Dual-lens: Model-Aware Data Curation for Efficient and Effective Knowledge Recovery in Pruned Language Models

Anonymous ACL submission

Abstract

Recovering capabilities in pruned language models typically requires fine-tuning on large datasets, but often yields suboptimal results since the original pretraining data is unavailable for state-of-the-art foundation models. In this paper, we propose *Dual-lens*, a data curation framework that identifies compact, high-utility subsets from public corpora. Dual-lens combines two criteria: *CE-lens*, which targets samples the pruned model finds difficult, and *SAE-lens*, which ensures semantic coverage via sparse autoencoders trained on latent concept distributions. By performing a pipelined fine-tuning procedure with the two lens, the proposed framework balances model-specific correction and representational diversity. Experiments across various models, pruning schemes, and downstream tasks show that Dual-lens outperforms full-data tuning and recent baselines while using significantly less data, e.g., LLaMA 2.1 13B, pruned with 35% pruning ratio, achieves a 22% improvement in accuracy for downstream reasoning tasks using only 10% of the full corpus of Alpaca dataset.

1 Introduction

Structured pruning (Ling et al., 2024; Hu et al., 2025; Sandri et al., 2025) is a widely used strategy to reduce the inference cost and memory footprint of large language models (LLMs). However, pruning often leads to substantial degradation in the reasoning and generalization capabilities of the original models. A common remedy is to apply post-pruning fine-tuning to recover the lost functionality. Yet, this recovery process is fundamentally limited by the inaccessibility of the original pretraining data, which is typically proprietary or non-public. In such settings, recovery efforts must rely on substitute datasets that are publicly available but only loosely aligned with the training distribution of the original model.

This gap between the demands of pruned models and available data introduces two critical challenges. First, pruned models exhibit *predictive failures distinct from those of their full-capacity counterparts* due to reduced representational and computational capacity. Even with existing data selection strategies such as scoring based on full-model responses (Liu et al., 2024b; Wang et al., 2024) and estimating the quality of data samples (Cao et al., 2023), the selected samples could be misaligned with the unique deficiencies of the pruned models, leading to suboptimal fine-tuning performance. Second, public datasets are typically *distributionally divergent from the original pretraining data* and can also be noisy and redundant, which is particularly problematic when the pretraining corpus is unavailable. These factors undermine the effectiveness of unfiltered large-scale fine-tuning and risk exhausting the limited capacity of the pruned model on uninformative samples.

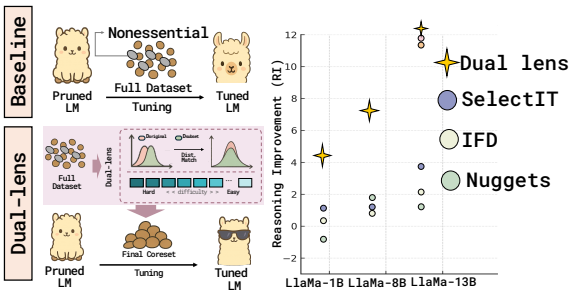

Figure 1: **Left:** Illustration of our subset–selection objective, which ranks and chooses the most informative data points for efficient model tuning. **Right:** Reasoning performance improvement (RI), computed as $RI = \text{Perform}_{\text{full}} - \text{Perform}_{\text{subset}}$ (higher scores indicate better performance) , for each data selection method. Here, $\text{Perform}_{\text{full}}$ is the score obtained when fine-tuning the LLM on the entire corpus, and $\text{Perform}_{\text{subset}}$ is the score achieved when tuning with the selected subset.

Based on these observations, we hypothesize that *a compact, strategically curated subset, selected using model-internal signals derived from the pruned model's own behavior can achieve more efficient and effective knowledge re-*

covery than full public dataset-based fine-tuning. In particular, we identify two complementary selection objectives aligned with the challenges outlined above without relying on the original training corpus: one that prioritizes examples where the model exhibits predictive uncertainty or failure, and another that ensures broad coverage of the model's internal semantic representations.

In this paper, we propose *Dual-lens*, a data curation framework that constructs a compact, informative subset for post-pruning fine-tuning. The framework integrates two complementary selection criteria that together capture distinct notions of sample utility.First, which we refer to as the *CE-lens*, identifies data points on which the pruned model exhibits high cross-entropy loss, thereby targeting samples that reveal its current weaknesses. This allows the fine-tuning process to concentrate on correcting specific knowledge gaps induced by pruning.Second, the *SAE-lens*, selects samples that preserve the internal concept distribution of the given model. We achieve this by training Sparse Autoencoders (SAEs) (Templeton et al., 2024; Karvonen et al., 2024) on the hidden representations of the pruned model to extract latent embeddings, then selecting a subset whose embedding distribution closely matches that of the full dataset.

These two criteria are intentionally orthogonal: the CE-lens weights pedagogical value by targeting difficult examples, while the SAE-lens promotes semantic coverage by modeling diversity in the latent concept space. Hence, the proposed framework integrates the two methods by applying SAE-lens first to identify a representative candidate pool, followed by CE-lens to select the most informative subset within it. It ensures that the selected data is both broadly representative and sharply focused on the residual deficiencies of the pruned model. Our evaluation results show that the complementarity is especially beneficial for aggressively pruned or capacity-constrained models.

To our knowledge, this is the first framework designed for knowledge recovery in pruned LLMs using explicitly model-sensitive data selection. Our contributions are summarized as follows:

1. We introduce *Dual-lens*, a principled data curation framework that unifies two model-aware criteria, cross entropy–based difficulty and latent-space coverage via sparse autoencoders, for efficient post-pruning recovery.

2. We devise two complementary selection mechanisms: *CE-lens*, which prioritizes samples that expose the residual deficiencies of the pruned model, and *SAE-lens*, which ensures semantic coverage by aligning the subset's latent distribution with that of the full corpus. Their integration balances distributional fidelity and corrective supervision, enabling high-quality recovery with as little as 10% of the original data.

3. We validate Dual-lens across various models (LLaMA 1B, 8B, 13B (Patterson et al., 2022; Grattafiori et al., 2024; Touvron et al., 2023)), pruning methods (LLM-Pruner (Ma et al., 2023), FLAP (An et al., 2023)), datasets (Alpaca (Taori et al., 2023), LaMini (Wu et al., 2023), Dolly (Ouyang et al., 2022)), and downstream tasks including reasoning (Clark et al., 2019; Zellers et al., 2019; Bisk et al., 2019; Clark et al., 2018) and math (Cobbe et al., 2021; Lewkowycz et al., 2022). Dual-lens consistently outperforms full-dataset tuning and state-of-the-art data selection methods such as IFD (Li et al., 2023a), SelectIT (Liu et al., 2024a), and Nuggets (Li et al., 2023b), achieving up to a 13% improvement in average reasoning accuracy and 25.5% reduction in perplexity compared to the full corpus tuning under a pruning ratio of 35%.

2 Related Work

2.1 Classical Data Selection

Prior work on data selection aimed to identify samples with high informational value for classification. Davis and Hwang (1992) applied geometric inversion techniques to select points near the decision boundary, improving classification accuracy by 6%. In parallel, Lewis (1995) introduced uncertainty sampling to identify difficult or ambiguous examples, a strategy that remains conceptually influential for modern LLM training.

2.2 Instruction-Tuning Data Selection

Recent work have developed various techniques to select a subset of data to train or finetune LLMs. For example, Deita (Liu et al., 2024b) introduces a framework that scores samples by complexity, quality, and diversity. Other approaches focus on diversity-aware objectives, such as determinantal point processes over gradient embeddings (Wang et al., 2024), or clustering-based refinement (Yu et al., 2024). Several works aim

to identify difficult examples for instruction tuning, including IFD (Li et al., 2023a), which proposes an Instruction-Following Difficulty metric, and Nuggets (Li et al., 2023b), a one-shot learning-based method that selects samples based on their anchor-set perplexity impact. SelectIT (Liu et al., 2024a) combines uncertainty estimation with self-reflection to score instruction samples. Instruction Mining (Cao et al., 2023) uses natural language indicators to identify useful subsets, motivated by the double descent phenomenon.

While these methods are effective for general-purpose instruction tuning, our work focuses on *pruned* models and introduces model-aware sample utility, i.e., difficulty from the loss landscape of a pruned model (*CE-lens*), and the coverage of its internal latent distribution (*SAE-lens*).

2.3 Curated Instruction-Tuning Benchmarks

Several public benchmarks provide curated instruction-tuning subsets. Alpagasus (Chen et al., 2023) contains 9,229 samples distilled from Alpaca using ChatGPT (OpenAI, 2023) to filter out low-quality examples. LIMA (Zhou et al., 2023) presents a 1,330-example dataset curated using the Superficial Alignment Hypothesis (Kirstain et al., 2021), focusing on high-quality, diverse instructions. While these datasets are valuable for evaluating tuning strategies, they are manually or heuristically filtered and not tailored to the needs of pruned models. In contrast, Dual-lens provides a generalizable, automated, and model-aware method for targeted knowledge recovery.

3 Methods

3.1 Overview of Dual-lens Sampling

Figure 2 illustrates the overall architecture of the Dual-lens framework, which constructs a compact yet effective data subset for fine-tuning pruned language models. Given a pruned model f and a publicly available dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, the objective is to identify a smaller subset $\mathcal{S}_{\text{dual-lens}} \subset \mathcal{D}$ that enables efficient and high-quality recovery of the model's lost capabilities.

To this end, Dual-lens combines two distinct selection strategies. The first strategy, called *CE-lens*, identifies training examples on which the pruned model incurs high cross-entropy loss, thereby capturing its current predictive weaknesses. The second strategy, called *SAE-lens*, selects examples that maintain coverage of the underlying semantic structure of the dataset, modeled through the latent activations of the pruned network. These two perspectives reflect complementary criteria, i.e., (i) *the difficulty* from the predictive behavior of the target model and (ii) *the representativeness* from its internal feature space. By integrating these two model-aware selection criteria, Dual-lens constructs a training subset that is both corrective and distributionally grounded.

3.2 Cross Entropy-based Difficulty Selection (CE-lens)

The CE-lens identifies samples that induce high predictive loss under the pruned model, focusing training on residual knowledge gaps. For each input $(x_i, y_i) \in \mathcal{D}$, we compute the cross-entropy loss:

$$\ell_i = \mathcal{L}_{\text{CE}}(f(x_i), y_i),$$

where f is the pruned model and ℓ_i reflects the discrepancy between the predicted distribution and the ground truth.

Samples are then ranked in descending order of loss (from hardest to easiest), and the top $M = \lfloor \rho N \rfloor$ are selected, where $\rho \in (0, 1)$ is a user-defined selection ratio. The resulting subset is:

$$\mathcal{S}_{\text{CE}} = \{(x_i, y_i) \in \mathcal{D} \mid \ell_i \text{ is among the top-}M \text{ losses}\}.$$

By focusing fine-tuning on high-loss samples, CE-lens encourages efficient gradient updates that address the weaknesses of the pruned model. Hence, this strategy helps the compact model make better use of its limited capacity, resulting in improved downstream performance.

We also explore a variant that computes ℓ_i using the original pretrained model f_{full} to examine how loss perception differs before and after pruning (see Section A.5). However, we observe that computing losses directly with the pruned model results in better alignment with its post-pruning recovery objective and leads to higher fine-tuning accuracy.

3.3 Latent Representation-based Coverage Selection (SAE-lens)

The SAE-lens aims to build a representative training subset by preserving the latent concept distribution of the full dataset. This is achieved by modeling the internal activations of the pruned model using a sparse autoencoder trained on Top-K neuron activations from the final transformer layer.

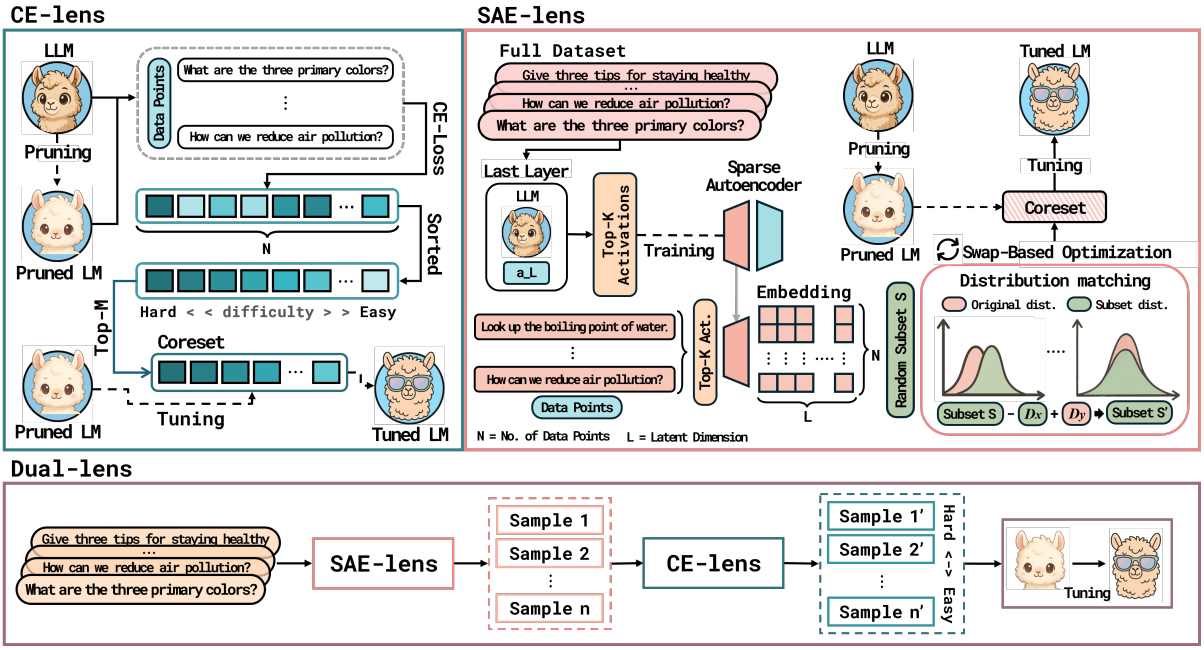

Figure 2: An overview of the Dual-lens framework for data selection. The top left panel illustrates CE-lens, where data samples are prioritized based on their difficulty and the top right panel depicts SAE-lens, which selects a data subset whose distribution closely matches that of the original corpus by utilizing a Sparse Autoencoder trained on the activations of the pruned model. The bottom panel shows Dual-lens, which integrates both approaches: it first uses SAE-lens to curate an initial set of samples and then applies CE-lens to further select a subset of these samples for fine-tuning the pruned model.

Latent activation extraction. Let L denote the final layer of the pruned model, and let $act^{(L)}(x)$ represent the hidden state vector at that layer for input x. Following Bhattacharyya and Kim (2025), we extract the K most salient activation dimensions based on their gradient magnitude, which helps eliminate noisy components and emphasizes the most informative features. Specifically, for each input x, we compute the element-wise gradient of the model's output with respect to the activations at layer L, and select the top K dimensions with the highest magnitudes. This filtering is performed per-example to focus on salient features. This yields:

$$A(x) = \text{TopK}\big(act^{(L)}(x),\, K\big) \in \mathbb{R}^K,$$

as a compact latent representation of x.

Training the sparse autoencoder. We train a sparse autoencoder f_θ to encode and reconstruct these latent representations. The encoder output $z = f_\theta(A(x))$ captures a low-dimensional, interpretable embedding of the input:

$$z = \text{ReLU}(W_e A(x) + b_e),$$

where $W_e \in \mathbb{R}^{d \times K}$. Sparsity is enforced on z to encourage factor disentanglement and compress information into a small number of dimensions, improving the interpretability and distributional coverage of the selected representations. This acts as a bottleneck that favors localized, semantically distinct features. The encoder is trained on the full dataset and used for all subsequent selection.

Subset selection via latent distribution alignment. After training the encoder, each data point x_i in the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ is mapped to an embedding $z_i = f_\theta(A(x_i))$. Let $\mathcal{S}_{SAE} \subset \mathcal{D}$ be a candidate subset. We denote by $\widehat{P}_{\mathcal{D}}$ and $\widehat{P}_{\mathcal{S}\text{SAE}}$ the empirical distributions over the latent embeddings $\{z_i\}_{i=1}^N$ for the full dataset and $\{z_j\}_{(x_j,y_j)\in\mathcal{S}_{\text{SAE}}}$ for the selected subset, respectively. To make $\mathcal{S}_{SAE}$ a distributionally faithful approximation of $\mathcal{D}$, we minimize the following discrepancy:

$$\Delta(\mathcal{S}_{SAE}) = w_B\, D_B\big(\widehat{P}_{\mathcal{D}}, \widehat{P}_{\mathcal{S}_{SAE}}\big) + w_{KS}\, D_{KS}\big(\widehat{P}_{\mathcal{D}}, \widehat{P}_{\mathcal{S}_{SAE}}\big),$$

where D_B denotes the Bhattacharyya distance (global overlap) and D_{KS} the two-sample Kolmogorov–Smirnov statistic (maximum quantile difference). w_B and w_{KS} are the weights corresponding for the two metrics, respectively. Lower values of $\Delta(\mathcal{S}_{SAE})$ indicate better alignment with the latent space distribution of the full dataset.

The subset $\mathcal{S}_{SAE}$ is initialized by random sampling and refined using a swap-based optimization

strategy that iteratively reduces $\Delta(\mathcal{S}_{SAE})$. The resulting selection maintains semantic diversity and coverage, enabling robust post-pruning recovery with a limited data budget.

3.4 Dual-lens Integration Strategy

The Dual-lens framework integrates the CE-lens and SAE-lens to combine their complementary objectives. While each lens can be applied independently, their integration provides a more balanced training subset that addresses both local model deficiencies and global distributional coverage.

Let $\mathcal{S}_{\text{SAE}} \subset \mathcal{D}$ denote the intermediate subset selected using the SAE-lens, so that $\mathcal{S}_{\text{SAE}} = \lfloor \rho' \cdot N \rfloor$ for a configurable ratio $\rho' \in (0, 1)$, which approximates the latent embedding distribution of the full dataset. From this subset, the CE-lens further identifies the most challenging samples based on their and CE losses computed with the pruned model. The final training subset $\mathcal{S}_{\text{dual-lens}}$ is defined as:

$$\mathcal{S}_{\text{dual-lens}} = \{(x_i, y_i) \in \mathcal{S}_{SAE} \mid \ell_i \text{ is among the top-}L \text{ losses}\}$$

where $\ell_i = \mathcal{L}_{\text{CE}}(f(x_i), y_i)$ and $L = \lfloor D_{pr} \cdot N \rfloor$ for a target data sampling ratio D_{pr}.

This integration ensures that the selected data subset is not only representative of the overall semantic space but also aligned with the specific learning needs of the pruned model.

4 Experiments

4.1 Experimental Setup

We evaluate Dual-lens on a diverse set of post-pruning fine-tuning tasks using three LLaMA-based models: LLaMA 2 13B (Touvron et al., 2023), LLaMA 3.1 8B (Grattafiori et al., 2024), and LLaMA 3.2 1B (Patterson et al., 2022). Unless otherwise specified, the hyperparameters for data selection in Dual-Lens and for fine-tuning are kept consistent across all methods to ensure a fair and controlled comparison. For dataset curation through Dual-lens and SAE-lens, we set w_B = 0.7 and w_{KS} = 0.3. To sample the initial Alpaca subset with SAE via Dual-lens, we select 90% of the original dataset. During finetuning the models, we set the learning rate to 5e-4, and lora ratio to 16.

To assess the generality of our approach across pruning techniques, we employ two distinct methods: LLM-Pruner (Ma et al., 2023), a structured pruning approach with tunable compression schedules, and FLAP (An et al., 2023), a training-free method that prioritizes architectural simplicity. The pruning ratio associated with each method is denoted as P_r throughout the remainder of the paper.

To benchmark the effectiveness of *Dual-lens*, we compare against two classes of baselines. The first includes full-dataset fine-tuning on Alpaca (Taori et al., 2023), LaMini (Wu et al., 2023), and GPT-J (Anand et al., 2023), as well as randomly sampled subsets from Alpaca (denoted "Random Selection"). The second group comprises state-of-the-art data selection methods which represent diverse selection philosophies, IFD (Li et al., 2023a), SelectIT (Liu et al., 2024a), and Nuggets (Li et al., 2023b). We evaluate performance on perplexity-based benchmarks (WikiText) and on commonsense reasoning datasets, including BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2019), and both the ARC-Easy and ARC-Challenge splits (Clark et al., 2018).

4.2 Overall Performance Evaluation

Table 1 presents the main evaluation results comparing Dual-lens against full-data fine-tuning, random sampling, and state-of-the-art (SOTA) data curation baselines across three pruned LLaMA variants (1B, 8B, and 13B), all under a fixed pruning ratio of $P_r = 0.35$. Across all model sizes and tasks, Dual-lens consistently achieves the best performance in both perplexity (Wikitext) and average reasoning accuracy. Compared to full-data tuning with Alpaca, it reduces perplexity by *25.47%* and improves average reasoning accuracy by *15.80%*. Relative to the untuned pruned models, the gains are even more pronounced: a *74.25%* reduction in perplexity and a *33.62%* improvement in reasoning accuracy.

We observed that both individual components, i.e., CE-lens and SAE-lens, also outperform full-data baselines and SOTA subset selection methods. SAE-lens usually yields better performance than CE-lens, implying that semantic coverage plays a slightly more critical role than difficulty targeting. However, their combination consistently yields the strongest results, confirming that the two strategies are complementary and the Dual-lens integration is effective for model-aware subset selection.

The SOTA baselines, i.e., SelectIT, IFD, and Nuggets show only modest gains over random selection but consistently fall short of Dual-lens as these methods could be misaligned with the limited capacity and altered representation space of the pruned model. Random selection performs poorly in most settings due to its lack of alignment with

	Methods	Wikitext ↓	HellaSwag ↑	BoolQ ↑	PIQA ↑	ARC-e ↑	ARC-c ↑	Average Reasoning ↑
LLaMA 3.2 1B | P_r = 0.35	Pruned Model (w/o tuning)	141.32	30.17	55.99	62.57	40.07	19.88	41.74
	Alpaca-Full	27.99	54.87	52.11	67.24	49.83	19.69	48.75
	LaMini-Full	28.21	52.41	51.83	57.30	48.36	19.55	45.89
	GPT-J-Full	26.00	54.57	51.70	55.99	37.85	21.33	44.29
	Random Selection	97.21	50.94	53.18	59.24	46.90	19.83	46.02
	SelectIT	24.00	53.95	55.20	65.51	50.27	21.08	49.20
	IFD	24.47	54.17	50.18	65.50	50.04	22.35	48.45
	Nuggets	27.22	44.66	52.23	66.43	49.11	23.94	47.27
	CE-lens	25.90	54.09	57.09	65.23	49.78	26.02	50.44
	SAE-lens	24.20	55.01	57.03	67.63	50.23	26.27	51.23
	Dual-lens	**23.14**	**56.09**	**57.23**	**68.91**	**52.11**	**27.19**	**52.30**
LLaMA 3.1 8B | P_r = 0.35	Pruned Model (w/o tuning)	56.79	35.59	56.76	66.10	41.79	25.17	45.08
	Alpaca-Full	29.48	62.98	63.33	71.22	58.92	28.67	57.02
	LaMini-Full	27.33	58.09	64.82	63.92	49.25	20.82	51.38
	GPT-J-Full	35.18	57.11	65.22	54.07	49.18	20.37	49.19
	Random Selection	41.09	51.39	62.74	49.05	40.29	18.20	44.33
	SelectIT	26.14	62.43	63.75	71.54	59.34	30.88	57.59
	IFD	27.94	63.08	62.57	70.62	59.51	28.75	56.91
	Nuggets	24.26	64.35	63.61	71.60	60.73	30.38	58.13
	CE-lens	29.41	63.18	64.25	74.53	63.24	41.08	61.25
	SAE-lens	25.01	62.97	64.34	74.00	67.77	42.68	62.35
	Dual-lens	**23.79**	**65.69**	**65.37**	**74.57**	**68.23**	**43.99**	**63.57**
LLaMA 2.1 13B | P_r = 0.35	Pruned Model (w/o tuning)	100.87	44.95	61.67	61.11	47.18	31.31	49.24
	Alpaca-Full	31.84	55.88	51.72	68.29	67.94	27.17	54.20
	LaMini-Full	32.96	50.96	47.07	62.93	61.19	26.16	49.66
	GPT-J-Full	32.64	51.07	50.17	67.29	67.91	28.14	52.92
	Random Selection	35.17	50.17	48.11	63.27	62.69	33.20	51.49
	SelectIT	30.43	50.33	49.87	65.52	57.06	32.44	51.04
	IFD	34.21	52.34	49.23	62.71	67.31	31.75	52.66
	Nuggets	33.71	49.97	54.31	64.13	68.66	31.01	53.62
	CE-lens	20.50	68.15	67.30	75.24	72.75	42.65	65.21
	SAE-lens	20.39	68.32	66.75	75.84	73.98	43.34	65.65
	Dual-lens	**19.17**	**68.75**	**67.63**	**75.94**	**74.26**	**44.73**	**66.26**

Table 1: Comparative evaluation of various dataset curation methods on pruned LLaMA models ($P_r = 0.35$) using LLM-Pruner for pruning. The Alpaca dataset served as the base for generating curated datasets for SAE-lens, CE-lens, Dual-lens, and other compared data selection techniques.

the model's post-pruning state. The one exception is the 13B model, where random subsets relatively perform well, likely due to higher representational capacity or favorable hyperparameter interactions. However, we note that the reported performance reflects the best among multiple random trials; variability is high, making it unreliable in practice.

Notably, Dual-lens uses only 10% of the Alpaca dataset, yet consistently outperforms full-data baselines, e.g., achieving superior performance only with 0.22% of the data compared to the number of samples in the full LaMini corpus. The findings suggest that *data quality*, not merely scale, is a more influential factor in restoring performance in pruned models. It supports our core hypothesis that a compact, model-aware subset can enable more effective recovery than full-corpus fine-tuning.

4.3 Data Efficiency Across Sampling Budgets

Figure 3 presents the relationship between the proportion of the dataset retained (D_{pr}) and average reasoning accuracy for three selection methods. We observe that Dual-lens achieves strong performance with remarkably little data: even at $D_{pr} = 0.1$ (10% of the data), it matches or exceeds the accuracy of full-data tuning in other methods. Accuracy continues to improve moderately up to $D_{pr} = 0.3$, but additional data yields diminishing returns. Beyond $D_{pr} = 0.5$, all methods converge in performance, suggesting that the benefits of targeted selection diminish. These results show that dual-lens can deliver near-peak performance with only a small fraction of the training data.

4.4 Robustness Across Pruning Schemes

Table 2 evaluates Dual-lens and other data selection methods under two pruning strategies: LLM-Pruner (Ma et al., 2023), a structured approach, and FLAP (An et al., 2023), a training-free pruning method. Despite producing different pruned architectures, Dual-lens consistently achieves the best performance in both perplexity and reasoning accuracy across both pruning schemes. These results demonstrate that Dual-lens adapts effectively to

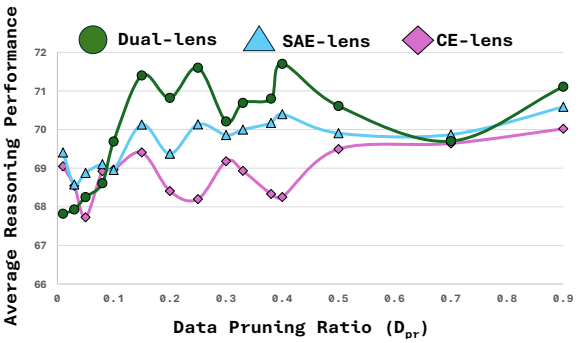

Figure 3: Effect of D_{pr} on model reasoning accuracy

Pruning Scheme: LLM-Pruner (P_r=0.15)		
Dataset Curation	**Wikitext (↓)**	**Reasoning (↑)**
Full Data	16.01	66.70
Random	16.41	65.81
Nuggets	16.61	66.84
IFD	15.81	65.44
SelectIT	16.01	66.64
SAE	15.19	67.34
CE	16.46	65.19
Dual-lens	**14.63**	**70.19**
Pruning Scheme: FLAP (P_r=0.15)		
Full Data	13.95	66.10
Random	14.16	68.02
Nuggets	13.06	68.19
IFD	13.26	67.79
SelectIT	13.46	68.11
SAE-lens	13.19	69.22
CE-lens	13.99	66.29
Dual-lens	**12.42**	**71.81**

Table 2: Evaluation of different data selection methods with LLaMA 3.1-8B under two pruning schemes (LLM-Pruner (Ma et al., 2023) and FLAP (An et al., 2023)). Metrics are reported on Wikitext (perplexity, lower is better) and Reasoning tasks (accuracy, higher is better). All subsets use 3k examples.

different pruned model states by leveraging model-aware signals to guide data selection. In contrast, full-data fine-tuning and other baselines show variable performance across pruning settings, indicating less robustness to architectural differences.

4.5 Impact of Source Dataset Scale

Table 3 compares Dual-lens and its components when constructed using 1k-sample subsets from three datasets with varying original sizes: LaMini (large), Alpaca (medium), and Dolly (small). We observe a consistent performance trend: subsets drawn from larger source datasets lead to stronger results across all metrics. Specifically, Dual-lens achieves the highest perrformance when selecting from LaMini, followed by Alpaca and then Dolly. This pattern holds for both SAE-lens and CE-lens components as well. The results implies that model-aware selection benefits from greater sample diversity, enabling more effective coverage and correction of the pruned model's deficiencies.

4.6 Generalization to Mathematical Tasks

We evaluate whether Dual-lens extends effectively to more domain-specific tasks by applying it to mathematical reasoning benchmarks: Minerva and GSM8K. Table 4 shows results for two pruned models (LLaMA 3.1 8B and LLaMA 2.1 13B) trained on subsets selected using different methods, all with a pruning ratio of $P_r = 0.25$. Dual-lens consistently outperforms all baselines, including its individual components and SOTA data selection methods. On both models and benchmarks, it achieves the highest accuracy, improving over the untuned pruned models by **22.9**× (8B) and **20.4**× (13B), and even surpassing full-data tuning. These findings demonstrate that Dual-lens is not only effective for general language understanding tasks, but also robust to domain shifts.

4.7 Effect of Pruning Ratio on Reasoning Performance

Figure 4 illustrates the impact of the pruning ratio (P_r) on reasoning improvement for LLaMA 3.1 8B. As expected, CE-lens performance declines steadily as P_r increases, since heavily pruned models lack the capacity to benefit from difficulty-targeted supervision alone. In contrast, SAE-lens exhibits a non-monotonic trend, e.g., its performance peaks near $P_r = 0.55$, suggesting that semantic coverage becomes increasingly important as model capacity diminishes. Dual-lens consistently achieves the highest improvement at moderate pruning levels ($P_r \approx 0.5$), where both predictive correction and representational alignment are valuable. However, at more extreme pruning levels ($P_r > 0.55$), SAE-lens slightly outperforms Dual-lens, likely because the CE-based criterion becomes less effective in identifying suitable samples for severely compressed models. This reflects a sensitivity to the sampling ratio applied at each stage of the pipeline; dynamically adapting this ratio based on pruning severity is a promising direction for future work.

5 Limitations

While Dual-lens demonstrates strong performance across pruning levels, it currently employs a fixed sampling ratio between the SAE-lens and CE-lens

	Dataset	Wikitext ↓	HellaSwag ↑	BoolQ ↑	PIQA ↑	ARC-e ↑	ARC-c ↑	Average Reasoning ↑
LLaMA 3.1-8B	**SAE-lens Alpaca**	32.18	67.11	61.92	72.69	74.50	43.27	63.89
	CE-lens Alpaca	33.80	66.25	62.35	72.20	72.73	41.74	63.05
	Dual-lens	**32.01**	**67.55**	**63.38**	**72.63**	**73.92**	**42.59**	**63.90**
	SAE-lens Lamini	30.22	67.10	65.04	73.93	76.22	43.31	65.11
	CE-lens Lamini	31.22	66.30	64.99	74.99	75.85	42.57	64.94
	Dual-lens	**27.17**	**69.27**	**65.57**	**75.09**	**77.91**	**44.59**	***66.49***
	SAE-lens Dolly	36.19	60.02	58.01	70.83	71.03	39.01	59.78
	CE-lens Dolly	37.01	60.00	60.01	71.50	69.81	37.02	59.67
	Dual-lens	**35.09**	**62.39**	**63.07**	**72.09**	**72.91**	**39.74**	**62.04**

Table 3: Performance evaluation of the pruned LLaMA-3.1 8B model fine-tuned on Dual-lens, SAE-lens, and CE-lens subsets derived from the Alpaca, LaMini, and Dolly datasets (1K samples from each dataset).

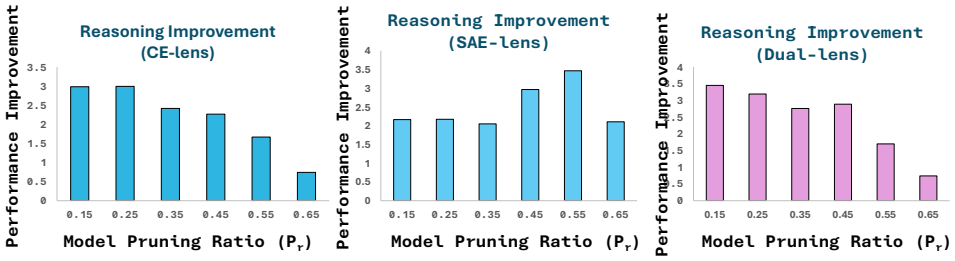

Figure 4: Change of model's reasoning accuracy improvement with varying P_r

	Methods	Minerva	GSM8K
LLaMA 3.1 8B	PM(w/o tuning)	4.29	1.00
	Full Dataset	13.42	35.29
	Random Selection	12.79	36.99
	SelectIT	15.89	40.67
	IFD	14.92	37.99
	Nuggets	14.99	39.12
	SAE-lens (MATH)	16.45	41.01
	CE-lens (MATH)	16.21	40.88
	Dual-lens (MATH)	**16.52**	**42.00**
LLaMA 2 13B	PM(w/o tuning)	6.12	1.10
	Full Dataset	17.01	42.88
	Random Selection	17.22	41.85
	SelectIT	18.95	42.44
	IFD	17.02	35.21
	Nuggets	18.73	41.00
	SAE-lens (MATH)	19.78	43.11
	CE-lens (MATH)	19.10	42.69
	Dual-lens (MATH)	**21.26**	**44.01**

Table 4: Performance evaluation of various methods, including our approach, on the Minerva and GSM8K benchmarks. Models were pruned using LLM-Pruner with a pruning ratio (P_r) of 0.25. The Camel-AI Math dataset was utilized for training the models.

stages. This static composition may not be optimal under all compression scenarios. In particular, when models are aggressively pruned, CE-based difficulty signals become less reliable due to reduced capacity, suggesting that dynamically adapting the lens weighting based on pruning severity could further improve robustness and efficiency.

Additionally, the effectiveness of SAE-lens relies on the assumption that the pruned model retains a coherent latent space from which meaningful concept representations can be extracted. While our results show that this generally holds across moderate pruning regimes, extremely compressed models may exhibit degraded internal activations, potentially limiting the representational fidelity required for effective distributional alignment. Exploring strategies to enhance or regularize latent structure in such cases remains an open direction.

6 Conclusion

We introduced *Dual-lens*, a model-aware data curation framework for efficient post-pruning recovery of language models. By combining CE-lens and SAE-lens, targeting predictive weaknesses and preserving latent semantic coverage, Dual-lens constructs compact subsets tailored to the residual capacity of pruned models. Extensive experiments show that Dual-lens consistently outperforms full-data fine-tuning and state-of-the-art selection methods, even when using only a fraction of the data. These findings support our hypothesis that a compact, model-guided subset can enable more effective recovery than conventional fine-tuning.

References

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. 2023. Fluctuation-based adaptive structured pruning for large language models. *Preprint*, arXiv:2312.11983.

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all.

Chaitali Bhattacharyya and Yeseong Kim. 2025. Finescope : Precision pruning for domain-specialized large language models using sae-guided self-data cultivation. *Preprint*, arXiv:2505.00624.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. 2023. Instruction mining: Instruction data selection for tuning large language models. *arXiv preprint arXiv:2307.06290*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and 1 others. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *Preprint*, arXiv:1905.10044.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

D.T. Davis and J.-N. Hwang. 1992. Attentional focus training by boundary region data selection. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, volume 1, pages 676–681 vol.1.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Hanyu Hu, Pengxiang Zhao, Ping Li, Yi Zheng, Zhefeng Wang, and Xiaoming Yuan. 2025. Fasp: Fast and accurate structured pruning of large language models. *arXiv preprint arXiv:2501.09412*.

Adam Karvonen, Can Rager, Samuel Marks, and Neel Nanda. 2024. Evaluating sparse autoencoders on targeted concept erasure tasks. *arXiv preprint arXiv:2411.18895*.

Yuval Kirstain, Patrick Lewis, Sebastian Riedel, and Omer Levy. 2021. A few more examples may be worth billions of parameters. *arXiv preprint arXiv:2110.04374*.

David D Lewis. 1995. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023a. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Ling-Hao Chen, Junhao Liu, Tongliang Liu, and 1 others. 2023b. One-shot learning as instruction data prospector for large language models. *arXiv preprint arXiv:2312.10302*.

Gui Ling, Ziyang Wang, and Qingwen Liu. 2024. Slimgpt: Layer-wise structured pruning for large language models. *Advances in Neural Information Processing Systems*, 37:107112–107137.

Liangxin Liu, Xuebo Liu, Derek F. Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024a. Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection. In *Advances in Neural Information Processing Systems*, volume 37, pages 97800–97825. Curran Associates, Inc.

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

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*.

OpenAI. 2023. Chatgpt. Accessed: 2025-05-19.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Preprint*, arXiv:2204.05149.

Fabrizio Sandri, Elia Cunegatti, and Giovanni Iacca. 2025. 2ssp: A two-stage framework for structured pruning of llms. *arXiv preprint arXiv:2501.17771*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Peiqi Wang, Yikang Shen, Zhen Guo, Matthew Stallone, Yoon Kim, Polina Golland, and Rameswar Panda. 2024. Diversity measurement and subset selection for instruction tuning datasets. *arXiv preprint arXiv:2402.02318*.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *arXiv preprint arXiv:2304.14402*.

Simon Yu, Liangyu Chen, Sara Ahmadian, and Marzieh Fadaee. 2024. Diversify and conquer: Diversity-centric data selection with iterative refinement. *arXiv preprint arXiv:2409.11378*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

A Appendix

A.1 Visualization of SAE-lens Selection

Figure 5 provides a qualitative comparison between samples selected by SAE-lens and those chosen via random sampling. The visualization indicates that SAE-lens selects a more diverse and semantically clustered set of samples, while random selection yields broader, less coherent distributions. This structural difference explains the performance gap observed between the two methods.

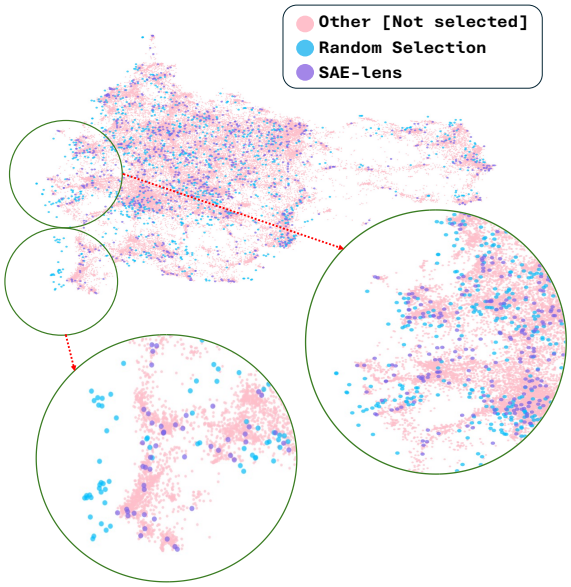

Figure 5: UMAP visualization of instruction embeddings from the Alpaca dataset. Pink points represent datapoints not selected by any method, blue points indicate randomly selected datapoints, and violet points denote datapoints selected using the SAE-lens.

A.2 Impact of Data Sampling Ratio

Figure 6 plots average reasoning performance against the data sampling ratio (D_{pr}) for three selection methods. Dual-lens achieves the highest performance across most data budgets. However, at very high sampling ratios (approaching 0.9), performance differences among methods converge, suggesting that when nearly all data is used, the impact of selection strategies diminishes.

A.3 Sample Overlap Across Pruning Ratios

Figure 7 shows the overlap among the top 1K samples selected by CE-lens at varying pruning ratios (P_r). As expected, sample overlap is higher between closer pruning levels (e.g., 0.15 vs. 0.25) than between more divergent ones (e.g., 0.15 vs. 0.35). This trend suggests that pruned models with

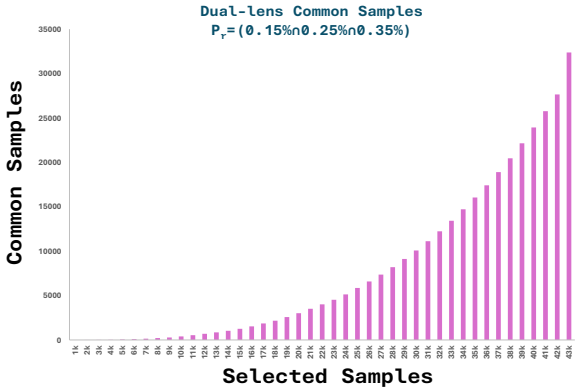

Figure 6: Change in number of common samples with varying number of selected samples.

similar capacity retain similar sensitivity patterns, resulting in consistent loss-based rankings.

A.4 Comparison with LIMA and ALPAGASUS

Table 5 presents a comparative analysis of Dual-lens and other SOTA models evaluated on the LIMA and Alpagasus datasets, using controlled dataset sizes (1k and 9k samples) to ensure fair comparison. Despite the small budget, Dual-lens demonstrates strong performance, indicating its effectiveness at identifying informative samples. Notably, Alpagasus was filtered using ChatGPT, while LIMA was manually curated with 1k diverse examples. These results suggest that Dual-lens can match or exceed the different curation methods through model-aware selection, even under strict size constraints.

A.5 Dual-lens with Different Model Sources

Figure 8 examines the impact of lens source on performance. Using CE-lens or SAE-lens derived from the *same* model (original or pruned) consistently yields better results than cross-model configurations. For instance, training SAE-lens on activations from LLaMA 3.2 1B provides stronger results for that model than using activations from LLaMA 3.1 8B. This suggests that model-specific characteristics are best captured when lenses are trained on the corresponding model.

	Dataset	Data size	Wikitext ↓	HellaSwag ↑	BoolQ ↑	PIQA ↑	ARC-e ↑	ARC-c ↑	Average Reasoning ↑
LLaMA 3.1-8B | P_r = 0.35	**LIMA**	1K	43.67	65.42	61.07	71.50	72.76	41.93	62.54
	Random	1K	43.44	65.12	62.44	71.98	73.10	40.49	62.63
	IFD	1K	44.03	63.64	62.22	72.36	73.57	42.67	62.89
	Nuggets	1K	42.44	66.32	61.96	72.57	72.14	42.50	63.09
	SelectIT	1K	41.09	65.71	62.79	71.87	74.39	42.33	63.42
	SAE-lens Alpaca	1K	32.18	67.11	61.92	72.69	74.50	43.27	63.89
	CE-lens Alpaca	1K	33.80	66.25	62.35	72.20	72.73	41.74	63.05
	Dual-lens Alpaca	**1K**	**32.01**	**67.55**	**63.38**	**72.63**	**73.92**	**42.59**	**63.90**
	Alpagasus	9K	36.22	65.00	62.87	72.63	76.69	43.45	64.13
	Random	9K	36.97	67.41	63.21	72.25	75.72	43.10	64.34
	IFD	9K	40.11	65.86	62.41	72.43	74.50	43.19	63.68
	Nuggets	9K	38.48	68.76	63.51	71.98	77.40	43.19	64.97
	SelectIT	9K	39.77	65.37	62.07	73.00	75.87	40.11	63.28
	SAE-lens Alpaca	9K	33.57	67.66	64.90	74.33	76.11	42.26	65.05
	CE-lens Alpaca	9K	35.41	66.32	63.83	74.31	74.71	40.97	64.03
	Dual-lens	**9k**	**32.03**	**70.92**	**67.34**	**77.35**	**77.91**	**44.97**	**67.69**

Table 5: Performance comparison of LIMA and Alpagasus datasets using an equal number of data instances corresponding to their original sizes curated using Dual-lens and other SOTA techniques. Model pruning ratio (P_r) was set to 0.35.

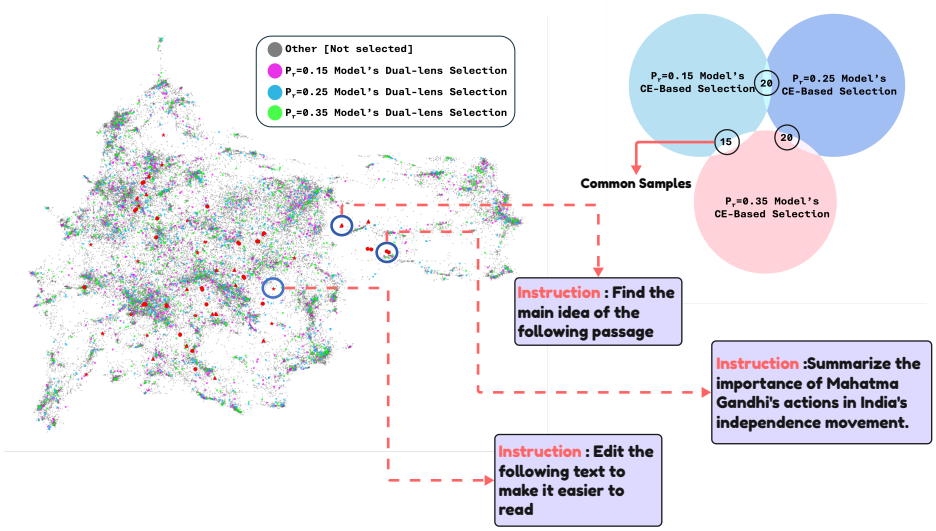

Figure 7: Visualization of common samples in different P_r setting.

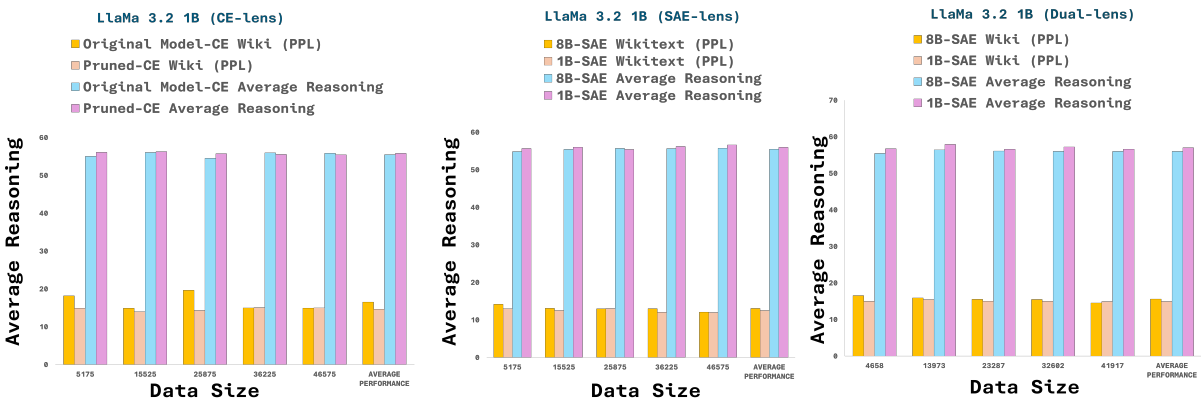

Figure 8: Effect of different setting on average reasoning accuracy