL 2026, pages 39271–39294

estimation via the change-of-variables formula, effectively mitigating the curse of dimensionality inherent to distance-based estimators. Our finite-sample bounds provide theoretical justification: the estimation error depends on the intrinsic effective dimension of the data manifold, not the embedding dimension. This ensures that FLARE remains reliable when scaling to the high-dimensional embeddings of modern LLMs.

We evaluate FLARE on 11 datasets across four task families. It attains Spearman's ρ up to 0.90 with supervised rankings, outperforms geometry-based and information-theoretic baselines, and remains stable for high-dimensional embeddings ($d \geq 3{,}584$) where existing methods fail.

Our contributions are summarized as follows:

- We introduce FLARE, a task-agnostic text embedding evaluation framework using normalizing flows to estimate information sufficiency without labeled data.
- We prove finite-sample generalization bounds whose dominant term scales with the intrinsic dimension of the data manifold rather than the raw embedding dimension, explaining the estimator's behaviour in high dimensions.
- Across 11 datasets and eight embedders, FLARE predicts downstream rankings more reliably than existing baselines, particularly for high-dimensional LLM-based embeddings.

2 Related Work

Embedding Evaluation. Modern text embedding models, often leveraging Large Language Model (LLM) architectures, provide high-dimensional representations that generalize across diverse semantic tasks without requiring task-specific fine-tuning (Neelakantan et al., 2022; Wang et al., 2024). Currently, the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) and its specialized variants like FinMTEB (Tang and Yang, 2025) serve as the primary evaluation standards. These benchmarks aggregate performance across diverse labeled tasks including clustering, retrieval, and semantic textual similarity (STS). All of them depend on ground-truth annotations and are therefore unsuitable when labels are absent, as is typical for proprietary, dynamic, or out-of-distribution corpora. To address this setting, we propose FLARE, a task-agnostic framework that quantifies embedding quality directly from the data distribution, offering a reliable performance proxy without the need for supervision.

Task-Agnostic Approaches. Task-agnostic metrics offer a long-standing alternative to supervised benchmarks by eliminating labeling costs. Classic indices such as the Silhouette Score (Rousseeuw, 1987) evaluate cluster separation, while more recent studies emphasize spectral and geometric properties. Examples include Effective Rank (Roy and Vetterli, 2007) for dimensionality estimation along with Uniformity (Wang and Isola, 2020) and IsoScore (Rudman et al., 2022) for spatial distribution analysis. A primary limitation of these metrics is their focus on geometric structure rather than semantic content. Such methods often rely on global priors like isotropy, which may not represent the intrinsic low-dimensional manifold structure characteristic of text embeddings. To address this, EMIR (Darrin et al., 2024) introduces the concept of Information Sufficiency to quantify how well one embedding model can reconstruct another. However, standard implementations of EMIR utilize Gaussian Mixture Models (GMM), which struggle with high-dimensional data and lack the generalization guarantees required for modern LLM-based embeddings.

Density Estimation. Calculating information-theoretic measures like differential entropy requires an accurate model of the underlying probability density. Traditional non-parametric approaches, notably Kernel Density Estimation (KDE), are fundamentally limited by the curse of dimensionality (Silverman, 2018; Beirlant et al., 1997). In high-dimensional spaces, these methods become statistically inefficient because the sample size needed to control estimation error grows exponentially with dimensionality. While neural variational estimators such as MINE (Ishmael Belghazi et al., 2018) and CLUB (Cheng et al., 2020) improve scalability, they optimize variational bounds instead of exact likelihoods. These bounds often suffer from a severe bias-variance trade-off, leading to loose estimates and optimization instability (Poole et al., 2019). Normalizing Flows provide a robust alternative by learning a sequence of invertible transformations that map complex data to a simple base distribution (Rezende and Mohamed, 2015; Dinh et al., 2016). A key advantage is their support for exact log-likelihood computation via the change-of-variables formula (Papamakarios et al., 2021).

Based on this property, we treat embedding evaluation as a problem of estimating probability densities directly. By integrating flows into the information-sufficiency framework (Darrin et al., 2024), we ensure that our metrics remain theoretically grounded and empirically stable even for high-dimensional embeddings.

3 Method

To evaluate embedding quality without task-specific labels, we propose FLARE, which quantifies representation quality by measuring information sufficiency using normalizing flows.

3.1 Problem Formulation

Given an unlabeled corpus $\mathcal{X}$, we consider a set of candidate embedding models $\mathcal{E} = \{E_1, \ldots, E_K\}$. Each model $E \in \mathcal{E}$ maps an input text $x \in \mathcal{X}$ to a high-dimensional representation:

$$z = E(x) \in \mathbb{R}^d. \tag{1}$$

To evaluate a specific model E_a, we pair it with a reference model E_b ($b \neq a$). We denote the embedding being evaluated as the source U, and the reference embedding as the target V:

$$U = E_a(x), \quad V = E_b(x). \tag{2}$$

Our objective is to derive a task-agnostic score for E_a based solely on these representations such that the resulting model ranking aligns with downstream supervised performance.

3.2 Information-Sufficiency Score

We build upon the information-sufficiency framework of (Darrin et al., 2024). The core intuition is that a high-quality source embedding U should act as a sufficient representation of the semantic space, enabling the reconstruction of the target representations V. We formalize this via Information-Sufficiency (I_s), which measures the reduction in uncertainty of V once U is observed.

Let $\mathcal{F}$ be a family of marginal densities and $\mathcal{K}$ be a family of conditional densities. The directional I_s score from U to V is defined as the difference between marginal and conditional entropy:

$$\mathrm{I_s}(U \to V) = \underbrace{\inf_{f \in \mathcal{F}} \mathbb{E}_v[-\log f(v)]}_{H(V)} - \underbrace{\mathbb{E}_u\left[\inf_{M \in \mathcal{K}} \mathbb{E}_{v|u}[-\log M(v \mid u)]\right]}_{H(V|U)}. \tag{3}$$

To obtain a single quality score for model E_a, we compute the normalized median of its pairwise scores against all other models in the pool:

$$\mathrm{I_{snorm}}(E_a) = \mathrm{median}_{b \neq a} \frac{\mathrm{I_s}(U_a \to U_b)}{\dim(U_b)}. \tag{4}$$

Normalization by the target dimension $\dim(U_b)$ is essential for comparability across reference models with varying output sizes, as raw entropy naturally scales with dimensionality.

3.3 Normalizing-Flow Implementation

We instantiate the density families in Eq. 3 using normalizing flows (Durkan et al., 2019). Flows enable exact log-likelihood computation via invertible transformations, providing stable estimation.

Two-Stage Training. As shown in Figure 1, we employ a progressive training strategy. We first train a marginal flow $p_\phi(v)$ on target embeddings to model distribution. Next, we initialize a conditional flow $p_\theta(v|u)$ by copying the marginal backbone weights. This warm-start strategy ensures the conditional model begins from a well-defined density baseline.

Low-Rank Conditioning. Standard conditional flows often use hypernetworks with $O(d^2)$ complexity, which is prohibitive for high-dimensional embeddings. We instead inject the source information u through a parameter-efficient low-rank residual branch. Let $\mathbf{h}_{\text{base}}$ be the intermediate features of the target flow. The conditional feature is computed as:

$$\mathbf{h}_{\text{cond}} = \mathbf{h}_{\text{base}} + B(A(u)), \tag{5}$$

where $A \in \mathbb{R}^{r \times d}$ projects the d-dimensional source embedding to a low-rank bottleneck of dimension $r = 64$, and B maps from the bottleneck to the output space. This mechanism allows the source u to adjust the transformation trajectory of the target v with minimal parameter overhead.

Zero Initialization. We initialize B to zeros and A with random initialization. At the start of training, the conditioning term $B(A(u))$ is zero, making the conditional flow $p_\theta(v|u)$ initially equivalent to the pre-trained marginal flow $p_\phi(v)$. The dependence on the source U is learned gradually, which prevents gradient instability and improves convergence speed.

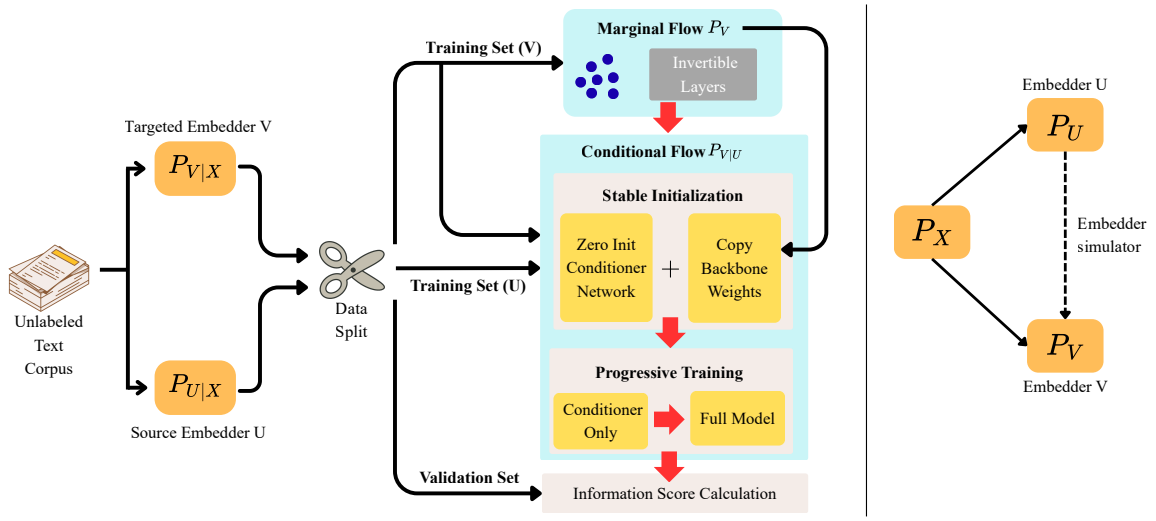

Figure 1: Overview of our two stage flow based estimation pipeline. **Stage 1:** Train a marginal flow $p_\phi(v)$ on target embeddings V to model their intrinsic distribution. **Stage 2:** Initialize a conditional flow $p_\theta(v|u)$ by copying the marginal backbone weights and adding a zero-initialized low-rank conditioning branch. This branch is then trained to capture the dependency between source embeddings U and target embeddings V, enabling computation of the information-sufficiency score via Eq. 3.

4 Theoretical Justification

Motivation. In real-world deployment, embedding models must generalize to vast amounts of unseen data that extend far beyond the validation set. Purely empirical evaluation on a fixed dataset cannot theoretically guarantee reliability on the underlying population distribution. To address this, we establish a theoretical framework relying on two pillars: the spectral stability of our flow architecture and the low-dimensional manifold hypothesis.

Core Assumptions. Our theory relies on two core assumptions (Appendix A.2): a low intrinsic dimension (Assumption 1) to enable scaling to high-dimensional embeddings, and layer-wise approximate independence (Assumption 2) to ensure stable gradient propagation. The latter, theoretically motivated by (Cohen et al., 2021), is practically realized by our Zero-Initialization strategy, which ensures the network exhibits stable, linear growth rather than exponential instability.

Finite-Sample Generalization Bound. Building on this stability, we verify the reliability of FLARE by bounding the gap between the training and validation losses.

Theorem 1 (Finite-sample generalization bound)**.** *Under Assumptions 1 and 2, for any fixed flow model p_θ and any $\delta \in (0,1)$, with probability at least $1-\delta$, the train-validation gap satisfies:*

$$\begin{aligned}\left|\hat{L}_{\text{val}}(\theta) - \hat{L}_{\text{train}}(\theta)\right| \leq \frac{2\tilde{C}_{\text{Rad}}\, L\, \bar{\sigma}\, \sqrt{d_{\text{eff}}}}{\sqrt{m}} \\ + M_{\text{val}}\sqrt{\frac{\log(2/\delta)}{2m_{\text{val}}}} \\ + 3M_{\text{train}}\sqrt{\frac{\log(2/\delta)}{2m}}.\end{aligned} \tag{6}$$

Interpretation. The bound scales with the intrinsic dimension d_{eff} rather than the raw embedding dimension d. Because $d_{\text{eff}} \ll d$ for semantic representations, reliable evaluation is possible on high-dimensional embeddings with moderate sample sizes. The error also depends linearly on depth L and spectral stability $\bar{\sigma}$, both kept small by our zero-initialization strategy.

5 Experiments

We design our experiments to evaluate the empirical utility of the Flow-based estimator across three primary dimensions: **(Q1)** its reliability in pre-

dicting model rankings across diverse downstream tasks; **(Q2)** its comparative performance against kernel-based baselines in high-dimensional embedding spaces; and **(Q3)** the alignment between empirical observations and our theoretical generalization guarantees.

5.1 Embedders and Datasets

Embedding Models. Our evaluation encompasses eight representative embedding models, covering a broad spectrum of architectures and dimensionalities. We focus specifically on high-dimensional space, including BGE-Multilingual-Gemma2 ($d = 3{,}584$) (Chen et al., 2024a), gte-Qwen2-7B-instruct ($d = 3{,}584$) (Li et al., 2023), and four Mistral-7B-based models ($d = 4{,}096$): Zeta-Alpha-E5-Mistral [1], GritLM-7B (Muennighoff et al., 2024), SFR-Embedding-Mistral (Meng et al., 2024), and Linq-Embed-Mistral (Kim et al., 2024). For lower-dimensional benchmarks, we include stella-base-en-v2 [2] ($d = 768$) and all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) ($d = 384$).

Dataset Selection and Task Categorization. To approximate the realistic setting in which models are evaluated on novel internal corpora, we deviate from standard benchmarks such as MTEB. Public benchmarks routinely leak into pre-training data and inflate reported scores, so we apply a temporal filter and select 11 Hugging Face datasets released *after* the training cutoff of every candidate model. FLARE itself is unsupervised; ground-truth labels are used only as an oracle to validate the predicted rankings. Dataset statistics are reported in Appendix B. The 11 datasets cover four task categories:

- **Classification:** Apt-eval (Saha and Feizi, 2025) (safety/robustness), GT-FintechLab (Shah et al., 2025) (finance), and BhashaBench-Finance (Devane et al., 2025) (multilingual finance).
- **Retrieval:** AIR-Bench-Finance (Chen et al., 2024b), LIMIT (Weller et al., 2025) (instruction-following), and ArXiv-Abstracts 2025 [3] (scientific literature). For retrieval tasks, we embed the passage corpus only (not queries) to simulate the label-free setting.
- **Semantic Textual Similarity (STS):** Augmented STS-B [4], LivNLP-STS (Zhang et al., 2025), and Philosophical-STS [5].
- **Clustering:** Clustered-FunPang Medical[6], and Reasoning-Clustering [7].

Evaluation Protocol. We assess the reliability of our unsupervised Information Sufficiency (IS) metric by measuring its alignment with ground-truth supervised rankings. Ground-truth performance is established using standard MTEB metrics (Muennighoff et al., 2023): F1 macro for classification, nDCG@10 for retrieval, Spearman correlation for STS, and V-measure for clustering. We quantify ranking alignment using Spearman's rank correlation (ρ) and Pearson correlation (r) between the predicted IS scores and the supervised metrics.

Baselines. We benchmark FLARE against existing unsupervised metrics, explicitly categorized into two types of unsupervised evaluation methods: geometric-based and information-theoretic. For the geometric-based methods, we considered (1) **Uniformity** (Wang and Isola, 2020) and (2) **IsoScore** (Rudman et al., 2022); and (3) **Silhouette Score** (Rousseeuw, 1987). For the information-theoretic methods, we considered (4) **EMIR** (Darrin et al., 2024). Our work aligns with the information-theoretic evaluation (specifically the framework established by EMIR), as we share the fundamental objective of quantifying representation quality via information retention. However, we diverge by employing normalizing flows to robustly estimate densities in high-dimensional spaces where the GMMs used in EMIR may fail.

5.2 Main Results

Table 1 summarizes the ranking correlations across eleven representative datasets, revealing three key findings. Detailed per-dataset results are shown in Appendix B.

[1] https://huggingface.co/zeta-alpha-ai/Zeta-Alpha-E5-Mistral

[2] https://huggingface.co/infgrad/stella-base-en-v2

[3] https://huggingface.co/datasets/almanach/arxiv_abstracts_2025

[4] https://huggingface.co/datasets/maiammar/Augmented_stsb_multi_mt

[5] https://huggingface.co/datasets/johnnyboycurtis/Philosophical-STS-Text-Pairs

[6] https://huggingface.co/datasets/mukulb/clustered_FUNPANG_dataset_with_groups

[7] https://huggingface.co/datasets/Ibisbill/Clustering_deduplicated_reasoning

	Spearman's ρ					Pearson's r				
Task	**Uni.**	**Iso.**	**Sil.**	**EMIR**	**Ours**	**Uni.**	**Iso.**	**Sil.**	**EMIR**	**Ours**
Class.	0.18	-0.40	-0.06	-0.06	**0.56**	**0.35**	-0.58	-0.08	-0.20	0.31
STS	0.01	-0.33	0.56	-0.06	**0.70**	**0.70**	-0.14	0.53	-0.08	**0.70**
Retr.	0.08	-0.45	-0.03	-0.22	**0.72**	0.43	-0.37	0.39	-0.18	**0.64**
Clust.	-0.14	0.29	-0.24	-0.16	**0.83**	-0.43	0.05	-0.25	-0.10	**0.69**
Avg	0.05	-0.27	0.08	-0.12	**0.69**	0.33	-0.29	0.18	-0.14	**0.58**

Table 1: Task-aggregated Comparison with Unsupervised Baselines.

Flow-Based Estimation Succeeds Where Kernel Methods Fail. FLARE achieves an average Spearman correlation of $\rho = 0.70$, outperforming all unsupervised baselines. EMIR, which shares our information-sufficiency framework but relies on Gaussian Mixture Model (GMM) density estimation, yields a negative average correlation ($\rho = -0.12$), i.e. systematically inverted rankings: in high-dimensional spaces ($d \geq 3{,}584$) kernel densities become vanishingly sparse and distance-based estimates unreliable. Our flow-based approach replaces these estimates with an explicit parametric density that adapts to the intrinsic manifold, and maintains positive correlations across all task categories.

Geometric Baselines Exhibit Inconsistent Performance. Table 1 reveals that existing unsupervised metrics struggle to maintain consistent correlations across task types. Uniformity achieves near-zero average correlation ($\rho = 0.05$), suggesting that embedding space uniformity alone is not predictive of downstream quality. IsoScore exhibits negative correlations across nearly all categories (Avg $\rho = -0.27$), as it penalizes anisotropy under the assumption that uniformity maximizes information capacity. However, high-quality semantic spaces are inherently anisotropic—meaningful concepts naturally form dense, non-uniform clusters on the manifold rather than populating the hypersphere uniformly. Silhouette shows task-specific success only on STS ($\rho = 0.56$) but fails elsewhere, indicating that geometric cohesion does not generalize as a universal quality indicator. As Figure 2 shows, these rigid geometric assumptions produce high variance and frequent ranking inversions. The information-theoretic objective behind FLARE instead adapts to the intrinsic data density rather than imposing a target shape on it, enabling FLARE to remain aligned with ground truth across heterogeneous task families.

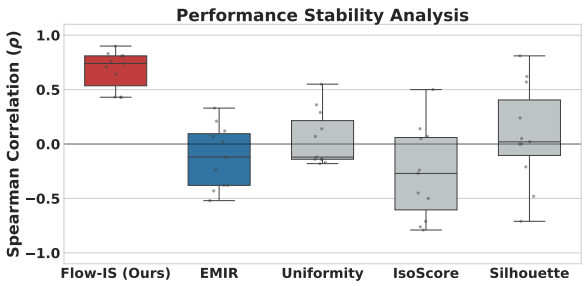

Figure 2: **Stability Analysis.** Distribution of Spearman correlations across all datasets. Geometric baselines (grey) and kernel-based EMIR (blue) exhibit high variance and frequent negative correlations, while FLARE (red) stays positively aligned with ground truth on every dataset.

Consistent Performance Across Task Families. A useful label-free metric must rank embedders across heterogeneous downstream tasks, not just on a single family. Existing baselines fail this requirement: each one has at least one task family on which it correlates negatively with ground truth (Table 1). FLARE correlates positively on every family, with $\rho = 0.83$ on Clustering, $\rho = 0.72$ on Retrieval, $\rho = 0.70$ on STS, and $\rho = 0.56$ on Classification, and the per-dataset breakdown in Table 7 (Appendix B) contains no negative entries. The signal is strongest on STS, Retrieval, and Clustering, which score embeddings directly through similarity in the embedding space. Classification involves an additional supervised predictor that can exploit features beyond the distributional structure captured by an unsupervised metric, which caps

Task Type	Bound Ratio	Rademacher %
Classification	11.0×	92.4%
STS	21.1×	94.5%
Retrieval	21.2×	95.5%
Clustering	18.4×	98.5%
Average	**17.9×**	**95.2%**

Table 2: **Theoretical validation.** Bound Ratio ($\Delta_{\text{theo}}/\Delta_{\text{emp}}$) and Rademacher complexity contribution, grouped by task type.

the achievable correlation.

Ranking Stability under Subsampling. We probe sample-size sensitivity by recomputing I_s on subsamples of the validation set at ratios $\alpha \in \{0.05, 0.1, \ldots, 1.0\}$ and measuring the deviation $\Delta_\rho(\alpha)=|\rho(\alpha)-\rho(1.0)|$ from the full-data ranking. For 8 of the 11 datasets, $\Delta_\rho < 0.05$ even at $\alpha=0.2$; the remaining three datasets are STS corpora where rank-based metrics are inherently more sample-sensitive, and even there $\Delta_\rho < 0.07$ once $\alpha \geq 0.2$. The flow therefore recovers the global manifold from far fewer observations than non-parametric estimators require, suggesting that 20–40% of the validation set is sufficient for reliable ranking on small, specialized corpora. Per-dataset stability curves are reported in Appendix C.

5.3 Generalization Bound Analysis

To empirically validate the guarantee provided by Theorem 1, we compare the derived theoretical bound against the observed empirical generalization gap. This validation is crucial for addressing the practical challenges of enterprise deployment, ensuring that the evaluation method generalizes reliably as new data continuously arrives.

Setup. Since the Information Sufficiency estimator decomposes into marginal and conditional components, its total estimation error is bounded by the sum of their respective generalization gaps. We therefore compute the empirical gap as the sum of the absolute differences between training and validation Negative Log-Likelihoods (NLL) for both flows, and compare this empirical quantity against the theoretical bound derived from our model architecture and sample complexity.

Analysis. As reported in Table 2, the theoretical bound Δ_{theo} upper-bounds the empirical gap Δ_{emp} by a margin ranging from $11.0\times$ to $21.2\times$ across task types (average ratio $17.9\times$). This confirms that our conservative linear bound remains informative, providing meaningful generalization guarantees without being vacuously loose. The Rademacher complexity term accounts for 92.4% to 98.5% of the total bound (average 95.2%), which aligns with its dependence on the effective intrinsic dimension d_{eff}. As expected, higher-dimensional embedding spaces incur larger complexity penalties. Notably, classification tasks exhibit tighter bounds ($11.0\times$) compared to retrieval and STS tasks ($\sim 21\times$), a discrepancy that reflects the richer representational capacity required for fine-grained semantic matching in the latter. Practically, these findings certify that FLARE provides a trustworthy and theoretically grounded signal on unseen data.

5.4 Ablation Study

Shuffle ablation. We partially shuffle the correspondence between source embeddings U and target embeddings V at a ratio $p \in [0, 1]$ while keeping the remaining pairs unchanged. As visualized in Figure 3 and detailed in Appendix D.1, the Spearman correlation with downstream performance degrades significantly as p increases, transitioning from positive to negative around $p = 0.2$ to 0.4. The degradation pattern varies across task types: Retrieval and Clustering exhibit a gradual decline, whereas Classification and STS show sharper transitions at lower shuffle ratios. At full shuffle ($p = 1.0$), all categories converge to negative correlations (averaging $\rho \approx -0.4$). This confirms that our metric relies heavily on the correct **alignment between source and target embeddings** rather than marginal statistics, and the varying sensitivity across tasks may reflect differences in the underlying embedding structure.

Conditional-only ablation. We investigate the contribution of the marginal term by comparing the full metric against a conditional-only variant, defined as $I_{\text{cond}}(u, v) = \log p_\phi(v \mid u)$. While this variant may be competitive when the target marginal distribution varies little, it lacks consistent reliability across diverse tasks. As quantified in Table 9 and visualized in Figure 5 (see Appendix D.2), removing the marginal term $\mathbb{E}[\log p_\phi(v)]$ reduces the average Spearman correlation substantially from 0.70 to 0.21. Both components are therefore necessary: the conditional term captures the cross-model mapping, the marginal term the intrinsic structure of the target space. We use the full score as default.

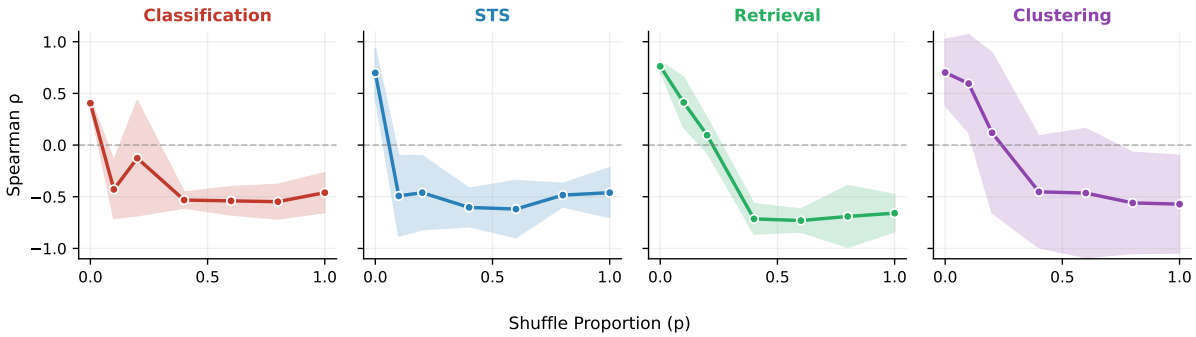

Figure 3: Partial shuffle ablation by task type. Increasing the shuffle proportion p causes correlation to degrade from positive ($p = 0$) to negative ($p = 1$), confirming that the metric relies on correct alignment.

Aggregation strategy. We aggregate the $N-1$ per-pair I_s scores of each source model into a single number using the median. The per-pair distribution is heavy-tailed: a few targets with very different geometry (e.g. tokenizer or pooling mismatches) inflate the mean and add noise unrelated to the source model's quality. Table 10 compares the mean, the median, and a 20% trimmed mean, and the median attains the highest or near-highest Spearman correlation with ground truth on every task family. The trimmed mean closes most of this gap but introduces an extra hyperparameter, so we keep the median as the default.

5.5 Additional Analysis

Assumption validation. Our theoretical guarantees in Theorem 1 are established under Assumption 2, which requires that the Jacobians of successive coupling layers act along approximately orthogonal directions. Under this condition, the overall Lipschitz constant grows approximately linearly with depth rather than exponentially (Lemma 2, Eq. 25). Although this assumption is idealized, it is supported by empirical evidence from three validation experiments; full results are reported in Appendix E.1. First, the singular vectors of inter-layer Jacobians exhibit a mean cosine similarity of 0.010, which is below the random baseline $1/\sqrt{d} = 0.016$, indicating near-orthogonality between layers. This behavior is consistent with the alternating binary masks and random permutations used in the NSF architecture. Second, the average per-layer spectral amplification is 1.049, showing that each layer contributes only a small and controlled perturbation, as encouraged by our zero-initialization strategy. Third, the cumulative perturbation amplification across all 18 atomic transforms in the composite flow (the 6 coupling layers, each followed by ActNorm and a random permutation) is only $2.38\times$, which is consistent with the approximately linear Lipschitz growth implied by Assumption 2. Overall, these results indicate that the structural regime required by Assumption 2 is well matched to the behavior of the trained models in practice.

Top-3 model identification. For practical model selection, the top of the ranking matters more than the tail: a practitioner mostly needs to know which two or three candidates to fine-tune or deploy, not the relative order of the weakest ones. We therefore compare the top-3 models ranked by I_s against the supervised top-3 on each of the 11 datasets. FLARE recovers the exact top-3 set on $4/11$ datasets and matches at least 2 out of 3 on $10/11$ datasets (the per-dataset breakdown is given in Appendix E.2). The single dataset on which only $1/3$ models agree is BhashaBench, where the top three supervised scores are all within one accuracy point and the ordering is essentially noise. In other words, whenever the supervised signal cleanly separates a leading group, FLARE identifies that group on its own, so the estimator is precise enough to drive model selection on an unlabeled corpus without any task-specific supervision.

Statistical robustness. To quantify ranking stability under a small pool ($N{=}8$), we run a leave-one-out bootstrap on the per-dataset 8×8 IS matrix. As reported in Table 3, $\rho_{\min}$ remains positive on every dataset, and the resampled Spearman ρ aggregates to $[0.583, 0.843]$, so the positive-correlation finding does not hinge on the inclusion of any particular model (full protocol in Appendix E.3). As a complementary test, converting rankings into $11 \times \binom{8}{2}{=}308$ pairwise preferences, FLARE matches the supervised ordering on 73.1% of pairs

(225/308; binomial p=1.37×10^{-16}, 95% CI lower bound 68.6%), well above the 50% chance baseline. Most residual disagreement comes from pairs with near-identical downstream performance, while the 90.9% Top-3 overlap confirms that FLARE reliably separates strong from weak models, the primary goal of model selection.

Dataset	Spearman ρ	Bootstrap range
FunPang	0.905	[0.857, 0.964]
Aug-STSB	0.833	[0.750, 0.964]
LivNLP-STS	0.833	[0.750, 0.964]
LIMIT	0.810	[0.714, 0.893]
BhashaBench	0.802	[0.739, 0.919]
arXiv '25	0.762	[0.643, 0.857]
Reasoning	0.762	[0.643, 0.857]
AIR-Bench	0.714	[0.607, 0.857]
apt-eval	0.429	[0.321, 0.750]
gtfintechlab	0.429	[0.143, 0.607]
Philo-STS	0.429	[0.250, 0.643]
Average	**0.701**	**[0.583, 0.843]**

Table 3: Leave-one-out bootstrap over the 8-model pool. "Spearman ρ" is the full 8-model estimate (matching Table 7); "Bootstrap range" is $[\rho_{\min}, \rho_{\max}]$ across the 8 resampled replicates obtained by recomputing ρ on each 7-model subset. The lower bound stays positive on every dataset.

Training stability. To test whether FLARE relies on a brittle optimum, we add Gaussian noise to each parameter tensor at σ=1−20% of its mean absolute value and re-evaluate the held-out NLL. Across the 616 conditional flows trained for the 11 datasets, the median relative NLL change is below 0.03% at $\sigma \leq 5\%$ and 1.22% at σ=20%; the median–mean gap indicates that the few sensitive pairs are outliers rather than the rule. The trained flows therefore lie in flat basins, and the resulting rankings are not driven by training randomness (Appendix E.4).

6 Conclusion

We presented FLARE, a label-free evaluator of text embedding models that estimates information sufficiency with normalizing flows. The framework replaces the kernel-density backbone of prior information-theoretic estimators with an exact log-likelihood from a normalizing flow, removing the dimensional sparsity that destabilises kernel methods on modern LLM-based embeddings. Finite-sample bounds derived via Rademacher complexity show that the estimation error scales with the intrinsic dimension of the data manifold rather than the raw embedding dimension, accounting for the estimator's behaviour in high-dimensional regimes. Across 11 datasets and eight embedders, FLARE correlates positively with supervised rankings on every task family (average Spearman ρ=0.70, up to 0.90), agrees with supervised pairwise preferences on 73.1% of pairs, and recovers $\geq 2/3$ of the supervised top-3 on 10/11 datasets. The induced ranking is also stable: the leave-one-out bootstrap keeps $\rho_{\min} > 0$ on every dataset, and 20–40% of the validation data is sufficient to reproduce the full-data ranking. Together, these properties make FLARE a practical alternative to annotated benchmarks for model selection on unlabeled corpora.

Practitioner's guide. The intended workflow is label-free model selection: a practitioner curates an unlabeled target corpus, picks a small pool of N candidate embedders, computes pairwise I_s on the corpus, and selects the top-scoring candidates. No downstream task or annotation is required. I_s is a relative quantity, comparable within a single dataset and candidate pool but not across datasets; the per-dimension normalization in Eq. 4 places encoders of differing widths on a common scale. The induced ranking is reliable when the per-dimension gap between adjacent candidates exceeds the leave-one-out bootstrap range (Appendix E.3); for tightly clustered candidates, the Top-3 overlap is a more robust summary than the full Spearman ρ. When a candidate exhibits an unexpectedly low score, the decomposition $I_s = H(V) - H(V \mid U)$ both localises the cause and indicates a remedy. A low marginal $H(V)$ reflects a poorly structured target density (e.g., collapsed or anisotropic), and motivates either substituting the candidate with a higher-capacity encoder or applying anisotropy-reducing post-processing such as whitening. A low conditional $H(V \mid U)$ accompanied by a healthy marginal indicates weak cross-model transferability, for which lightweight alignment such as a learned linear projection or distillation from a stronger reference is typically sufficient. Before attributing a depressed score to embedding quality, however, one should verify that the marginal flow's validation NLL has converged, as underfitting biases I_s downward for reasons unrelated to representation quality.

Limitations

FLARE has three main limitations. First, as a learning-based evaluation method, FLARE incurs non-trivial training cost: one conditional flow on Reasoning (d=3,584) takes $\sim$1.33 GPU hours on a V100, and the full N=8 pairwise evaluation costs $\sim$74.5 GPU hours. Caching the marginal flow makes adding a new model $O(N)$ ($\sim$10 GPU hours) rather than $O(N^2)$, and we further mitigate inference latency by using discrete normalizing flows instead of continuous-time approaches like flow matching; however, training a generative model remains a bottleneck for resource-constrained scenarios. Training-free baselines such as KDE-based EMIR cost only $\sim$6 GPU hours, though they degrade in high-dimensional spaces ($d\geq$3,584) where density estimates can collapse to a single kernel component.

Second, we observe a performance discrepancy across task types. While FLARE excels on geometry-centric tasks like STS and Retrieval, its correlation is lower for Classification ($\rho = 0.56$). This suggests that global information sufficiency captures the overall semantic manifold but may miss fine-grained, class-specific features required for linear separability.

Third, our theoretical analysis relies on simplifying assumptions. Specifically, regarding Assumption 2, parameter coupling during end-to-end backpropagation inevitably introduces inter-layer correlations, making the approximate independence assumption an idealization that is computationally intractable to verify in high dimensions.

References

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Peter L Bartlett and Shahar Mendelson. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482.

Jan Beirlant, Edward J Dudewicz, László Györfi, Edward C Van der Meulen, and 1 others. 1997. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jianlyu Chen, Nan Wang, Chaofan Li, Bo Wang, Shitao Xiao, Han Xiao, Hao Liao, Defu Lian, and Zheng Liu. 2024b. Air-bench: Automated heterogeneous information retrieval benchmark. *Preprint*, arXiv:2412.13102.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR.

Alain-Sam Cohen, Rama Cont, Alain Rossier, and Renyuan Xu. 2021. Scaling properties of deep residual networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2039–2048. PMLR.

Maxime Darrin, Philippe Formont, Ismail Ayed, Jackie CK Cheung, and Pablo Piantanida. 2024. When is an embedding model more promising than another? *Advances in Neural Information Processing Systems*, 37:68330–68379.

Vijay Devane, Mohd Nauman, Bhargav Patel, Aniket Mahendra Wakchoure, Yogeshkumar Sant, Shyam Pawar, Viraj Thakur, Ananya Godse, Sunil Patra, Neha Maurya, Suraj Racha, Nitish Kamal Singh, Ajay Nagpal, Piyush Sawarkar, Kundeshwar Vijayrao Pundalik, Rohit Saluja, and Ganesh Ramakrishnan. 2025. Bhashabench v1: A comprehensive benchmark for the quadrant of indic domains. *Preprint*, arXiv:2510.25409.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. 2019. Neural spline flows. *Advances in neural information processing systems*, 32.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Flat minima. *Neural Computation*, 9(1):1–42.

Wassily Hoeffding. 1994. *Probability Inequalities for sums of Bounded Random Variables*, pages 409–426. Springer New York, New York, NY.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. 2018. Mine: mutual information neural estimation. *arXiv e-prints*, pages arXiv–1801.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelinskiy, and Ping Tak Peter Tang. 2017. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*.

Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, Jy yong Sohn, and Chanyeol Choi. 2024. Linq-embed-mistral: Elevating text retrieval with improved gpt data through task-specific control and quality refinement. Linq AI Research Blog.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, and 1 others. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *Preprint*, arXiv:2402.09906.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, and 1 others. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 2021. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International conference on machine learning*, pages 5171–5180. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Danilo Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. Isoscore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339.

Shoumik Saha and Soheil Feizi. 2025. Almost AI, almost human: The challenge of detecting AI-polished writing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25414–25431, Vienna, Austria. Association for Computational Linguistics.

Agam Shah, Siddhant Sukhani, Huzaifa Pardawala, Saketh Budideti, Riya Bhadani, Rudra Gopal, Siddhartha Somani, Rutwik Routu, Michael Galarnyk, Soungmin Lee, Arnav Hiray, Akshar Ravichandran, Eric Kim, Pranav Aluru, Joshua Zhang, Sebastian Jaskowski, Veer Guda, Meghaj Tarte, Liqin Ye, and 8 others. 2025. Words that unite the world: A unified framework for deciphering central bank communications globally. *Preprint*, arXiv:2505.17048.

Bernard W Silverman. 2018. *Density estimation for statistics and data analysis*. Routledge.

Yixuan Tang and Yi Yang. 2025. FinMTEB: Finance massive text embedding benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3620–3638, Suzhou, China. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. 2025. On the theoretical limitations of embedding-based retrieval. *Preprint*, arXiv:2508.21038.

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

Gaifan Zhang, Yi Zhou, and Danushka Bollegala. 2025. Annotating training data for conditional semantic textual similarity measurement using large language models. *Preprint*, arXiv:2509.14399.

A Theoretical Assumptions and Proofs

A.1 Problem setup

Let $\mathcal{D}$ be an unknown distribution on $\mathbb{R}^d$. We observe an i.i.d. training sample

$$S_{\text{train}} = \{v_1^{\text{train}}, \ldots, v_m^{\text{train}}\} \sim \mathcal{D}^m. \tag{7}$$

A normalizing flow T_θ induces a density

$$p_\theta(v) = p_Z\big(T_\theta(v)\big)\,\big|\det J_{T_\theta}(v)\big|, \tag{8}$$

where p_Z is a fixed base density and $J_{T_\theta}(v)$ is the Jacobian of the flow transformation. We use the negative log-likelihood loss $\ell_\theta(v) = -\log p_\theta(v)$. The empirical training risk is

$$\hat{L}_{\text{train}}(\theta) = \frac{1}{m}\sum_{i=1}^{m} \ell_\theta\big(v_i^{\text{train}}\big). \tag{9}$$

The population risk is

$$L(\theta) = \mathbb{E}_{v\sim\mathcal{D}}\big[\ell_\theta(v)\big]. \tag{10}$$

Given a validation set

$$S_{\text{val}} = \{v_1^{\text{val}}, \ldots, v_{m_{\text{val}}}^{\text{val}}\} \sim \mathcal{D}^{m_{\text{val}}}, \tag{11}$$

the empirical validation risk is

$$\hat{L}_{\text{val}}(\theta) = \frac{1}{m_{\text{val}}}\sum_{j=1}^{m_{\text{val}}} \ell_\theta\big(v_j^{\text{val}}\big). \tag{12}$$

We aim to bound the generalization gap between the validation risk (observable proxy for performance) and the training risk (optimization objective):

$$\Delta(\theta) = \big|\hat{L}_{\text{val}}(\theta) - \hat{L}_{\text{train}}(\theta)\big|. \tag{13}$$

Note that bounding this gap involves controlling the deviation of both empirical risks from the population risk $L(\theta)$.

We first consider a marginal flow for a fixed embedder, then apply the same argument to a conditional flow. The Information Sufficiency (I_s) score is defined as

$$IS = L_{\text{marg}}(V) - L_{\text{cond}}(V \mid U), \tag{14}$$

and the empirical $\widehat{IS}$ score replaces the population risks with their validation counterparts. Establishing a finite-sample bound on the generalization gaps of the marginal and conditional flows directly yields a bound for the estimation error of the I_s score.

A.2 Core assumptions

Assumption 1 (Low intrinsic dimension)**.** *The embedding distribution is supported on a compact subset $\mathcal{M} \subset \mathbb{R}^d$ with intrinsic dimension $d_{\text{eff}} \ll d$ and bounded diameter. This type of low-dimensional structure for learned representations is consistent with empirical measurements of intrinsic dimensionality in deep features (Ansuini et al., 2019).*

Assumption 2 (Approximate layer independence)**.** *The flow T_θ is a composition of L invertible blocks with **perturbation rank** at most r. Assuming a regime of near-identity mappings (e.g., via residual connections or zero-initialization), the Jacobians of different blocks are approximately independent in their dominant singular directions. Consequently, the overall Lipschitz behaviour of the composition **grows linearly** in depth rather than exhibiting exponential growth. This assumption is consistent with analyses of signal propagation and dynamical isometry in deep architectures (Cohen et al., 2021).*

Remark: Validity via Zero-Initialization. While assuming approximate independence between layer Jacobians is non-trivial for generic deep networks, our Zero-Initialization strategy (Section 3.3) provides a rigorous structural justification. Specifically: (1) **Identity Start**: By initializing the conditioner's projection matrix B to zero, the conditional flow starts as an identity transformation relative to the backbone ($T_\theta = T_\phi$), ensuring that initial layer-wise Jacobians satisfy $J_l = I$. (2) **Linear Accumulation**: During progressive training, the deviation Δ_l from identity (where $J_l = I + \Delta_l$) is explicitly constrained via L_2 regularization. In this small-perturbation regime, the norm of the composed Jacobian follows a first-order approximation $\|\prod_{l=1}^{L}(I+\Delta_l)\| \approx \|I+\sum_{l=1}^{L}\Delta_l\|$, which implies that the Lipschitz constant grows **linearly** with depth L rather than exponentially. This design effectively aligns the flow's architectural behavior with the stability postulated in Assumption 2.

A.3 Rademacher complexity on a manifold

Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{M}$. Given a sample $S = \{v_1, \ldots, v_m\} \subset \mathcal{M}$, the em-

pirical Rademacher complexity of $\mathcal{F}$ is

$$\mathcal{R}_m(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(v_i)\right], \tag{15}$$

where σ_i are i.i.d. Rademacher variables.

Lemma 1 (Rademacher complexity on a manifold)**.** *Let $\mathcal{F}$ be the class of **linear functions** $f(v) = \langle w, v\rangle$ restricted to $\mathcal{M}$, where $\|w\| \le L_f$. Under Assumption 1, there exists a constant $C_{\mathrm{Rad}} > 0$ such that*

$$\mathcal{R}_m(\mathcal{F}) \;\le\; \frac{C_{\mathrm{Rad}}\, L_f\, D\, \sqrt{d_{\mathrm{eff}}}}{\sqrt{m}}. \tag{16}$$

Proof. The manifold $\mathcal{M}$ admits a covering number bound of the form

$$\mathcal{N}(\mathcal{M}, \varepsilon) \;\le\; C_0 \left(\frac{D}{\varepsilon}\right)^{d_{\mathrm{eff}}} \quad \text{for all } \varepsilon > 0. \tag{17}$$

Here, $\mathcal{N}(\mathcal{M}, \varepsilon)$ represents the covering number of the manifold, and C_0 is a geometry-dependent constant.

For the class of linear functions $\mathcal{F}$ with bounded norm $\|w\| \le L_f$, the covering number of the function space is controlled by the covering number of the domain $\mathcal{M}$ via a duality argument. Specifically, distinguishing two linear functions on $\mathcal{M}$ is equivalent to covering the domain at a finer scale ε/L_f. Consequently, we have:

$$\mathcal{N}(\mathcal{F}, \varepsilon) \le \mathcal{N}\Big(\mathcal{M}, \tfrac{\varepsilon}{L_f}\Big) \tag{18}$$

$$\le C_0 \left(\frac{DL_f}{\varepsilon}\right)^{d_{\mathrm{eff}}}. \tag{19}$$

Taking logarithms yields

$$\log \mathcal{N}(\mathcal{F}, \varepsilon) \;\le\; \log C_0 + d_{\mathrm{eff}}\, \log\Big(\frac{DL_f}{\varepsilon}\Big). \tag{20}$$

Let $F = L_f D$ be the uniform bound on $|f(v)|$. Dudley's entropy integral gives

$$\mathcal{R}_m(\mathcal{F}) \;\le\; \frac{12}{\sqrt{m}} \int_0^F \sqrt{\log \mathcal{N}(\mathcal{F}, \varepsilon)}\, d\varepsilon. \tag{21}$$

Substituting the covering number bound and simplifying (absorbing $\log C_0$ into the constant factor for asymptotic behavior), we obtain:

$$\mathcal{R}_m(\mathcal{F}) \le \frac{12 L_f}{\sqrt{m}} \int_0^D \sqrt{d_{\mathrm{eff}} \log\Big(\frac{D}{t}\Big)}\, dt, \tag{22}$$

where we utilized the substitution $t = \varepsilon/L_f$. Let $I = \int_0^D \sqrt{\log(D/t)}\, dt$. Using the change of variables $u = \log(D/t)$, we have $t = De^{-u}$, which yields $I = D\int_0^\infty u^{1/2} e^{-u} du = D \cdot \Gamma(3/2) = D\frac{\sqrt{\pi}}{2}$. Substituting this back yields:

$$\mathcal{R}_m(\mathcal{F}) \;\le\; \frac{12 L_f \sqrt{d_{\mathrm{eff}}}}{\sqrt{m}} \left(\frac{D\sqrt{\pi}}{2}\right) \;=\; \frac{6\sqrt{\pi}\, L_f\, D\, \sqrt{d_{\mathrm{eff}}}}{\sqrt{m}}. \tag{23}$$

By setting $C_{\mathrm{Rad}} = 6\sqrt{\pi}$, we recover the bound in (16). □

A.4 Architectural stability of the flow

We now turn to the flow T_θ. Let $J_\ell(v)$ be the Jacobian of the ℓ-th invertible block at input v, and let $J_{\mathrm{tot}}(v)$ denote the input–output Jacobian of T_θ. The Lipschitz constant is defined as

$$\mathrm{Lip}(T_\theta) = \sup_v \big\| J_{\mathrm{tot}}(v) \big\|_2. \tag{24}$$

Lemma 2 (Architectural stability via Zero-Initialization)**.** *Consider the flow T_θ composed of L blocks. Under the Zero-Initialization strategy, the Jacobian of the ℓ-th block takes the form $J_\ell = I + \Delta_\ell$. Let $\bar\sigma = \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}[\|\Delta_\ell\|_2]$ be the mean spectral norm of the perturbations. Consistent with the conservative estimation used in our implementation, the expected Lipschitz constant of the flow is bounded by:*

$$\mathbb{E}\big[\mathrm{Lip}(T_\theta)\big] \;\le\; 1 + L \cdot \bar\sigma. \tag{25}$$

Proof. The total Jacobian J_{tot} is the product of layer-wise Jacobians:

$$J_{\mathrm{tot}} = \prod_{\ell=1}^{L} J_\ell = \prod_{\ell=1}^{L} (I + \Delta_\ell). \tag{26}$$

We aim to bound the Lipschitz constant $\mathrm{Lip}(T_\theta) = \|J_{\mathrm{tot}}\|_2$. Recall that $J_{\mathrm{tot}} = \prod_{\ell=1}^{L}(I + \Delta_\ell)$. Performing a first-order expansion of this product yields:

$$J_{\mathrm{tot}} \approx I + \sum_{\ell=1}^{L} \Delta_\ell. \tag{27}$$

Applying the triangle inequality separates the identity transformation from the perturbations:

$$\|J_{\mathrm{tot}}\|_2 \approx \left\| I + \sum_{\ell=1}^{L} \Delta_\ell \right\|_2 \le \|I\|_2 + \sum_{\ell=1}^{L} \|\Delta_\ell\|_2. \tag{28}$$

Noting that $\|I\|_2 = 1$, we take the expectation:

$$\mathbb{E}[\|J_{\text{tot}}\|_2] \leq 1 + \sum_{\ell=1}^{L} \mathbb{E}[\|\Delta_\ell\|_2] \tag{29}$$

$$= 1 + L \cdot \left(\frac{1}{L} \sum_{\ell=1}^{L} \mathbb{E}[\|\Delta_\ell\|_2] \right). \tag{30}$$

Substituting the definition of the mean spectral norm $\bar{\sigma}$, we directly obtain:

$$\mathbb{E}[\text{Lip}(T_\theta)] \leq 1 + L \cdot \bar{\sigma}. \tag{31}$$

This confirms that under the small-perturbation regime enforced by zero-initialization, the Lipschitz constant grows linearly with depth L, aligning with the conservative bound used in our implementation. □

A.5 Finite-sample generalization bound

We now derive a finite-sample bound that matches the decomposition used in the main report. For clarity, we distinguish the training and validation samples

$$S_{\text{train}} = \{v_1^{\text{train}}, \ldots, v_m^{\text{train}}\}, \qquad S_{\text{val}} = \{v_1^{\text{val}}, \ldots, v_{m_{\text{val}}}^{\text{val}}\}, \tag{32}$$

and define the corresponding empirical risks

$$\hat{L}_{\text{train}}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \ell_\theta\big(v_i^{\text{train}}\big). \tag{33}$$

$$\hat{L}_{\text{val}}(\theta) = \frac{1}{m_{\text{val}}} \sum_{j=1}^{m_{\text{val}}} \ell_\theta\big(v_j^{\text{val}}\big). \tag{34}$$

The loss class is

$$\mathcal{L} = \big\{\ell_\theta(\cdot) = -\log p_\theta(\cdot) : \theta \in \Theta\big\}. \tag{35}$$

By Assumption 2 and Lemma 2, the loss functions are Lipschitz with respect to v. Using the architectural bound derived in Lemma 2 (which establishes linear growth with depth L), the Rademacher complexity of $\mathcal{L}$ admits the bound:

$$\mathcal{R}_m(\mathcal{L}) \;\leq\; \frac{\tilde{C}_{\text{Rad}}\, L\, \bar{\sigma}\, \sqrt{d_{\text{eff}}}}{\sqrt{m}}, \tag{36}$$

where $\tilde{C}_{\text{Rad}}$ collects the constants from Lemma 1 and the affine identity term. Note that this scales linearly with L, consistent with our implementation.

We now control the difference between training and validation risks. Introduce the population risk

$$L(\theta) = \mathbb{E}_{v \sim \mathcal{D}}\big[\ell_\theta(v)\big]. \tag{37}$$

By the triangle inequality,

$$\big|\hat{L}_{\text{val}}(\theta) - \hat{L}_{\text{train}}(\theta)\big| \leq \big|L(\theta) - \hat{L}_{\text{train}}(\theta)\big| + \big|\hat{L}_{\text{val}}(\theta) - L(\theta)\big|. \tag{38}$$

The two terms on the right-hand side are treated separately.

Training to population. A standard Rademacher complexity bound (see for example (Bartlett and Mendelson, 2002)) states that if $|\ell_\theta(v)| \leq M_{\text{train}}$ for all $\theta \in \Theta$ and all v in the support of the training distribution, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of S_{train} one has

$$\big|L(\theta) - \hat{L}_{\text{train}}(\theta)\big| \leq 2\mathcal{R}_m(\mathcal{L}) + 3M_{\text{train}} \sqrt{\frac{\log(2/\delta)}{2m}}. \tag{39}$$

Combining (39) with (36) gives

$$\big|L(\theta) - \hat{L}_{\text{train}}(\theta)\big| \leq \frac{2\tilde{C}_{\text{Rad}}\, L\, \bar{\sigma}\, \sqrt{d_{\text{eff}}}}{\sqrt{m}} + 3M_{\text{train}} \sqrt{\frac{\log(2/\delta)}{2m}}. \tag{40}$$

Validation to population. For the validation set, the model parameter θ is fixed, so we only need a concentration bound for bounded random variables. Let $X_j = \ell_\theta(v_j^{\text{val}})$ with $X_j \in [a, b]$ and define $M_{\text{val}} = b - a$. Hoeffding's inequality (Hoeffding, 1994) yields

$$\mathbb{P}\left(\left| \hat{L}_{\text{val}}(\theta) - L(\theta) \right| \geq t \right) \leq 2 \exp\left(-\frac{2m_{\text{val}} t^2}{M_{\text{val}}^2} \right). \tag{41}$$

Setting the right-hand side to δ and solving for t gives that, with probability at least $1 - \delta$,

$$\big|\hat{L}_{\text{val}}(\theta) - L(\theta)\big| \;\leq\; M_{\text{val}} \sqrt{\frac{\log(2/\delta)}{2m_{\text{val}}}}. \tag{42}$$

Final bound. Combining (38), (40) and (42), and applying a union bound, we obtain the following result.

Theorem 1 (Finite-sample generalization bound)**.** *Under Assumptions 1 and 2, for any fixed flow model p_θ and any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draws of the training set of size m and the validation set of size m_{val}, the train–validation gap satisfies*

$$\left|\hat{L}_{\text{val}}(\theta) - \hat{L}_{\text{train}}(\theta)\right| \le \frac{2\tilde{C}_{\text{Rad}}\, L\, \bar{\sigma}\, \sqrt{d_{\text{eff}}}}{\sqrt{m}} + M_{\text{val}}\sqrt{\frac{\log(2/\delta)}{2m_{\text{val}}}} + 3M_{\text{train}}\sqrt{\frac{\log(2/\delta)}{2m}}. \tag{43}$$

A.6 Extension to the IS score

Let L_{marg} and L_{cond} denote the population loss of the marginal and conditional flows, and let $\hat{L}_{\text{marg}}$ and $\hat{L}_{\text{cond}}$ be the corresponding validation losses. The population I_s score is

$$IS = L_{\text{marg}}(V) - L_{\text{cond}}(V|U), \tag{44}$$

and the empirical I_s score is

$$\widehat{IS} = \hat{L}_{\text{marg}}(V) - \hat{L}_{\text{cond}}(V|U). \tag{45}$$

The difference is

$$IS - \widehat{IS} = (L_{\text{marg}} - \hat{L}_{\text{marg}}) - (L_{\text{cond}} - \hat{L}_{\text{cond}}). \tag{46}$$

By the triangle inequality,

$$|IS - \widehat{IS}| \le |L_{\text{marg}} - \hat{L}_{\text{marg}}| + |L_{\text{cond}} - \hat{L}_{\text{cond}}|. \tag{47}$$

Applying Theorem 1 (Eq. 43) separately to the marginal and conditional flows yields a finite-sample upper bound for the generalization error of the I_s estimator as the sum of the two individual generalization gaps.

B Experiment Detail

In this section, we provide the experimental details necessary to reproduce these experiments.

B.1 Evaluated Model and Datasets Details

In Table 4, we provide the metadata of the evaluated models and their score on the 11 datasets. We provide the statistics of the datasets used to evaluate $\bar{I}_s$ in table 5.

Model	Dim.	$\bar{I}_S$
Zeta_Alpha_E5_Mistral	4096	**0.20**
Linq_Embed_Mistral	4096	**0.20**
SFR_Embedding_Mistral	4096	**0.19**
bge_multilingual_gemma2	3584	**0.19**
GritLM_7B	4096	**0.18**
gte_Qwen2_7B_instruct	3584	**0.17**
stella_base_en_v2	768	**0.13**
all_MiniLM_L6_v2	384	**0.13**

Table 4: Information sufficiency of the evaluated models by FLARE .

B.2 Downstream Task Evaluation

We deliberately select datasets that are either newly released or underexplored to minimize the risk of data leakage during embedding model pretraining. As no established performance benchmarks for embedding models exist on these datasets, we evaluate downstream task performance ourselves following the MTEB evaluation (Muennighoff et al., 2023) protocol: F1 Macro for classification, Spearman for STS, nDCG@10 for retrieval, and V-measure for clustering.

Table 6 reports the average downstream task performance of the eight evaluated embedding models, aggregated by task type. Overall, 7B-scale instruction-tuned models (GritLM, SFR, Linq, Zeta-Alpha) consistently outperform smaller models across all task categories. GritLM-7B achieves the best classification performance (0.61), while SFR-Embedding-Mistral leads on STS (0.67). The two smaller models, stella-base-en-v2 and all-MiniLM-L6-v2, show competitive performance on classification but lag significantly on retrieval and STS tasks.

Retrieval scores exhibit the largest variance, partly because LIMIT is designed to stress-test embedding-model capacity and yields low absolute scores for all candidates.

B.3 Training Configuration

The flow architecture itself is described in Section 3.3. Here we list the optimization hyperparameters, which are shared across all 11 datasets and all N=8 embedders without any per-dataset tuning, so that the same training recipe is used end-to-end. Both the marginal flow $p_\phi(v)$ and the conditional flow $p_\theta(v \mid u)$ are trained with AdamW under mixed precision (AMP), weight decay 1×10^{-3},

Dataset	Task	Train	Val	Total
apt-eval	Classification	13,185	1,465	14,650
gtfintechlab	Classification	12,250	2,625	14,875
BhashaBench-Finance	Classification	12,105	1,346	13,451
Aug-STSB	STS	33,635	3,738	37,373
LivNLP-STS	STS	12,758	1,418	14,176
Philosophical-STS	STS	58,560	6,507	65,067
AIR-Bench	Retrieval	23,639	2,627	26,266
LIMIT	Retrieval	45,000	5,000	50,000
arXiv '25	Retrieval	2,610	290	2,900
FunPang	Clustering	28,716	3,191	31,907
Reasoning	Clustering	50,772	5,642	56,414

Table 5: Statistics of the datasets used in our experiments.

Model	Classification	STS	Retrieval	Clustering
bge_multilingual_gemma2	0.58	0.63	0.46	0.30
Zeta_Alpha_E5_Mistral	0.57	0.65	0.53	0.31
GritLM_7B	0.61	0.64	0.50	0.32
SFR_Embedding_Mistral	0.57	0.67	0.53	0.26
Linq_Embed_Mistral	0.60	0.66	0.52	0.32
gte_Qwen2_7B_instruct	0.57	0.64	0.50	0.28
stella_base_en_v2	0.49	0.45	0.05	0.28
all_MiniLM_L6_v2	0.50	0.51	0.47	0.21

Table 6: Summary of the evaluated embedders and their performance on downstream datasets.

and EMA decay 0.999. The marginal flow uses initial learning rate 2×10^{-2}, batch size 256, gradient accumulation 2, and at most $1{,}000$ epochs; the conditional flow uses initial learning rate 1×10^{-1}, batch size 64, gradient accumulation 4, and at most 500 epochs.

B.4 Comprehensive Results

The primary goal of our method is to *rank* candidate embedding models so that practitioners can select the best one for a given downstream task without access to labeled data. The correlation coefficients reported in Table 7 validate that the rankings produced by FLARE align with ground-truth task performance.

We compare FLARE against four unsupervised baselines: Uniformity, IsoScore, Silhouette Score, and EMIR. Across 11 datasets spanning classification, STS, retrieval, and clustering, FLARE attains the highest average Spearman correlation ($\rho = 0.70$) and is the only method whose correlation is positive on every dataset.

EMIR achieves an average Spearman correlation of only -0.12, indicating that its rankings frequently contradict ground-truth performance. This is consistent with the well-known degradation of kernel-based density estimators in high-dimensional spaces. On individual datasets, EMIR shows high variance: while achieving moderate positive correlations on some tasks (e.g., $\rho = 0.33$ on LivNLP-STS), it produces strongly negative correlations on others (e.g., $\rho = -0.52$ on FunPang, $\rho = -0.43$ on arXiv '25). This instability limits its practical utility for model selection.

IsoScore exhibits consistently poor performance with an average of $\rho = -0.27$. These results confirm that FLARE provides the most reliable unsupervised signal for embedding model selection.

C Ranking Stability under Subsampling

We evaluate whether the proposed FLARE yields stable model rankings when the evaluation set is randomly subsampled. For each dataset, we subsample the evaluation set at ratios $\alpha \in$

		Spearman's ρ					Pearson's r				
Dataset	**Task**	**Uni.**	**Iso.**	**Sil.**	**EMIR**	**Ours**	**Uni.**	**Iso.**	**Sil.**	**EMIR**	**Ours**
apt-eval	Class.	0.36	-0.24	-0.21	0.07	**0.43**	0.50	-0.70	-0.42	-0.01	0.20
gtfintechlab	Class.	-0.12	-0.50	0.00	-0.38	0.43	0.11	-0.42	-0.18	-0.46	0.14
BhashaBench	Class.	0.29	-0.45	0.02	0.12	0.81	0.44	-0.62	0.35	-0.12	0.59
Aug-STSB	STS	0.14	0.05	0.24	-0.12	**0.83**	0.51	0.25	0.06	-0.11	**0.68**
LivNLP-STS	STS	-0.18	-0.27	0.81	0.33	**0.83**	0.77	-0.41	0.83	0.39	**0.94**
Philo-STS	STS	0.07	-0.76	**0.62**	-0.38	0.43	**0.82**	-0.26	0.70	-0.51	0.49
AIR-Bench	Retr.	-0.14	-0.79	0.05	-0.24	**0.71**	**0.70**	-0.27	0.45	0.07	0.69
arXiv '25	Retr.	0.55	0.14	-0.71	-0.43	**0.76**	**0.72**	-0.27	0.10	-0.33	0.62
LIMIT	Retr.	-0.17	-0.71	0.57	0.02	**0.81**	-0.14	-0.57	**0.62**	-0.28	**0.62**
FunPang	Clust.	-0.14	0.50	-0.48	-0.52	**0.90**	-0.52	0.82	-0.75	-0.57	**0.83**
Reasoning	Clust.	-0.14	0.07	0.00	0.21	**0.76**	-0.33	-0.73	0.26	0.37	**0.55**
Average		0.05	-0.27	0.08	-0.12	**0.70**	0.33	-0.29	0.18	-0.14	**0.58**

Table 7: Comparisons with unsupervised baselines. FLARE achieves the highest consistency and average correlation.

$\{5\%, 10\%, 20\%, 40\%, 60\%, 80\%, 100\%\}$ without replacement, and repeat this process 20 times. For each α, we recompute scores using the same pretrained marginal and conditional flows and obtain a ranking of embedding models. We then compute the Spearman rank correlation $\rho(\alpha)$ between the ranking induced by the subsampled evaluation set and the reference ranking computed on the full evaluation set ($\alpha = 1.0$), and report the deviation

$$\Delta_\rho(\alpha) \;=\; |\rho(\alpha) - \rho(1.0)|\,. \tag{48}$$

We focus on Spearman correlation because our goal is to assess the stability of *model ranking* for model selection, rather than the linear agreement of raw scores. Rank-based measures directly quantify whether the relative ordering of models is preserved under subsampling, which is the main quantity of interest in this analysis.

Figure 4 shows the ranking deviation $\Delta_\rho(\alpha) = |\rho_\alpha - \rho_{\text{full}}|$ as a function of the subsampling ratio α across all 11 datasets, grouped by task type. Overall, I_s rankings remain highly stable under subsampling: for 8 out of 11 datasets, $\Delta_\rho < 0.05$ even when using only 20% of the validation data.

Among task types, **clustering** exhibits the highest stability, with both datasets maintaining $\Delta_\rho < 0.025$ across all subsampling ratios. **Classification** and **retrieval** tasks also demonstrate strong robustness, with most datasets showing only minor deviations ($\Delta_\rho < 0.045$) even at very low α.

In contrast, **STS** datasets display larger variance at small sample sizes; *Aug-STSB* shows the highest deviation ($\Delta_\rho \approx 0.12$) at $\alpha = 0.05$, consistent with correlation-based evaluation being more sample-sensitive. However, these deviations diminish rapidly: once $\alpha \geq 0.2$, all STS datasets achieve $\Delta_\rho < 0.07$.

These results demonstrate that I_s selection is robust to evaluation subsampling, with 20–40% of validation data typically sufficient for reliable model ranking across diverse task types.

D Ablation Study

In this section, we conduct a set of ablation studies to better understand the factors that contribute to the effectiveness of our method. Unless otherwise specified, all experiments are conducted using the same evaluation protocol and datasets as in the main experiments.

Experimental setup. For all ablation experiments, we use the same pretrained models and evaluation datasets as in the main experiments. Unless otherwise stated, we recompute the evaluation scores under modified settings while keeping all other components unchanged. Performance is measured using Spearman correlation between the predicted ranking and the ground-truth ranking.

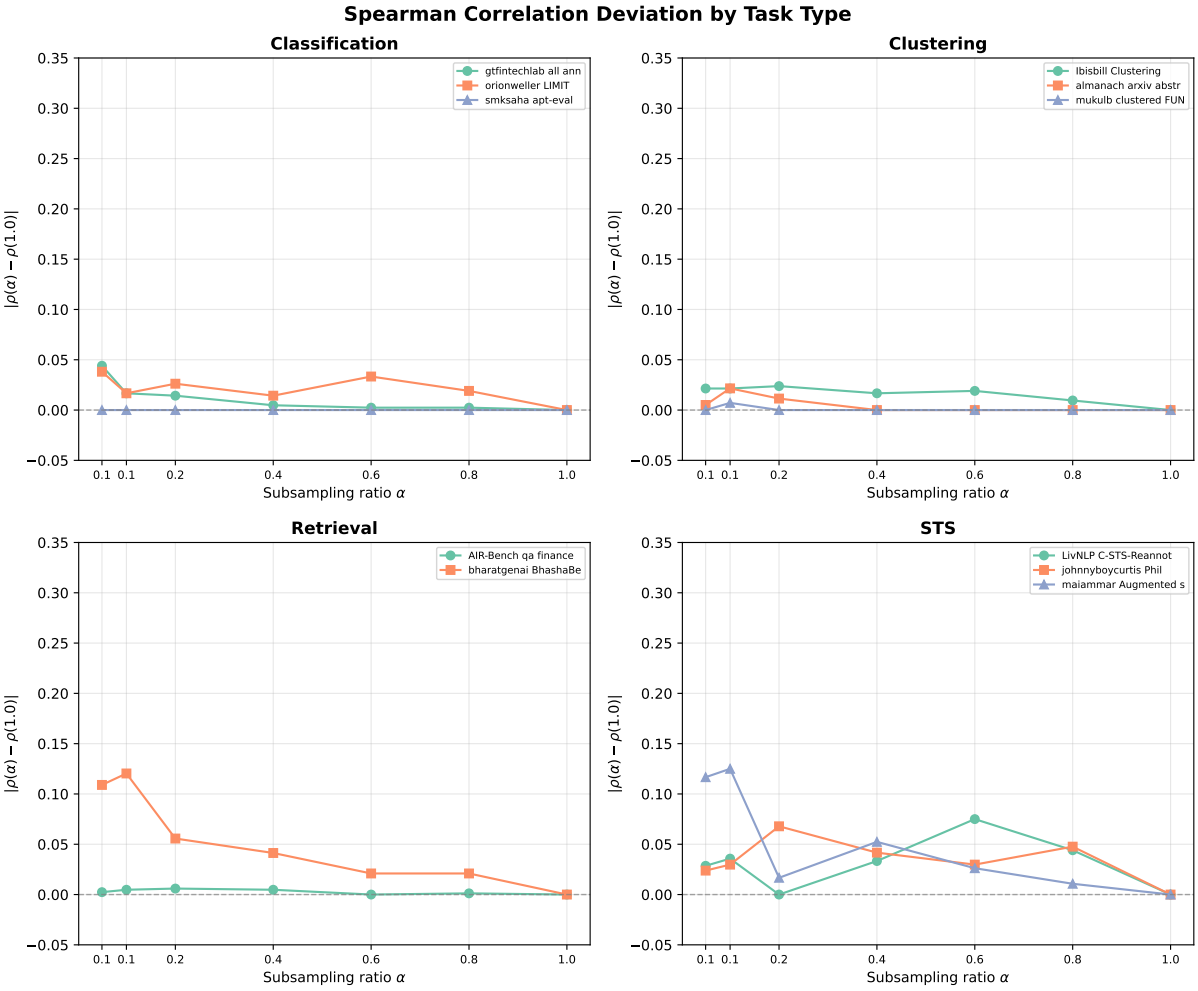

Figure 4: **Ranking stability under evaluation subsampling.** We report the deviation $\Delta_\rho(\alpha) = |\rho(\alpha) - \rho(1.0)|$, where $\rho(\alpha)$ is the Spearman rank correlation between the model ranking induced by IS scores computed on a subsampled evaluation set (ratio α) and the ranking computed on the full evaluation set ($\alpha = 1.0$). Smaller values indicate more stable rankings.

D.1 Shuffle Ablation

To examine whether the proposed metric truly relies on the correspondence between paired representations, we conduct a shuffle-based ablation with varying shuffle ratios. Specifically, for a given ratio $p \in \{0, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$, we randomly select a p fraction of the evaluation samples and permute the correspondence between U and V within this subset, while keeping the remaining $(1 - p)$ portion unchanged. This procedure preserves the marginal distributions of both U and V but progressively destroys their pairwise alignment as p increases.

For each shuffle ratio, we recompute the proposed score and evaluate the resulting ranking against the ground-truth downstream performance. As shown in Table 8, the Spearman correlation generally degrades as p increases across datasets. At $p = 0$, all datasets exhibit positive correlations (avg. $\rho = 0.70$); at $p = 1.0$, correlations become predominantly negative (avg. $\rho \approx -0.40$). The transition point varies by task type: Classification and STS datasets tend to flip sign at lower shuffle ratios ($p \approx 0.1$), while Retrieval and Clustering datasets maintain positive correlations longer (up to $p \approx 0.2$–0.4). This behavior confirms that the proposed metric critically depends on correct U-V alignment rather than marginal statistics alone.

D.2 Full I_s vs. Conditional-Only

To investigate the contribution of the marginal term in the I_s, we conduct an ablation study comparing the full I_s formulation against a conditional-only variant that omits the marginal likelihood component.

Results. Table 9 and Figure 5 present the comparison between Full I_s and Cond Only across all 11 datasets. The full I_s formulation achieves an average Spearman correlation of $\rho = 0.70$, substantially outperforming the Cond Only variant ($\rho = 0.21$). Notably, Full I_s outperforms Cond Only on 9 out of 11 datasets. The advantage is particularly pronounced on retrieval and clustering tasks: on LIMIT, Full I_s achieves $\rho = 0.81$ compared to -0.21 for Cond Only; on FunPang, the gap is 0.90 vs. -0.36. These results indicate that the conditional term alone captures only partial information about embedding quality. It measures how well the source embedding predicts the target, but fails to account for the intrinsic structure of the target embedding space.

Interestingly, on a few datasets (apt-eval, gtfintechlab), Cond Only slightly outperforms Full I_s. However, its performance is highly inconsistent, with five datasets showing negative correlations. Full I_s, in contrast, stays positive on all 11 datasets.

These findings confirm that both terms in the I_s formulation are necessary: the marginal term captures target-side embedding quality, while the conditional term measures cross-model information transfer. Their combination yields a more reliable and consistent signal for unsupervised model selection.

D.3 Aggregation Strategy

We investigate the impact of different aggregation strategies for combining IS scores into a single model score. We compare three methods: (1) arithmetic mean, (2) median, and (3) 10% trimmed mean (Trim10), which discards the top and bottom 10% of values before averaging.

As shown in Table 10, median aggregation consistently outperforms alternatives, achieving the highest Spearman correlation on 9 of 11 datasets (average $\rho = 0.70$ vs. 0.52 for mean, a relative improvement of 34.6%). For Pearson correlation, median also leads with an average of $r = 0.56$ compared to 0.54 for mean and 0.53 for trimmed mean.

The advantage of median aggregation is most pronounced on STS, where it reaches $\rho = 0.83$ on both Aug-STSB and LivNLP-STS, compared to $\rho \leq 0.45$ for the mean. The mean is pulled by a small number of pairs with anomalously high or low IS scores, while the median is unaffected.

Exceptions include Philo-STS and arXiv '25, where mean or trimmed mean aggregation outperforms the median. For Philo-STS, mean aggregation ($\rho = 0.62$) surpasses median ($\rho = 0.43$), likely due to the smaller dataset size reducing the prevalence of outlier scores. Similarly, on arXiv '25, trimmed mean achieves the highest correlation ($\rho = 0.84$), suggesting that while some outliers exist, the distribution tails might contain valuable signal for this specific retrieval task. Even with these exceptions, the median is the safest default across task families.

Based on these findings, we adopt median aggregation as the default strategy throughout our experiments.

Dataset	Task	$p = 0$	$p = 0.1$	$p = 0.2$	$p = 0.4$	$p = 0.6$	$p = 0.8$	$p = 1.0$
apt-eval	Cls	0.43	−0.36	0.48	−0.48	−0.43	−0.48	−0.48
gtfintechlab	Cls	0.43	−0.19	−0.24	−0.62	−0.69	−0.74	−0.64
BhashaBench	Cls	0.81	−0.74	−0.62	−0.50	−0.50	−0.43	−0.26
Aug-STSB	STS	0.83	−0.83	−0.81	−0.81	−0.93	−0.57	−0.71
LivNLP-STS	STS	0.83	−0.07	−0.10	−0.55	−0.52	−0.52	−0.43
Philo-STS	STS	0.43	−0.57	−0.48	−0.45	−0.40	−0.36	−0.24
AIR-Bench	Ret	0.71	0.76	0.24	0.69	−0.02	0.02	−0.55
arXiv '25	Ret	0.76	0.71	0.36	0.64	0.43	0.50	0.48
LIMIT	Ret	0.81	0.31	0.29	−0.81	−0.67	−0.36	−0.48
FunPang	Clust	0.90	0.93	0.67	−0.83	−0.90	−0.90	−0.90
Reasoning	Clust	0.76	0.26	−0.43	−0.07	−0.02	−0.21	−0.24

Table 8: Per-dataset Spearman correlation (ρ) under partial shuffle ablation. The column $p = 0$ corresponds to the full method without shuffling. As shuffle proportion p increases, correlation with downstream performance generally degrades.

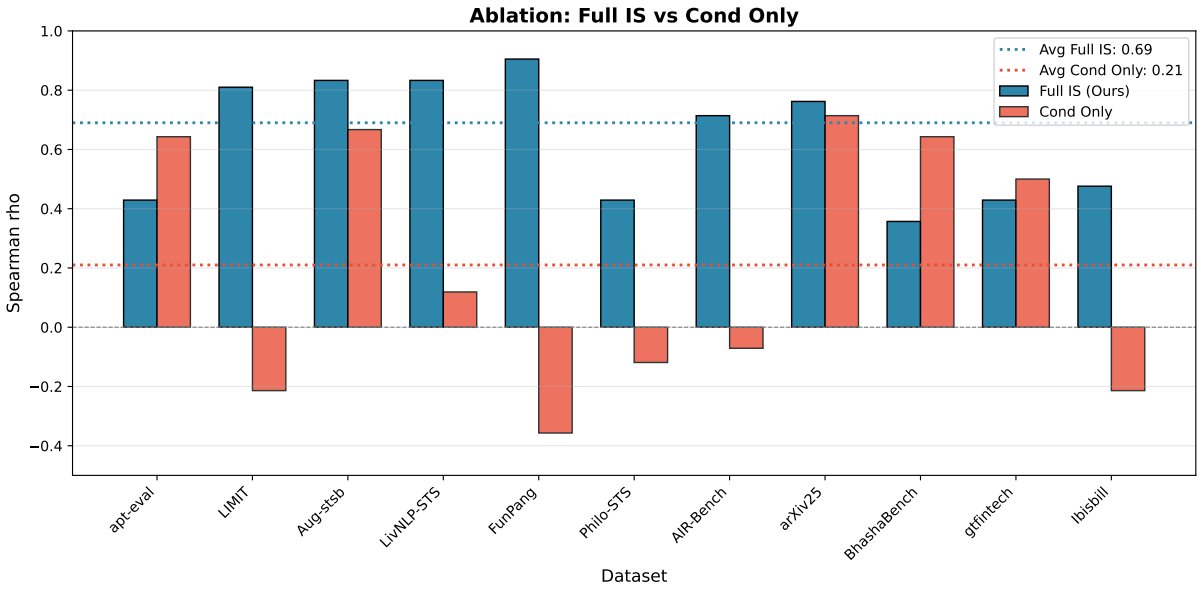

Figure 5: Ablation: Comparison of correlation of Full I_s vs. conditional-only component.

Dataset	Full IS	Cond Only
apt-eval	0.43	**0.64**
LIMIT	**0.81**	−0.21
Aug-STSB	**0.83**	0.67
LivNLP-STS	**0.83**	0.12
FunPang	**0.90**	−0.36
Philo-STS	**0.43**	−0.12
AIR-Bench	**0.71**	−0.07
arXiv '25	**0.76**	0.71
BhashaBench	**0.81**	0.64
gtfintechlab	0.43	**0.50**
Reasoning	**0.76**	−0.21
Average	**0.70**	0.21

Table 9: Ablation comparing Full I_s with Cond Only. Full I_s achieves substantially higher correlation with ground truth rankings (avg. $\rho = 0.70$ vs. 0.21), confirming the importance of the marginal term.

E Additional Analysis Details

This section provides full experimental results for the analyses summarized in Section 5.5.

E.1 Assumption Validation

This section presents the full experimental details for the three validation experiments summarized in Section 5.5.

Logical structure of the test. Assumption 2 states that the dominant Jacobian singular directions of different coupling layers are approximately independent. Together with a per-layer Lipschitz factor close to 1, this implies that the total Lipschitz constant of the L-layer flow grows approximately linearly in L rather than exponentially, which is the regime under which Theorem 1 provides a useful generalization bound. The three experiments are organized to verify this chain from complementary perspectives. Experiment 1 tests the core geometric assumption by checking that the dominant Jacobian directions of different layers are not aligned. Experiment 2 verifies the required local condition by showing that per-layer amplification remains close to 1 while layers still perform nontrivial transformations. Experiment 3 examines the global consequence by directly probing whether total perturbation amplification remains $O(1)$ in trained flows. The assumption itself cannot be exhaustively verified in high-dimensional parameter spaces; instead, the quantitative agreement among these three empirical observations is what makes Assumption 2 plausible in practice. In terms of scope, Experiment 1 is conducted on a small set of representative conditional flows from the FUNPANG dataset, while Experiments 2 and 3 aggregate statistics across conditional flows from all 11 datasets.

Experiment 1: Singular direction independence. We first estimate each coupling layer's Jacobian dominant singular vector via randomized finite-difference probing and compute pairwise $|\cos\theta|$ across all layer pairs (Table 11). Two independent random vectors in $\mathbb{R}^d$ have expected $|\cos\theta| \approx 1/\sqrt{d} = 0.016$, which provides the natural *independence baseline*: values at or below 0.016 indicate that dominant directions of different layers are no more aligned than random vectors.

The mean $|\cos\theta| = 0.010$ is consistently *below* this baseline (ratio $0.65\times$ on average across the three pairs), and the worst single layer pair only reaches 0.040. In other words, dominant singular directions of different coupling layers are no more aligned than randomly drawn unit vectors. We attribute this to the alternating binary masks and the random coordinate permutation applied after every coupling block in our NSF backbone (Section 3.3): each permutation rotates the "active" coordinate subspace, so consecutive layers structurally cannot inherit each other's dominant directions. The three model pairs span very different source/target architectures yet produce nearly identical numbers, supporting our claim in Section 5.5 that this is a property of the backbone rather than of any particular embedding.

To strengthen the test beyond a single direction, we also compute principal angles between the top-3 singular subspaces of adjacent layers (Table 12). Every reported angle lies in $[87.6°, 89.7°]$: the top-3 subspace of one layer projects onto that of the next layer with magnitude at most $\cos(87.6°) \approx 0.04$, so even when we widen the comparison from one direction to a three-dimensional subspace the layers remain essentially orthogonal.

Experiment 2: Per-layer behavior. We characterize each coupling layer by two complementary quantities: the relative point-wise displacement $\|f(y)-y\|/\|y\|$, which measures whether the layer performs a non-trivial transformation rather than collapsing to the identity, and the per-layer amplification factor (the empirical Lipschitz factor probed in random directions), which measures how much it can stretch perturbations. Both are computed for

		Spearman ρ			Pearson r		
Dataset	**Task**	Mean	Median	Trim10	Mean	Median	Trim10
apt-eval	Class.	0.29	0.43	0.29	0.26	0.20	0.16
gtfintechlab	Class.	0.38	0.43	0.17	0.36	0.14	-0.05
BhashaBench	Class.	0.53	**0.81**	0.36	0.69	0.59	0.45
Aug-STSB	STS	0.40	**0.83**	0.69	0.54	0.68	0.63
LivNLP-STS	STS	0.45	**0.83**	0.52	0.55	**0.94**	0.53
Philo-STS	STS	**0.62**	0.43	0.52	**0.66**	0.49	0.63
AIR-Bench	Retr.	0.43	**0.71**	0.48	0.52	0.70	**0.80**
arXiv '25	Retr.	0.70	0.76	**0.84**	0.25	0.46	**0.56**
LIMIT	Retr.	0.62	**0.81**	**0.81**	0.61	0.62	**0.65**
FunPang	Clust.	0.81	**0.90**	0.88	**0.95**	0.83	0.93
Reasoning	Clust.	0.52	**0.76**	0.60	**0.56**	0.55	**0.57**
Average		0.52	**0.70**	0.56	0.54	**0.56**	0.53

Table 10: Aggregation ablation comparing mean, median, and trimmed mean (10%). The median aggregation (Ours) achieves the highest consistency ($\rho = 0.70$).

Model Pair	**Mean** $\lvert\cos\theta\rvert$	**Max**	**Ratio**
bge $\rightarrow$ Zeta	0.012	0.040	0.75$\times$
GritLM $\rightarrow$ SFR	0.011	0.030	0.69$\times$
gte $\rightarrow$ Linq	0.009	0.024	0.56$\times$
Average	**0.010**	**0.040**	**0.65**$\times$

Table 11: Inter-layer cosine similarity of dominant singular vectors. The random baseline is $1/\sqrt{d} = 0.016$; all observed values are below it.

Layer Pair	**Principal Angles**
$L_0 \leftrightarrow L_1$	87.8°, 88.4°, 89.6°
$L_1 \leftrightarrow L_2$	87.6°, 88.6°, 89.7°
$L_2 \leftrightarrow L_3$	87.8°, 88.2°, 89.2°
$L_3 \leftrightarrow L_4$	88.3°, 88.8°, 89.7°
$L_4 \leftrightarrow L_5$	88.7°, 89.6°, 89.7°

Table 12: Principal angles between adjacent layers' top-3 singular subspaces.

every coupling layer of conditional flows trained on all 11 datasets (Table 13). The mean displacement of 0.390 shows that layers do meaningful work, while the geometric-mean amplification of 1.049 confirms that the per-layer Lipschitz factor is close to 1, satisfying the precondition required for Assumption 2 to deliver a useful linear-growth bound.

Metric	**Value**
Mean displacement per layer	0.390
Median	0.383
Max	0.586
Per-layer amplification (geo. mean)	1.049

Table 13: Per-layer relative displacement $\|f(y) - y\|/\|y\|$ across the conditional flows.

Experiment 3: Perturbation amplification. We test the consequence directly. For each conditional flow we add a random-direction perturbation of norm ε=0.01 to the input and measure the resulting relative output change $\|f(y + \varepsilon u) - f(y)\|/\varepsilon$ after the full sequence of 18 atomic transforms (Table 14). Random-direction probing does not certify a worst-case Lipschitz bound, but, combined with the near-orthogonality of dominant singular directions verified in Experiment 1, it provides a faithful empirical estimate of the typical amplification a perturbation experiences in the trained flow.

The mean probed amplification of $2.38\times$ closely matches the per-layer factor of 1.049 compounded over 18 layers ($1.049^{18} \approx 2.38$), and is many orders of magnitude smaller than what even mildly larger per-layer factors would produce (Table 15).

Metric	Value
Mean total amplification	$2.38\times$
Std	0.22
Min / Max	1.94 / 2.79
Per-layer geometric mean	1.049

Table 14: Perturbation amplification across the 18 atomic transforms. The observed $O(1)$ amplification is consistent with linear Lipschitz growth.

Per-layer σ	Total (σ^{18})
1.049 (ours)	$2.38\times$
1.1	$5.6\times$
1.5	$1{,}478\times$
2.0	$262{,}144\times$

Table 15: Comparison of observed amplification with hypothetical exponential growth at different per-layer rates.

The three experiments together pin down distinct links of the same chain: non-alignment between layer Jacobians (Experiment 1), near-unit per-layer amplification (Experiment 2), and bounded global amplification of the full flow (Experiment 3). Their numerical agreement places the trained flows in the regime where the linear Lipschitz bound underlying Theorem 1 is empirically realized.

E.2 Top-3 Model Identification

For each dataset we form two unordered sets of size three: GT-TOP3, the three highest-scoring models under the supervised metric, and FLARE-TOP3, the three highest-scoring models under FLARE's information-sufficiency score I_s. The reported overlap is the number of models shared by the two sets, ranging from 0 to 3, and is order-agnostic. Ties in either ranking are broken in a fixed lexicographic order over model names; with the gaps observed in our pool ($N{=}8$) this only affects fewer than three borderline assignments.

The Top-3 overlap targets the practitioner-relevant question of whether the ranker surfaces the genuinely best models for deployment. In practice only the top few candidates are actually deployed, so a low full-ranking Spearman ρ does not by itself indicate a poor selector: as long as the Top-3 is correct, mid- and bottom-rank shuffles are inconsequential. By contrast, Spearman ρ over the full $N{=}8$ list (used in the main results) over-penalises such harmless permutations, while Top-1 is too sensitive to a single GT measurement. Top-3 strikes a compromise: with $N{=}8$ the random-selection baseline yields an expected overlap of only $9/8{=}1.125$ and produces a perfect $3/3$ match less than 2% of the time, so even $2/3$ already strongly suggests a non-trivial signal.

Table 16 reports the per-dataset breakdown. Across datasets, the four with $3/3$ overlap (FunPang, LIMIT, Aug-STSB, LivNLP-STS) are also those with the largest GT gap between the top-3 cluster and the rest, which makes the decision boundary unambiguous. Conversely, datasets where the supervised top-3 scores are tightly clustered tend to yield smaller overlaps, suggesting that the difficulty of Top-3 identification is largely set by the GT separability of the candidate models.

E.3 Bootstrap Confidence Intervals

The most actionable risk for a small candidate pool ($N{=}8$) is that the ranking depends on which models happen to be included. To probe this, we run a leave-one-out bootstrap over the 8-model pool on the per-dataset 8×8 information-sufficiency matrix: for each of the 8 source models we drop one model at a time, recompute the per-source IS score and Spearman ρ against the supervised ranking on the remaining 7 models, and report the range $[\rho_{\min}, \rho_{\max}]$ across the 8 resampled replicates as a non-parametric confidence interval. The aggregate row in Table 3 averages the 11 per-dataset point estimates and the per-dataset $[\rho_{\min}, \rho_{\max}]$ bounds.

The width of the resampled range (Table 3) is itself informative. The narrowest ranges (FunPang $[0.857, 0.964]$, BhashaBench $[0.739, 0.919]$) appear on datasets where the IS scores are well separated, so removing any single model leaves the remaining ordering largely intact. Conversely, the widest range (gtfintechlab $[0.143, 0.607]$) is on a small classification benchmark with tightly clustered ground-truth scores, where one model's inclusion can flip several adjacent ranks. In every dataset $\rho_{\min}$ remains positive, and the aggregate lower bound (0.583) stays well above 0, so the positive-correlation finding does not hinge on the inclusion of any particular model.

Dataset	Task	GT Top-3	FLARE Top-3	Overlap
Aug-STSB	STS	Linq, SFR, GritLM	SFR, GritLM, Linq	3/3
LivNLP-STS	STS	Linq, SFR, BGE	SFR, Linq, BGE	3/3
LIMIT	Retrieval	Zeta, Linq, SFR	Linq, SFR, Zeta	3/3
FunPang	Clustering	Linq, Zeta, SFR	Zeta, Linq, SFR	3/3
AIR-Bench	Retrieval	Zeta, SFR, Qwen2	Zeta, GritLM, SFR	2/3
arXiv '25	Retrieval	GritLM, Linq, Zeta	Linq, SFR, Zeta	2/3
Reasoning	Clustering	SFR, GritLM, Zeta	Zeta, SFR, Linq	2/3
gtfintechlab	Class.	Zeta, BGE, Linq	Zeta, Linq, SFR	2/3
apt-eval	Class.	GritLM, Linq, BGE	Linq, BGE, Qwen2	2/3
Philo-STS	STS	Linq, Qwen2, SFR	SFR, Linq, Zeta	2/3
BhashaBench	Class.	GritLM, Qwen2, BGE	Zeta, GritLM, SFR	1/3

Table 16: Top-3 model identification across all 11 datasets. FLARE achieves exact $3/3$ overlap on $4/11$ (36.4%) datasets and at least $2/3$ overlap on $10/11$ (90.9%) datasets.

E.4 Weight Perturbation Robustness

For each trained conditional flow we form a perturbed copy by adding i.i.d. Gaussian noise to every parameter tensor, scaled to the tensor's own magnitude: for tensor W with mean absolute magnitude $\overline{|W|}$, $\Delta W \sim \mathcal{N}(0, (\sigma\,\overline{|W|})^2 I)$. Per-tensor scaling avoids a single global noise level overwhelming small-magnitude tensors (e.g. ActNorm scales) while leaving large-magnitude ones untouched, and gives the standard flat-minimum probe interpretation (Hochreiter and Schmidhuber, 1997; Keskar et al., 2017). We sweep $\sigma \in \{1\%, 2\%, 5\%, 10\%, 20\%\}$ with 3 noise draws per configuration, evaluate the conditional NLL of the clean and perturbed flows on the held-out validation split, and report the relative change $\Delta\mathrm{NLL}/\mathrm{NLL}$ in Table 17.

The median relative NLL change stays below 0.03% for $\sigma \leq 5\%$ and reaches only 1.22% at σ=20%, indicating that trained flows sit in flat basins and are essentially insensitive to small parameter perturbations. The widening gap between median and mean shows that sensitivity is concentrated in a small number of pairs rather than being a generic property of the ensemble.

σ	Median	Mean	Std	Max
1%	+0.00%	+0.09%	0.54%	5.11%
2%	+0.01%	+0.29%	1.13%	9.58%
5%	+0.02%	+1.21%	3.52%	38.3%
10%	+0.13%	+3.89%	8.97%	81.3%
20%	+1.22%	+16.4%	36.6%	389%

Table 17: Relative NLL change (%) under weight perturbation across 616 model pairs from 11 datasets. Median values are reported alongside mean, std, and max. At $\sigma \leq 5\%$, the median NLL change remains below 0.03%, indicating that the vast majority of trained flows reside in flat, stable minima.