H-MSAE: HIERARCHICAL SPARSE CONCEPT COVERAGE FOR EFFICIENT AND DIVERSE INSTRUCTION DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large instruction-tuning corpora contain substantial semantic redundancy, motivating the selection of compact subsets that preserve useful supervision while reducing training cost. Existing diversity-aware selectors often rely on generated semantic tags and explicit graph structures, introducing considerable annotation and preprocessing overhead. We introduce H-MSAE, a hierarchical sparse instruction selection framework that organizes multi-resolution concepts learned by a Matryoshka sparse autoencoder into a coarse-to-fine hierarchy. Quality-weighted, diminishing-return coverage with ancestor propagation promotes high-quality and non-redundant semantic coverage, while pool-relative length constraints regulate subset composition. We motivate this design through a local quality-diversity analysis of subset SFT utility and examine it using example-quality and representation-space diversity diagnostics. On Tulu 3 data pool, H-MSAE achieves the highest Final Avg among the evaluated 50K selectors across scales and initialization. The selected subset also exhibits the highest mean quality and effective rank among the evaluated methods. By avoiding full-pool generative tagging and supporting efficient sparse selection, H-MSAE reduces measured end-to-end runtime by approximately $5\times$ relative to MIG.

1 INTRODUCTION

Large instruction-tuning corpora contain substantial semantic redundancy, making it desirable to select compact subsets that preserve useful supervision while reducing training cost. Effective selection must account for both example quality and semantic diversity: high-quality examples can provide reliable supervision, while non-redundant examples can expose the model to a broader range of training signals. Existing approaches address these properties through learned quality scores, generated semantic tags, taxonomies, or graph-based information gain (Chen et al., 2024; Liu et al., 2024a;b; Lu et al., 2023; Chen et al., 2025; Niu et al., 2026). However, diversity-aware selectors often rely on expensive full-pool annotation and explicit semantic structure construction. This motivates learning the semantic structure needed for selection directly from frozen conversation representations.

We introduce H-MSAE, a hierarchical sparse instruction selector built on multi-resolution concepts learned by a Matryoshka sparse autoencoder (MSAE). Cross-resolution co-activations establish soft hierarchical relations between sparse concepts, allowing an example to contribute to both fine-grained features and their broader ancestors. We then maximize a concave coverage objective that rewards underrepresented concepts while introducing diminishing returns for redundant examples. Unlike tag-based approaches, H-MSAE does not require generated semantic labels or an explicit graph over training examples.

Semantic novelty alone, however, does not guarantee useful supervision: rare or noisy examples can receive high coverage gains. We therefore incorporate example quality directly into the coverage objective, allowing quality to influence both immediate marginal gains and subsequent concept saturation. Since the adopted quality scorer exhibits a preference toward shorter examples, we additionally impose pool-relative length constraints to regulate subset composition. These design choices are motivated by our local SFT analysis, which identifies complementary contributions from

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.

supervision quality and directional diversity to expected subset utility. We examine this interpretation empirically through example-quality and representation-space diversity diagnostics.

Experiments on Tulu 3 show that H-MSAE achieves the highest Final Avg among the evaluated 50K selectors across three Qwen3 model scales, with particularly pronounced aggregate gains in subjective evaluation. Its selected subset also exhibits the highest mean quality and effective rank among the evaluated methods, while component ablations and data-scaling experiments support the combined design. Beyond selection effectiveness, H-MSAE achieves approximately $5\times$ lower end-to-end runtime than MIG by avoiding full-pool generative tagging. With prerequisite representations cached, lightweight hierarchy construction and sparse stochastic-greedy selection further reduce structure-construction and core sampling time by $4.8\times$ and $44.9\times$, respectively, consistent with the theoretical complexity analysis.

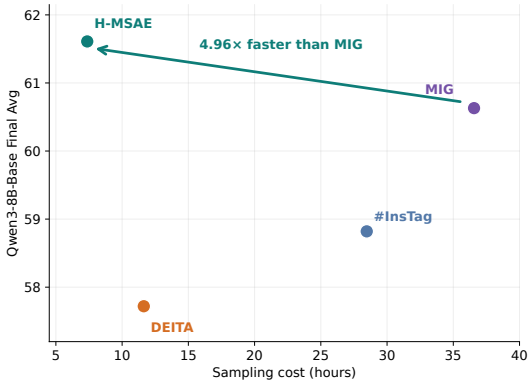


Figure 1: Performance–cost frontier on Qwen3-8B-Base. H-MSAE achieves a 0.98-point higher Final Avg than MIG while using approximately one-fifth of its end-to-end wall-clock time.

Our contributions are:

- **Hierarchical sparse instruction selection with quality and composition control.** We introduce H-MSAE, which builds on multi-resolution sparse concepts obtained from a Matryoshka sparse autoencoder and organizes them into a coarse-to-fine hierarchy through cross-resolution relations. We perform diminishing-return coverage over this hierarchy with ancestor propagation, and further integrate quality-weighted coverage and pool-relative length constraints into a unified stochastic-greedy framework that jointly controls semantic coverage, example quality, and subset composition.
- **A quality–diversity view of instruction subset utility.** We interpret instruction subset utility through the joint role of example quality and semantic diversity: useful subsets should retain diverse semantic structure while avoiding low-quality supervision. This view is motivated by a local utility analysis and examined empirically through example quality and effective rank.
- **Strong downstream performance with efficient reusable selection.** H-MSAE achieves the highest overall average among the evaluated 50K selectors across model scales and initialization settings. The learned sparse representation enables lightweight post-representation selection and can be reused across selection budgets and downstream models. Unlike tag-based selectors, H-MSAE avoids full-pool generative tagging and explicit tag-level graph construction, substantially reducing the measured end-to-end selection cost.

2 Related Work

Instruction data selection. Instruction selectors use external judges, uncertainty, learned quality/complexity, open-set intents, gradients, or diversity objectives (Chen et al., 2024; Liu et al., 2024a;b; Lu et al., 2023; Xia et al., 2024). DEITA combines quality and complexity scoring, while #InsTag uses open-set semantic tags. MIG combines tags and quality in a graph-based information-gain objective, and TAGS constructs a fine-grained taxonomy (Chen et al., 2025; Niu et al., 2026). At larger scale, sophisticated selectors do not always dominate random selection, making strong random controls important (Ivison et al., 2025). Our focus is a fixed-pool, downstream-label-free selector that replaces tag-based semantic structure with reusable sparse concepts.

Sparse autoencoders and representations. Sparse autoencoders (SAEs) learn sparse latent features from unlabeled neural representations, traditionally through sparsity regularization (Ng et al., 2011) and more recently through architectures such as k-sparse/Top-k autoencoders (Makhzani & Frey,

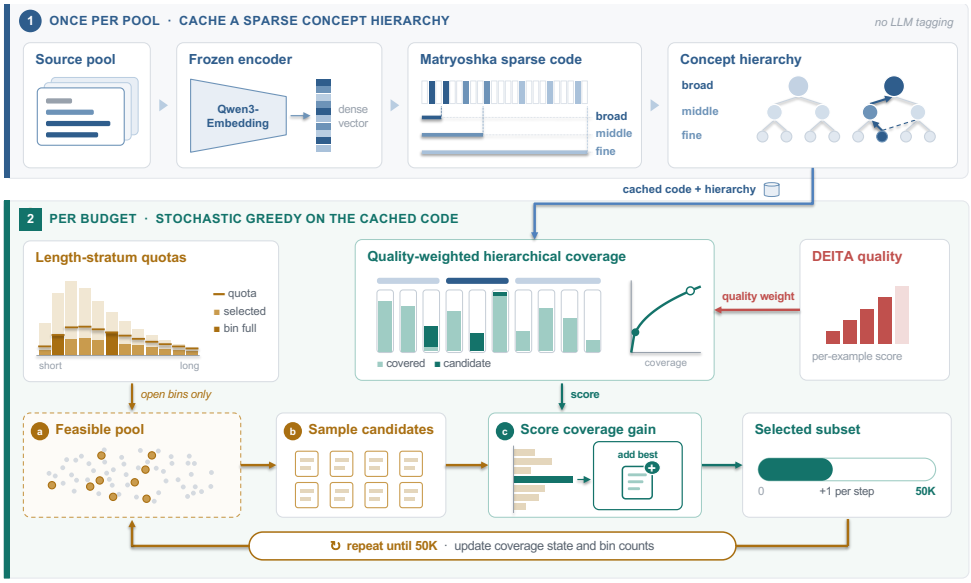


Figure 2: Overview of H-MSAE. Frozen conversation embeddings are converted into multi-resolution sparse concepts and organized through cross-resolution relations. Quality-weighted hierarchical coverage determines semantic utility, while pool-derived length quotas constrain subset composition; stochastic greedy selects the final subset.

2013; Gao et al., 2025). Recent work has further studied the semantic quality, robustness, and theoretical limitations of SAE features (Minegishi et al., 2025; Cui et al., 2026). SAE representations have also been explored for diversity-aware tuning and data engineering (Yang et al., 2025; Jing et al., 2026). Matryoshka Sparse Autoencoders (MSAEs) provide nested sparse representations at multiple capacities (Bussmann et al., 2025), which we adopt as the representation backbone of H-MSAE. On top of MSAE, we organize the resulting multi-resolution concepts through cross-resolution relations, optimize quality-weighted hierarchical concave coverage, and incorporate length-aware composition constraints for instruction selection.

Length and compute in instruction tuning. Response length is a nontrivial selection variable: selecting long responses can itself be a strong baseline (Zhao et al., 2024), and packing makes realized optimization steps depend on the selected length distribution. Compute-aware data selection can also be posed explicitly (Wan et al., 2025). Our length mechanism is narrower: it constrains subset composition using only pool statistics rather than optimizing against a target token total.

3 H-MSAE: Hierarchical Sparse Instruction Selection

3.1 Problem Setup and Overview

Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ be an eligible instruction-tuning pool and let $B \ll N$ denote the selection budget. We seek an index set $\mathcal{I} \subseteq [N]$ with $|\mathcal{I}| = B$ that preserves broad semantic coverage while avoiding redundant or low-quality supervision, without access to downstream evaluation labels.

H-MSAE consists of three components. First, we obtain multi-resolution sparse concepts using MSAE and organize them into a cross-resolution hierarchy. Second, we define a quality-weighted diminishing-return coverage objective over these concepts. Third, we maximize this utility using stochastic greedy under pool-derived length constraints.

3.2 Hierarchical Sparse Concept Representation

Multi-resolution sparse concepts. Each complete conversation is serialized and encoded using a frozen representation model as $e_i = E(\text{serialize}(x_i)) \in \mathbb{R}^d$. We adopt the Matryoshka Sparse

Autoencoder (MSAE) (Bussmann et al., 2025), which learns nested sparse representations at multiple widths within a shared model. It exposes coarse-to-fine sparse concepts from the same representation space, providing a natural basis for constructing cross-resolution semantic relations. A shared encoder produces the maximum-width sparse activation $z_i = \mathrm{BatchTopK}(\sigma(W_{\mathrm{enc}} e_i + b_{\mathrm{enc}}))$ with nested widths $D_1 < \cdots < D_L$. The Matryoshka sparse autoencoder is trained by

$$\mathcal{L}_{\mathrm{MSAE}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \alpha_\ell \left\| e_i - W_{\mathrm{dec}}^{(\ell)} z_i^{(1:D_\ell)} \right\|_2^2. \tag{1}$$

Only newly introduced latent units are treated as concepts at each resolution: $\mathcal{F}_1 = \{1, \ldots, D_1\}$ and $\mathcal{F}_\ell = \{D_{\ell-1} + 1, \ldots, D_\ell\}$ for $\ell > 1$. Thus, each latent concept belongs to exactly one resolution.

Cross-resolution hierarchy. For concept f, let $\mathcal{S}_f = \{i : z_{if} > 0\}$ denote its activation support. For a child concept $f \in \mathcal{F}_\ell$ and candidate parent $p \in \mathcal{F}_{\ell-1}$, we define asymmetric containment and the corresponding soft parent weight by

$$\mathrm{cont}(f, p) = \frac{|\mathcal{S}_f \cap \mathcal{S}_p|}{|\mathcal{S}_f|}, \qquad a_{fp} = \frac{\exp(\mathrm{cont}(f, p)/T_p)}{\sum_{p' \in \mathcal{P}_{K_p}(f)} \exp(\mathrm{cont}(f, p')/T_p)}, \tag{2}$$

where $\mathcal{P}_{K_p}(f)$ contains the top-K_p candidate parents in the immediately coarser resolution and T_p is the parent-assignment temperature.

We propagate fine-grained activations toward their ancestors so that one example can contribute simultaneously to specific and broader concepts:

$$\widetilde{z}_{iu} = z_{iu} + \sum_{f \in \mathrm{Desc}(u)} \pi_{f \to u} z_{if}, \tag{3}$$

where $\mathrm{Desc}(u)$ denotes all direct and indirect descendants of concept u, and $\pi_{f \to u}$ is the total propagation weight along all paths from f to u, determined by the soft parent weights and optional depth decay. The propagated activation $\widetilde{z}_i$ is used directly by the selection objective below.

3.3 Quality-Weighted Hierarchical Coverage

Let q_i denote the quality score of example i and define its normalized quality weight as $\omega_i = (q_i/\bar{q})^\beta$, where $\bar{q} = N^{-1} \sum_{i=1}^{N} q_i$ and $\beta \geq 0$ controls the strength of quality weighting. We use the released DEITA quality scorer in our implementation (Liu et al., 2024b).

The subset utility is defined by quality-weighted concave coverage across all hierarchy levels:

$$F(\mathcal{I}) = \sum_{\ell=1}^{L} \lambda_\ell \sum_{f \in \mathcal{F}_\ell} \phi\left(\sum_{i \in \mathcal{I}} \omega_i \widetilde{z}_{if} \right), \qquad \phi(t) = (t + \epsilon)^\gamma - \epsilon^\gamma, \quad 0 < \gamma < 1. \tag{4}$$

Here $\widetilde{z}_{if}$ determines which concepts an example covers across resolutions, while ω_i controls the strength of its contribution. The concavity of ϕ introduces diminishing returns, favoring examples that cover insufficiently represented concepts over repeated coverage of already saturated regions. We denote the marginal gain of candidate i by

$$\Delta_F(i \mid \mathcal{I}) = F(\mathcal{I} \cup \{i\}) - F(\mathcal{I}). \tag{5}$$

3.4 Length-Constrained Stochastic-Greedy Selection

Because quality-aware selection can alter the response-length distribution, particularly when the DEITA quality scorer exhibits a preference toward shorter examples, we use length to constrain subset composition rather than incorporating it into the semantic utility. We partition the eligible pool into m quantile bins $\{\mathcal{B}_b\}_{b=1}^{m}$ according to serialized conversation length and define the feasible family

$$\mathcal{M} = \{\mathcal{I} \subseteq [N] : |\mathcal{I} \cap \mathcal{B}_b| \leq \kappa_b, \ \forall b\}, \qquad \kappa_b \propto \bar{L}_b^\tau, \qquad \sum_{b=1}^{m} \kappa_b = B, \tag{6}$$

where $\bar{L}_b$ is the mean serialized length in bin b and $\tau \geq 0$ controls the strength of length reweighting.

Starting from $\mathcal{I}_0 = \varnothing$, at iteration t we sample a candidate set R_t from the currently feasible examples $\{i \notin \mathcal{I}_t : \mathcal{I}_t \cup \{i\} \in \mathcal{M}\}$ and select

$$i_t^\star = \arg\max_{i \in R_t} \Delta_F(i \mid \mathcal{I}_t), \qquad \mathcal{I}_{t+1} = \mathcal{I}_t \cup \{i_t^\star\}. \tag{7}$$

Following Mirzasoleiman et al. (2015), we use at most $s = \lceil (N/B) \log(1/\varepsilon_{\text{sg}}) \rceil$ sampled candidates per iteration. The standard approximation guarantee associated with this rule applies to the cardinality-constrained setting; under our length-aware constraints, we use the same sampling rule with explicit feasibility checks. After B iterations, the algorithm returns the final selected index subset $\mathcal{I}_B$.

4 Analysis of Subset Utility and Efficiency

4.1 A Quality–Diversity Analysis of Subset SFT

We analyze why H-MSAE can yield effective instruction subsets from a quality–diversity perspective. Its objective combines quality weighting with diminishing-return coverage over hierarchical sparse concepts, encouraging both reliable supervision and non-redundant semantic content. We show that, under a local SFT model, these two properties contribute complementary positive terms to expected target-risk reduction.

Let θ_0 denote the model before SFT and

$$R_{\mathcal{T}}(\theta) = \mathbb{E}_{z \sim \mathcal{T}}[\ell(z; \theta)] \tag{8}$$

the population risk on a target instruction distribution $\mathcal{T}$. We denote its local gradient and Hessian by $g_{\mathcal{T}} = \nabla_\theta R_{\mathcal{T}}(\theta_0)$, and $H_{\mathcal{T}} = \nabla_\theta^2 R_{\mathcal{T}}(\theta_0) \succeq 0$. For a fixed-size subset S, $|S| = B$, let g_i denote the local per-example training gradient, modeled stochastically below. We approximate the update induced by subset SFT as $\Delta_S = -\frac{\eta}{B} \sum_{i \in S} g_i$, where $\eta > 0$ is an effective local step size, and define the utility of S by its reduction in target risk,

$$U_{\mathcal{T}}(S) = R_{\mathcal{T}}(\theta_0) - R_{\mathcal{T}}(\theta_0 + \Delta_S). \tag{9}$$

Local gradient model. The population gradient $g_{\mathcal{T}}$ specifies the local direction of the target objective, whereas an individual training example contributes a stochastic gradient g_i that need not coincide with $g_{\mathcal{T}}$. We write $\bar{g}_i := \mathbb{E}[g_i]$ for its expected training direction and associate each example with a latent quality $q_i \in [0, 1]$, where higher quality corresponds to lower stochastic uncertainty around $\bar{g}_i$. The precise covariance model are given in Appendix A.3.

Subset quality. For a subset S of size B, we define its average quality as

$$Q(S) = \frac{1}{B} \sum_{i \in S} q_i. \tag{10}$$

$Q(S)$ measures the reliability of the supervision contained in the subset: a larger value means that, on average, the selected examples induce training gradients with lower stochastic uncertainty.

Directional diversity. To characterize redundancy between examples, we compare their expected training directions in the local geometry induced by the target-risk Hessian $H_{\mathcal{T}}$. For each example, define the curvature-normalized direction as $v_i = H_{\mathcal{T}}^{1/2} \bar{g}_i / \|H_{\mathcal{T}}^{1/2} \bar{g}_i\|_2$. We define the directional diversity of a subset S as

$$D(S) = \operatorname{tr}\left(\widehat{\operatorname{Cov}}_{i \in S}[v_i] \right), \tag{11}$$

where $\operatorname{tr}(\cdot)$ denotes the trace and $\widehat{\operatorname{Cov}}_{i \in S}[v_i]$ denotes the sample covariance of the normalized directions $\{v_i\}_{i \in S}$. Intuitively, $D(S)$ measures their total directional dispersion: it is small when the selected examples induce highly redundant training directions and increases as these directions spread across the local target-risk geometry.

Throughout the analysis, $\mathbb{E}_g$ denotes expectation over the stochastic training gradients $\{g_i\}_{i \in S}$ for a fixed subset S. Thus, $Q(S)$ and $D(S)$ are deterministic properties of the selected examples, while their contributions to expected utility can vary with the choice of S.

Theorem 4.1 (Quality–diversity decomposition of subset SFT utility)**.** *Under the local gradient modeling, the expected target-risk reduction of a given subset S of size B admits the decomposition*

$$\mathbb{E}_g[U_{\mathcal{T}}(S)] = C_B + \lambda_Q Q(S) + \lambda_D D(S) + o(\eta^2) \tag{12}$$

where C_B, λ_Q, λ_D are specified in the proof.

Theorem 4.1 shows that expected subset utility increases with both average gradient reliability and directional dispersion. Here, $\mathbb{E}_g$ denotes expectation over the stochastic training gradients $\{g_i\}_{i\in S}$ for a fixed subset S. Thus, $Q(S)$ and $D(S)$ are deterministic properties of the selected examples and can vary with the subset composition. The proof is provided in Appendix A.3.

This decomposition provides an interpretation of the two signals combined by H-MSAE: quality-weighted coverage favors examples with higher estimated supervision quality, while diminishing-return hierarchical coverage promotes non-redundant coverage across latent semantic regions. Since the implementation uses DEITA quality scores, we additionally impose length constraints to mitigate their bias toward shorter examples. In Section 5.4, we empirically examine whether the selected subsets exhibit high example quality and broad representation-space diversity.

4.2 Computational Complexity and Efficiency

We analyze the efficiency of H-MSAE from three complementary aspects: lightweight hierarchy construction, sparse subset selection, and the avoidance of full-pool generative annotation.

Lightweight hierarchy construction. H-MSAE constructs cross-resolution relations by accumulating sparse co-activations between adjacent resolutions and assigning each child concept to a bounded constant number of candidate parents. With at most $O(k)$ active concepts per example, each example contributes at most $O(k^2)$ active cross-resolution pairs. Over a pool of N examples, hierarchy construction therefore requires approximately

$$O(Nk^2 + E), \tag{13}$$

where E is the number of distinct observed cross-resolution concept pairs. Since each non-root concept has a constant number of parents, the final hierarchy contains $O(D)$ edges and requires $O(D)$ storage, with $O(E)$ temporary storage during construction. In contrast to MIG's explicit graph over generated semantic tags, H-MSAE directly constructs a compact hierarchy over learned sparse concepts.

Sparse subset selection. Given the hierarchy, marginal-gain computation only accesses the active concepts of a candidate and their propagated ancestors. With $k \ll D$ active concepts and L hierarchy resolutions, the per-candidate cost is approximately $O(kL)$. Stochastic greedy evaluates only $s = \left\lceil \frac{N}{B} \log \frac{1}{\varepsilon_{\text{sg}}} \right\rceil$ candidates at each of the B selection steps, yielding an approximate selection cost of

$$O(Bskl). \tag{14}$$

Length-aware quotas add only constant-time feasibility checks when bin assignments and remaining capacities are maintained. The sampling budget follows standard stochastic greedy; we do not assume that its original cardinality-constrained approximation guarantee directly extends to the length-constrained setting.

Avoiding full-pool annotation. The above analysis concerns hierarchy construction and subset selection. End-to-end cost also depends on how the semantic representation is obtained. Tag-based selectors require generative annotation over the candidate pool, whereas H-MSAE learns sparse concepts from frozen conversation embeddings through an MSAE and therefore avoids full-pool generative tagging. This provides an additional source of end-to-end efficiency beyond sparse hierarchy construction and selection.

Section 5.3 empirically evaluates these advantages through end-to-end subset-construction time.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We select 50K examples from the 939K-example Tulu 3 pool and compare H-MSAE with Uniform Random (Xia et al., 2025), ZIP (Yin et al., 2024), IFD (Li et al., 2024), #InsTag (Lu et al., 2023), DEITA (Liu et al., 2024b), CaR (Ge et al., 2024), QDIT (Bukharin et al., 2024), and MIG (Chen et al., 2025). Full Tulu 3 is included only as a full-data reference.

Our primary comparisons use three-epoch full-parameter SFT with a 4K context window and packing. We evaluate six objective benchmarks and three subjective benchmarks; Avg_{obj} and Avg_{sub} denote their respective aggregates, and Final Avg is their equally weighted mean. Detailed model, training, selection, and evaluation configurations are provided in Appendix A.1.

5.2 MAIN RESULTS

Table 1: Qwen3-Base results across 1.7B, 4B, and 8B scales. All rows use 50K selected examples unless marked Full. Best, second-best distinct, and third-best distinct comparable 50K results are shown in bold, underline, and italics, respectively.

Base Model	Method	Data Size	ARC	BBH	GSM	HE	MMLU	IFEval	Avg_{obj}	AE	MT	Wild	Avg_{sub}	Avg
Qwen3-1.7B-Base	Full Tulu 3	939K	76.27	55.29	66.57	61.59	61.66	56.75	63.02	2.36	5.70	-60.71	26.33	44.68
	Uniform Random	50K	73.56	57.41	69.37	57.93	59.11	51.76	61.52	1.61	5.66	-65.64	25.13	43.33
	ZIP	50K	74.24	57.26	69.14	60.98	59.73	47.32	61.45	1.24	5.84	-71.71	24.59	43.02
	IFD	50K	74.24	58.11	67.55	59.76	61.42	48.24	61.55	2.73	5.74	**-57.47**	27.13	44.34
	#InsTag	50K	*74.58*	57.73	66.64	56.10	60.02	49.72	60.80	2.11	5.99	-63.17	26.81	43.80
	DEITA	50K	72.88	55.69	69.83	**61.59**	59.70	48.80	61.41	1.49	*6.14*	-65.04	26.79	44.10
	CaR	50K	**77.63**	**58.34**	68.54	56.71	61.44	49.91	62.10	3.11	5.71	-58.60	26.97	44.53
	QDIT	50K	76.27	57.63	*70.36*	57.32	61.66	48.43	61.95	2.61	6.06	*-57.83*	*28.10*	*45.02*
	MIG	50K	72.20	58.32	71.49	57.32	*61.49*	*51.02*	*61.97*	*2.98*	**6.39**	-61.78	28.66	45.32
	H-MSAE	50K	73.22	*58.13*	**72.25**	*60.37*	**62.30**	**52.87**	**63.19**	**3.35**	6.38	-57.78	**29.42**	**46.31**
Qwen3-4B-Base	Full Tulu 3	939K	88.47	72.89	82.87	78.66	72.46	72.46	77.97	4.60	7.10	-33.27	36.32	57.15
	Uniform Random	50K	**86.44**	75.54	83.47	68.90	71.68	72.09	76.35	3.85	7.10	-41.31	34.73	55.54
	ZIP	50K	85.42	76.81	85.52	73.17	71.04	71.16	*77.19*	3.48	6.54	-50.10	31.28	54.23
	IFD	50K	86.10	*76.82*	82.18	**73.78**	73.20	68.58	76.78	6.34	6.96	-33.15	36.45	56.62
	#InsTag	50K	**86.44**	75.89	84.61	69.51	71.73	*73.38*	76.93	*6.27*	7.30	-35.24	*37.20*	*57.06*
	DEITA	50K	*85.76*	75.60	**87.64**	69.51	71.53	71.90	76.99	4.22	6.62	-40.61	33.37	55.18
	CaR	50K	*85.76*	75.12	*85.82*	71.34	72.12	70.98	76.86	4.84	7.01	-32.39	36.25	56.55
	QDIT	50K	*85.76*	**77.16**	84.31	71.34	*73.08*	71.53	77.20	5.34	7.04	*-32.59*	36.48	56.84
	MIG	50K	81.02	76.87	86.43	68.29	72.75	**74.49**	76.64	**7.33**	**7.38**	-33.41	38.14	57.39
	H-MSAE	50K	82.71	**77.16**	83.62	*72.56*	**73.76**	74.31	**77.35**	**7.33**	*7.22*	**-25.07**	**39.00**	**58.18**
Qwen3-8B-Base	Full Tulu 3	939K	90.85	73.98	83.24	83.54	77.94	70.61	80.03	6.83	5.87	-33.04	33.00	56.51
	Uniform Random	50K	*91.53*	74.94	70.43	85.98	66.74	64.88	75.75	5.59	7.41	-25.93	38.91	57.33
	ZIP	50K	90.85	47.97	80.14	78.05	78.18	48.61	70.63	1.74	6.58	-43.22	31.98	51.30
	IFD	50K	*91.53*	**79.63**	**87.19**	84.15	77.76	59.15	79.90	*10.81*	7.29	-24.64	*40.46*	60.18
	#InsTag	50K	91.19	73.66	82.18	83.54	77.99	64.70	78.88	5.22	7.31	*-24.05*	38.76	58.82
	DEITA	50K	90.51	73.22	82.56	81.10	**78.28**	66.91	78.76	5.09	7.19	-33.88	36.68	57.72
	CaR	50K	*91.53*	75.52	82.49	82.32	77.89	58.78	78.09	9.81	7.51	-28.86	40.16	59.12
	QDIT	50K	90.85	78.87	87.04	**86.59**	77.66	59.15	*80.03*	11.55	*7.42*	-27.00	40.75	*60.39*
	MIG	50K	92.88	*78.49*	*86.20*	*85.37*	77.53	**67.84**	**81.38**	8.57	7.25	-22.87	39.88	60.63
	H-MSAE	50K	**93.22**	78.44	84.08	81.10	*78.01*	*66.73*	80.26	**12.42**	**7.55**	**-18.10**	**42.96**	**61.61**

Consistent performance across scales. Table 1 shows that H-MSAE achieves the highest Final Avg among the evaluated 50K selectors across all three Qwen3 scales, indicating that its effectiveness is not confined to a particular model size. Compared with competing selection methods, H-MSAE exhibits more pronounced gains in subjective evaluation than in objective performance across the three scales. This suggests that quality-weighted hierarchical coverage may be particularly beneficial for subjective evaluation, highlighting the importance of considering both dimensions when assessing instruction data selection.

Random and full-data controls. Across all three scales, H-MSAE consistently outperforms both Uniform Random and Full Tulu 3, despite using substantially fewer examples than the latter. These results highlight the importance of data selection beyond simply increasing training data volume under the reported training protocol. Additional results on different model initializations and architectures are provided in Appendix A.2.

5.3 SELECTION EFFICIENCY

We measure the cost of selecting 50K examples from Tulu 3 using complete end-to-end pipelines and separately profile structure construction and core sampling after method-specific prerequisites are available. All experiments are conducted on a single B200 GPU with 32 CPU cores and 256 GB of host memory.

Table 2: End-to-end and selection-stage efficiency for selecting 50K examples from the full Tulu 3 pool. Relative costs are normalized by H-MSAE within each timing scope. End-to-end measurements include all method-specific prerequisites. Running times are medians over three profiling runs.

	Accounted total		Major prerequisite costs			Structure construction		Core sampling / filtering	
Method	Total	Rel.	Embedding	LLM tagging	Scoring	Time	Rel.	Time	Rel.
H-MSAE	**7h22m**	1.00×	3h13m	–	4h02m	**7.86 s**	1.00×	28.89 s	1.00×
MIG	36h34m	4.96×	–	28h28m	7h40m	37.44 s	4.76×	1,296.16 s	44.87×
#InsTag	28h28m	3.86×	–	28h28m	–	–	–	**7.65 s**	0.26×
DEITA	11h38m	1.58×	3h47m	–	7h40m	–	–	76.02 s	2.63×

†In the prerequisite stage, H-MSAE requires additionally only 5m18s for MSAE training and sparse encoding. Dashes indicate inapplicable costs.

End-to-end efficiency. As shown in Table 2, H-MSAE reduces end-to-end runtime by approximately $4.96\times$ relative to MIG and $3.86\times$ relative to #InsTag. The main difference lies in prerequisite computation: #InsTag and MIG require expensive full-pool generative tagging, while DEITA incurs substantial embedding and scoring costs. Although quality scoring and embedding account for approximately 98% of H-MSAE's own runtime, MSAE training and sparse encoding add only modest overhead. By learning sparse concepts from frozen conversation embeddings, H-MSAE avoids the full-pool generative annotation required by tag-based selectors. Moreover, H-MSAE uses only quality scores, whereas DEITA and MIG require both quality and complexity scores, reducing its scoring cost by approximately half. Notably, even without quality scoring, H-MSAE achieves a Final Avg of 60.54, comparable to MIG's 60.63, while reducing end-to-end runtime to approximately 3h20m, an $11\times$ speedup over MIG.

Structure construction and sampling. With prerequisite representations and annotations cached, H-MSAE further reduces structure-construction and core sampling time relative to MIG by $4.8\times$ and $44.9\times$, respectively. The former reflects lightweight cross-resolution relations derived from sparse co-activations, avoiding MIG's explicit tag-graph construction. The latter is consistent with the combination of stochastic greedy candidate evaluation and sparse marginal-gain computation, which accesses only active concepts and their propagated ancestors. These results support the two sources of computational savings discussed in Section 4.2: avoiding expensive full-pool annotation and reducing the cost of structure construction and subset selection through sparse computation. While #InsTag employs a faster core sampling procedure, its overall runtime is dominated by generative tagging.

5.4 Quality and Diversity Diagnostics

Our local utility analysis in Section 4.1 suggests that example quality and directional diversity contribute complementary terms to expected subset utility. To examine whether the selected subsets exhibit these desirable properties, we use the official DEITA quality score as a proxy for supervision quality and the effective rank (eRank) of frozen conversation embeddings as a measure of representation-space diversity. These empirical metrics are motivated by, but do not directly measure, the gradient-based quantities $Q(S)$ and $D(S)$ in Theorem 4.1.

Table 3 reveals distinct quality–diversity profiles among the evaluated selectors. Random selection retains relatively broad representation-space diversity but has lower average quality, whereas MIG selects higher-quality examples with more concentrated representations. In contrast, H-MSAE achieves the highest mean quality and eRank simultaneously, indicating that quality-weighted hierarchical coverage can favor high-quality examples without sacrificing broad latent support.

These findings are consistent with the quality–diversity interpretation in Section 4.1, providing empirical support for combining quality weighting with diminishing-return hierarchical coverage.

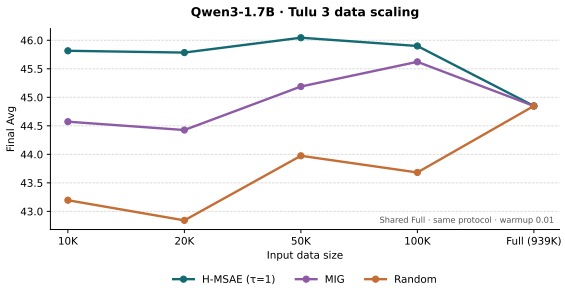
5.5 Ablation and Scaling

Component ablation. We conduct controlled ablations on Qwen3-1.7B to examine four components of H-MSAE: multi-resolution reconstruction, ancestor propagation, quality weighting, and length control. All variants use the same frozen conversation embeddings, candidate pool, 50K selection budget, and 256-step SFT protocol. We remove quality weighting and length control individually

Table 3: Quality and diversity diagnostics on the same frozen 50K selections. Q mean is the average DEITA quality score, while eRank measures the effective rank of the selected representation space.

Method	Q mean ↑	eRank ↑
Random	3.848	156.07
#InsTag	4.049	154.10
DEITA	3.776	150.12
MIG	4.598	149.13
H-MSAE	**4.993**	**158.15**

Figure 3: Data scaling on Qwen3-1.7B-Base. Final Avg across selection budgets of 10K, 20K, 50K, and 100K, with Full Tulu 3 as a reference.



and jointly, disable ancestor propagation, and separately ablate multi-resolution reconstruction while retaining the MSAE architecture.

Table 4 shows that the full configuration achieves the highest overall Avg. Removing either quality weighting or length control reduces performance, with their joint removal producing the largest degradation. This is consistent with their complementary roles: quality weighting favors informative supervision, while length constraints regulate the resulting subset composition. Disabling ancestor propagation also reduces both evaluation aggregates, supporting the use of cross-resolution relations rather than treating sparse concepts independently. Finally, multi-resolution reconstruction improves the overall result primarily through objective evaluation, while its removal yields a slightly higher subjective score. These findings support the combined design, although the contributions of individual components differ across evaluation dimensions.

Table 4: Controlled component ablation on Qwen3-1.7B. The multi-resolution reconstruction ablation removes the broad- and middle-resolution reconstruction objectives by assigning reconstruction loss with weights $(0, 0, 1)$.

Variant	Multi-res. recon.	Ancestor prop.	Quality	Length	Avg_{obj} ↑	Avg_{sub} ↑	Avg ↑
Full H-MSAE	✓	✓	✓	✓	**63.53**	30.06	**46.79**
w/o quality	✓	✓	×	✓	62.60	28.11	45.35
w/o length control	✓	✓	✓	×	62.56	28.84	45.70
w/o quality & length	✓	✓	×	×	62.71	27.44	45.07
w/o ancestor prop.	✓	×	✓	✓	62.56	29.27	45.92
w/o multi-res. recon.	×	✓	✓	✓	62.36	30.23	46.29

Data scaling. We independently rerun H-MSAE, MIG, and Uniform Random at selection budgets of 10K, 20K, 50K, and 100K, recomputing the length quotas and selected subsets for each budget. As shown in Figure 3, H-MSAE consistently outperforms both baselines across the evaluated budgets, indicating that its effectiveness is not specific to the default 50K setting. Performance improves substantially from 10K to 50K examples but largely saturates thereafter, suggesting diminishing returns from further increasing the selection budget under the current training protocol.

6 Conclusion

We introduced H-MSAE, a hierarchical sparse instruction selection framework that combines multi-resolution semantic coverage, quality weighting, and length-aware subset composition. On Tulu 3, H-MSAE achieves the highest Final Avg among the evaluated 50K selectors across three scales and an instruction-tuned initialization. Quality and diversity diagnostics, component ablations, and data-scaling experiments further support the combined design. In addition to its downstream performance, H-MSAE reduces end-to-end selection time by approximately $5\times$ relative to MIG, with substantially lower structure-construction and core sampling costs. Together, these results demonstrate the potential of hierarchical sparse concepts as an effective and computationally efficient alternative to generated semantic tags and explicit tag-graph construction for instruction data selection. Future work will investigate the transferability of learned sparse concepts across instruction domains and model architectures, as well as their application to larger-scale and dynamically evolving training corpora.

AI USE STATEMENT

In this work, we used generative AI tools to assist with intermediate mathematical derivations and the writing and checking of parts of the theoretical proofs, the implementation and debugging of the proposed method and experimental code, and the interpretation and organization of experimental results. All AI-assisted mathematical derivations and proofs were manually verified and revised by the authors, and AI-assisted implementations were tested for correctness.

We did not use generative AI tools to develop the conceptual or methodological framework, formulate research hypotheses or mathematical claims, or design or provide feedback on the experiments. We also did not use generative AI tools to generate synthetic datasets, perform translation, clean or reformat datasets, or conduct qualitative or thematic data analysis.

Additionally, we used generative AI tools to assist with searching and summarizing related literature, manuscript drafting and organization, and improving readability and grammar.

We reviewed all AI-assisted work before incorporating it into this paper. Mathematical derivations and proofs were independently checked and revised by the authors, experimental implementations and scripts were tested for correctness, numerical results were verified against the corresponding experimental outputs, and references and manuscript claims were manually reviewed. We take responsibility for the final content of this work, including text, claims, code, analyses, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings in Appendix A.1, including data preprocessing, sparse representation learning, hierarchy construction, quality weighting, length-aware selection, and evaluation protocols. The appendix specifies the principal selection hyperparameters, benchmark suites, and aggregate metric definitions. Our experiments use publicly released baseline implementations, with downstream comparisons conducted under our reported training and evaluation settings. We will release our code and experimental configurations upon acceptance.

ETHICS STATEMENT

This work is a fundamental algorithmic study of instruction data selection using publicly available datasets. It does not involve human-subject experiments or the collection of new personal data.

REFERENCES

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. Data diversity matters for robust instruction tuning. In *EMNLP*, 2024.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. *ICML*, 2025.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a better alpaca with fewer data. In *ICLR*, 2024.

Yicheng Chen, Yining Li, Kai Hu, Zerun Ma, Haochen Ye, and Kai Chen. Mig: Automatic data selection for instruction tuning by maximizing information gain in semantic space. In *Findings of ACL*, 2025.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy. In *ICLR*, 2026.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *ICLR*, 2025.

Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Mahong Xia, Zhang Li, Boxing Chen, Hao Yang, et al. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. In *EMNLP*, 2024.

Hamish Ivison, Muru Zhang, Faeze Brahman, Pang Wei Koh, and Pradeep Dasigi. Large-scale data selection for instruction tuning. *arXiv preprint arXiv:2503.01807*, 2025.

Yi Jing, Zao Dai, Jinwu Hu, Zijun Yao, Lei Hou, Juanzi Li, and Xiaozhi Wang. Guiding llm post-training data engineering with model internals from sparse autoencoders. *arXiv preprint arXiv:2605.27354*, 2026.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *NAACL*, 2024.

Liangxin Liu, Xuebo Liu, Derek F. Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection. In *NeurIPS*, 2024a.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *ICLR*, 2024b.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. #InsTag: Instruction tagging for analyzing supervised fine-tuning of large language models. *arXiv preprint arXiv:2308.07074*, 2023.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Gouki Minegishi, Hiroki Furuta, Yusuke Iwasawa, and Yutaka Matsuo. Rethinking evaluation of sparse autoencoders through the representation of polysemous words. In *ICLR*, 2025.

Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrak, and Andreas Krause. Lazier than lazy greedy. In *AAAI*, 2015.

Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

Zihan Niu, Wenping Hu, Junmin Chen, Xiyue Wang, Tong Xu, and Ruiming Tang. From tags to trees: Structuring fine-grained knowledge for controllable data selection in llm instruction tuning. *arXiv preprint arXiv:2601.13995*, 2026.

Weilin Wan, Weizhong Zhang, and Cheng Jin. Computational budget should be considered in data selection. *arXiv preprint arXiv:2510.16806*, 2025.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. In *ICML*, 2024.

Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. Rethinking data selection at scale: Random selection is almost all you need. In *EMNLP (Findings)*, 2025.

Xianjun Yang, Shaoliang Nie, Lijuan Liu, Suchin Gururangan, Ujjwal Karn, Rui Hou, Madian Khabsa, and Yuning Mao. Diversity-driven data selection for language model tuning through sparse autoencoder. *arXiv preprint arXiv:2502.14050*, 2025.

Mingjia Yin, Chuhan Wu, Yufei Wang, Hao Wang, Wei Guo, Yasheng Wang, Yong Liu, Ruiming Tang, Defu Lian, and Enhong Chen. Entropy law: The story behind data compression and llm performance. *arXiv preprint arXiv:2407.06645*, 2024.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. In *ICML*, 2024.

A Appendix

A.1 Detailed Experimental Settings

Data and sparse representations. We use the Tulu 3 instruction pool containing $939{,}343$ conversations. After filtering examples whose serialized character length exceeds $22{,}500$, the eligible pool contains $929{,}647$ examples. Each complete conversation is encoded into a frozen 4096-dimensional representation using Qwen3-8B. We adopt a three-resolution MSAE (Bussmann et al., 2025) with nested widths of $1{,}024$, $4{,}096$, and $16{,}384$, using ReLU activations and global BatchTopK sparsification. We retain up to 64 active features per example for subsequent hierarchy construction and selection.

Hierarchy and coverage. We construct the hierarchy using activation-support containment between adjacent resolutions. Each child concept retains its top $K_p = 4$ candidate parents, with soft parent weights computed using temperature $T_p = 0.1$. Ancestor propagation uses a depth-decay factor of $\rho = 0.5$. For quality weighting, we use the released DEITA quality scorer (Liu et al., 2024b) and compute the expected score over its six output labels rather than taking the highest-scoring label. We normalize the resulting scores to $[0, 1]$ and set the quality exponent to $\beta = 0.5$. The concave coverage function uses $\gamma = 0.85$.

Length-aware selection. We partition the eligible pool into $m = 40$ quantile bins according to serialized conversation length. By default, we set $\tau = 1$, allocating each bin a quota proportional to its mean length, with integer quotas normalized to the selection budget. Our primary experiments select $B = 50{,}000$ examples. For stochastic greedy, we use $\varepsilon_{\mathrm{sg}} = 0.1$, corresponding to at most 43 sampled feasible candidates per iteration. The final selected index set $\mathcal{I}_B$ contains $50{,}000$ unique source indices.

SFT training. For the primary comparisons, we perform three-epoch full-parameter SFT with a 4K context window and sequence packing. All downstream SFT experiments are conducted on HC accelerator cards, whereas data selection and efficiency profiling are performed on a single NVIDIA B200 GPU with 32 CPU cores and 256 GB of host memory.

Evaluation. We evaluate objective performance on ARC-Challenge, BBH, GSM8K, HumanEval pass@1, MMLU, and IFEval, reporting their unweighted mean as $\mathrm{Avg}_{\mathrm{obj}}$. Preference evaluation uses CompassJudger-1-32B on AlpacaEval-v2, MT-Bench, and WildBench. Following the comparison protocol, we compute

$$\mathrm{Avg}_{\mathrm{sub}} = \frac{\mathrm{AE} + 10\,\mathrm{MT} + (\mathrm{Wild} + 100)/2}{3}, \qquad \mathrm{Avg} = \frac{\mathrm{Avg}_{\mathrm{obj}} + \mathrm{Avg}_{\mathrm{sub}}}{2}. \tag{15}$$

A.2 Additional Experimental Results

Different model initialization. Table 5 extends the comparison to Qwen3-8B-Instruct. H-MSAE achieves the highest objective and subjective aggregates among the evaluated selectors, indicating that its selected subset remains effective when SFT starts from an instruction-tuned model rather than a Base initialization.

Table 5: Results on Qwen3-8B-Instruct under the 50K subset-selection setting. Full Tulu 3 is shown as a full-data reference and excluded from ranking. Best, second-best distinct, and third-best distinct 50K results are shown in bold, underline, and italics, respectively.

Base Model	Method	Data Size	ARC	BBH	GSM	HE	MMLU	IFEval	$\mathrm{Avg}_{\mathrm{obj}}$	AE	MT	Wild	$\mathrm{Avg}_{\mathrm{sub}}$	Avg
Qwen3-8B-Instruct	Full Tulu 3	939K	86.10	75.96	82.26	73.78	77.19	75.05	78.39	4.84	7.29	-25.99	38.25	58.32
	Uniform Random	50K	*89.49*	79.65	84.91	76.83	76.44	*75.23*	80.42	4.22	7.11	-30.17	36.74	58.58
	ZIP	50K	90.51	80.44	87.34	36.59	76.53	72.83	74.04	3.23	7.03	-42.51	34.09	54.07
	IFD	50K	87.12	80.53	83.02	**82.32**	76.89	72.64	80.42	9.19	*7.46*	-23.88	40.62	60.52
	#InsTag	50K	**91.19**	79.10	83.47	*77.44*	*76.72*	*75.23*	80.53	5.34	7.41	-27.35	38.59	59.56
	DEITA	50K	90.51	78.93	**87.64**	76.83	76.54	74.49	80.82	5.47	7.15	-32.88	36.84	58.83
	CaR	50K	90.51	78.33	86.28	76.83	**77.18**	73.38	80.42	7.83	7.47	*-25.95*	*39.85*	60.13
	QDIT	50K	89.15	*80.89*	84.76	78.05	76.64	74.49	80.66	*7.95*	7.47	-26.51	39.80	*60.23*
	MIG	50K	87.46	**81.17**	87.34	75.61	75.75	**76.71**	*80.67*	*7.95*	7.40	-28.24	39.28	59.98
	H-MSAE	50K	89.15	81.10	*86.50*	78.05	76.69	75.97	**81.24**	**9.94**	**7.75**	**-19.14**	**42.62**	**61.93**

Table 6: Cross-family evaluation on Llama-3.1-8B and Mistral-7B-v0.3 under our reproduced 4K-context SFT protocol. All selection methods use 50K examples unless marked Full. Best, second-best distinct, and third-best distinct results among the evaluated 50K selectors are shown in bold, underline, and italics, respectively. Full Tulu 3 is a full-data reference and is excluded from ranking.

Base Model	Method	Data Size	ARC	BBH	GSM	HE	MMLU	IFEval	Avg_{obj}	AE	MT	Wild	Avg_{sub}	Avg
Llama-3.1-8B	Full Tulu 3	939K	80.68	69.27	76.65	62.20	66.14	67.65	70.43	6.58	6.75	-27.70	36.74	53.59
	Uniform Random	50K	77.97	67.55	64.75	**57.32**	65.09	*54.71*	**64.56**	7.08	6.88	-35.42	36.06	50.31
	ZIP	50K	79.66	66.89	65.05	44.51	65.78	56.01	62.98	7.83	6.76	-45.96	34.15	48.57
	IFD	50K	*80.00*	65.58	*67.32*	*52.44*	65.32	49.72	63.40	12.17	6.85	-25.99	*39.22*	51.31
	#InsTag	50K	**83.39**	67.41	65.43	47.56	**66.09**	**56.19**	64.35	7.95	6.47	-32.98	35.39	49.87
	DEITA	50K	*80.00*	66.55	67.55	46.95	65.05	56.01	63.69	9.69	6.91	-37.14	36.74	50.21
	CaR	50K	82.03	**68.25**	66.11	51.83	*65.70*	51.76	*64.28*	11.43	6.97	-28.79	38.91	51.60
	QDIT	50K	78.98	67.39	66.34	*52.44*	65.38	51.94	63.74	**14.04**	6.81	*-28.03*	39.38	*51.56*
	MIG	50K	69.15	*67.47*	**68.54**	53.05	64.71	52.87	62.63	*13.29*	*6.95*	-31.94	38.94	50.79
	H-MSAE	50K	72.88	65.75	65.35	53.05	65.12	56.01	63.03	13.79	**7.09**	**-25.59**	**40.63**	**51.83**
Mistral-7B-v0.3	Full Tulu 3	939K	35.25	58.72	72.55	62.80	61.55	73.01	60.65	9.07	6.97	-22.54	39.17	49.91
	Uniform Random	50K	75.59	*59.46*	60.50	46.34	60.60	64.33	61.14	5.84	*6.99*	-28.78	37.12	49.13
	ZIP	50K	78.98	59.78	51.78	34.15	60.98	64.51	58.36	5.96	6.61	-39.83	34.05	46.21
	IFD	50K	75.59	49.57	53.98	42.68	62.03	57.30	56.86	*14.16*	6.72	-22.88	*39.97*	48.42
	#InsTag	50K	76.27	**59.81**	61.49	40.24	62.01	*66.91*	*61.12*	5.59	**7.15**	-27.09	37.85	49.49
	DEITA	50K	*76.61*	59.07	61.49	38.41	61.38	62.85	59.97	6.58	*6.99*	-32.18	36.80	48.38
	CaR	50K	*76.61*	59.06	59.51	**48.17**	62.50	58.23	60.68	11.93	6.85	-24.92	39.32	50.00
	QDIT	50K	**80.00**	59.02	60.80	40.24	**62.91**	61.55	60.75	14.29	6.55	-20.43	39.86	*50.31*
	MIG	50K	72.20	58.48	*61.11*	46.34	61.39	**68.95**	**61.41**	13.17	7.08	*-27.92*	41.00	51.21
	H-MSAE	50K	75.25	50.70	**62.62**	*45.73*	*62.04*	67.65	60.66	**15.53**	6.96	**-18.91**	**41.89**	**51.28**

Cross-family evaluation. Table 6 extends our evaluation to Llama-3.1-8B and Mistral-7B-v0.3. On Llama-3.1-8B, H-MSAE achieves the highest Final Avg among the evaluated 50K selectors, outperforming MIG by 1.04 points, with a more pronounced improvement in subjective evaluation. On Mistral-7B-v0.3, H-MSAE and MIG achieve comparable Final Avg, with H-MSAE achieving a slight 0.07-point advantage. These results indicate that the selected subset remains effective beyond the Qwen3 family, although the magnitude of the improvement varies across architectures.

Note that here we use the publicly released implementations of the competing methods and evaluate them under our unified training and evaluation protocol. All methods are retrained on our HC hardware with sequence packing to ensure consistent comparisons. The resulting scores can be a bit different from those reported in the original MIG paper, potentially due to differences in hardware, packing, or other implementation details. We therefore focus on comparisons within our experimental setting rather than directly comparing absolute scores with the published results.

Additional data pools. We further evaluate H-MSAE against MIG on OpenHermes and Xsota to examine its effectiveness beyond the primary Tulu 3 pool. We use $\tau = 2$ for this setting. As shown in Table 7, H-MSAE outperforms MIG on both OpenHermes and Xsota using Qwen3-4B.

Table 7: Additional evaluation on OpenHermes and Xsota. Final Avg is compared with MIG under the same model and data pool. Δ is the difference relative to MIG.

Model	Data pool	Configuration	H-MSAE	MIG	Δ
Qwen3-4B-Base	OpenHermes	BF16	53.57	50.37	+3.19
Qwen3-4B-Base	Xsota	BF16	53.93	52.20	+1.74

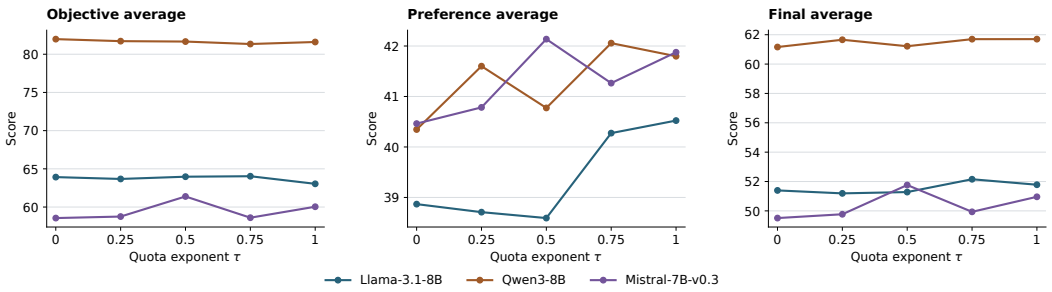


Figure 4: Sensitivity to the length-quota exponent τ across three model families. The panels report objective, subjective, and Final Avg scores for selection configurations with $\tau \in \{0, 0.25, 0.5, 0.75, 1\}$.

Sensitivity to length-stratification. We investigate the sensitivity of H-MSAE to the length-quota exponent τ. As shown in Figure 4, objective performance remains relatively stable across different values of τ, while subjective performance exhibits greater variation and tends to stabilize around $\tau = 0.75$. Despite these differences, Final Avg is comparatively insensitive to τ, showing no consistent improvement with stronger length reweighting. These results suggest that H-MSAE is relatively robust to the choice of τ in terms of overall performance, although length allocation has a more noticeable effect on subjective evaluation.

A.3 Proof of the Quality–Diversity Decomposition

This section presents the modeling details and proof of Theorem 4.1, using the notation introduced in Section 4.1.

Modeling details. We model the covariance of each stochastic training gradient as

$$\mathrm{Cov}(g_i) = (1 - q_i)\Sigma, \qquad \Sigma \succeq 0, \tag{16}$$

where $q_i \in [0, 1]$ represents latent example quality and Σ is a shared covariance matrix. Thus, higher-quality examples have lower gradient uncertainty. We assume uncorrelated gradient fluctuations across distinct examples:

$$\mathrm{Cov}(g_i, g_j) = \mathbb{E}\big[(g_i - \bar{g}_i)(g_j - \bar{g}_j)^\top\big] = 0, \qquad i \neq j. \tag{17}$$

To isolate the effects of quality and directional diversity, we consider a homogeneous local regime:

$$g_{\mathcal{T}}^\top \bar{g}_i = \mu, \qquad \bar{g}_i^\top H_{\mathcal{T}} \bar{g}_i = \chi > 0, \qquad \forall i. \tag{18}$$

These conditions fix the first-order target alignment and self-curvature across candidate examples. We further define

$$\nu = \mathrm{tr}(H_{\mathcal{T}}\Sigma) \geq 0. \tag{19}$$

Finally, we assume that $R_{\mathcal{T}}$ has a locally Lipschitz-continuous Hessian and that the stochastic subset-average gradient has a finite third moment. These conditions ensure that the expected remainder in the second-order Taylor expansion of $R_{\mathcal{T}}(\theta_0 + \Delta_S)$ is $o(\eta^2)$ as $\eta \to 0$, for each fixed subset S.

An equivalent form of directional diversity. Because $\|v_i\|_2 = 1$ and $D(S) = \mathrm{tr}(\widehat{\mathrm{Cov}}_{i \in S}[v_i])$ uses the $1/(B-1)$ sample-covariance normalization,

$$D(S) = 1 - \frac{2}{B(B-1)} \sum_{\substack{i,j \in S \\ i<j}} v_i^\top v_j. \tag{20}$$

Equivalently,

$$2 \sum_{\substack{i,j \in S \\ i<j}} v_i^\top v_j = B(B-1)\big(1 - D(S)\big). \tag{21}$$

This identity is used only to simplify the cross-example terms below.

Proof of Theorem 4.1. Applying a second-order Taylor expansion of the target risk around θ_0 to the local subset update, we obtain

$$U_{\mathcal{T}}(S) = -g_{\mathcal{T}}^\top \Delta_S - \frac{1}{2} \Delta_S^\top H_{\mathcal{T}} \Delta_S + o(\eta^2). \tag{22}$$

For the first-order term,

$$-\mathbb{E}[g_{\mathcal{T}}^\top \Delta_S] = \frac{\eta}{B} \sum_{i \in S} g_{\mathcal{T}}^\top \bar{g}_i = \eta\mu. \tag{23}$$

For the second-order term,

$$\mathbb{E}[\Delta_S^\top H_{\mathcal{T}} \Delta_S] = \frac{\eta^2}{B^2} \left[\sum_{i \in S} \mathbb{E}[g_i^\top H_{\mathcal{T}} g_i] + 2 \sum_{\substack{i,j \in S \\ i<j}} \mathbb{E}[g_i^\top H_{\mathcal{T}} g_j] \right]. \tag{24}$$

For each diagonal term,

$$\mathbb{E}[g_i^\top H_{\mathcal{T}} g_i] = \bar{g}_i^\top H_{\mathcal{T}} \bar{g}_i + \operatorname{tr}(H_{\mathcal{T}} \operatorname{Cov}(g_i)) = \chi + (1 - q_i)\nu. \tag{25}$$

Therefore,

$$\sum_{i \in S} \mathbb{E}[g_i^\top H_{\mathcal{T}} g_i] = B\chi + B\nu\big(1 - Q(S)\big). \tag{26}$$

For $i \neq j$, Eq. (17) gives

$$\mathbb{E}[g_i^\top H_{\mathcal{T}} g_j] = \bar{g}_i^\top H_{\mathcal{T}} \bar{g}_j. \tag{27}$$

From the definition $v_i = H_{\mathcal{T}}^{1/2} \bar{g}_i / \|H_{\mathcal{T}}^{1/2} \bar{g}_i\|_2$ and the homogeneous self-curvature assumption,

$$v_i^\top v_j = \frac{\bar{g}_i^\top H_{\mathcal{T}} \bar{g}_j}{\sqrt{\bar{g}_i^\top H_{\mathcal{T}} \bar{g}_i} \sqrt{\bar{g}_j^\top H_{\mathcal{T}} \bar{g}_j}} = \frac{\bar{g}_i^\top H_{\mathcal{T}} \bar{g}_j}{\chi}. \tag{28}$$

Hence,

$$\bar{g}_i^\top H_{\mathcal{T}} \bar{g}_j = \chi v_i^\top v_j. \tag{29}$$

Using Eq. (21),

$$2 \sum_{\substack{i,j \in S \\ i<j}} \mathbb{E}[g_i^\top H_{\mathcal{T}} g_j] = \chi B(B-1)\big(1 - D(S)\big). \tag{30}$$

Combining Eqs. (26) and (30) yields

$$\mathbb{E}[\Delta_S^\top H_{\mathcal{T}} \Delta_S] = \eta^2 \chi + \frac{\eta^2 \nu}{B} - \frac{\eta^2 \nu}{B} Q(S) - \frac{\eta^2 \chi (B-1)}{B} D(S). \tag{31}$$

Substituting into Eq. (22),

$$\mathbb{E}[U_{\mathcal{T}}(S)] = \eta\mu - \frac{\eta^2 \chi}{2} - \frac{\eta^2 \nu}{2B} + \frac{\eta^2 \nu}{2B} Q(S) + \frac{\eta^2 \chi (B-1)}{2B} D(S) + o(\eta^2) \tag{32}$$

$$= C_B + \lambda_Q Q(S) + \lambda_D D(S) + o(\eta^2), \tag{33}$$

where

$$C_B = \eta\mu - \frac{\eta^2 \chi}{2} - \frac{\eta^2 \nu}{2B}, \qquad \lambda_Q = \frac{\eta^2 \nu}{2B} \geq 0, \qquad \lambda_D = \frac{\eta^2 \chi (B-1)}{2B} > 0. \tag{34}$$

This proves Theorem 4.1. □