Learning What Matters: End-to-End Adaptive Calibration Data Selection for LLM Pruning

Mengting Ai[1] **Sirui Chen**[1] **Xinrui He**[1]
Maggie Zarekarizi[2] **Kommy Weldemariam**[2] **Jingrui He**[1]
[1]University of Illinois Urbana-Champaign [2]Amazon
{mai10,jingrui}@illinois.edu

Abstract

Pruning is a critical technique for compressing Large Language Models (LLMs), yet its effectiveness is highly sensitive to the small calibration set used for parameter-importance estimation. Current selection often relies on heuristics such as random sampling or entropy, leading to substantial variation in post-pruning perplexity. We analyze this instability at two levels. Calibration samples have unequal effects, and we operationally group them as **Golden**, **Mediocre**, or **Detrimental** according to their individual post-pruning outcomes. These labels are not intrinsic data properties: membership and sample interactions depend on the model and pruning algorithm. We then introduce **OASIS**, an end-to-end framework that optimizes calibration-data weights using the language-modeling loss of a differentiably soft-pruned proxy on a separate optimization corpus. The learned weights select a discrete subset that is passed to the original hard-pruning algorithm. Experiments across Llama and Qwen models show the largest benefits in calibration-sensitive structured pruning and smaller, sometimes statistically inconclusive, gains in unstructured pruning. The results position adaptive selection as a practical alternative to fixed calibration heuristics.

1 Introduction

Due to the rapidly increasing size and computational cost of Transformer-based Large Language Models (LLMs) (Vaswani et al., 2017), model compression techniques, particularly pruning, have become essential for efficient deployment and broader accessibility (Frantar and Alistarh, 2023; An et al., 2024; Ma et al., 2023; Xia et al., 2024a). However, the effectiveness of most pruning methods, both structured and unstructured, critically depends on a small set of calibration data, which is used to estimate parameter-importance saliency scores and guide pruning decisions. A growing body of

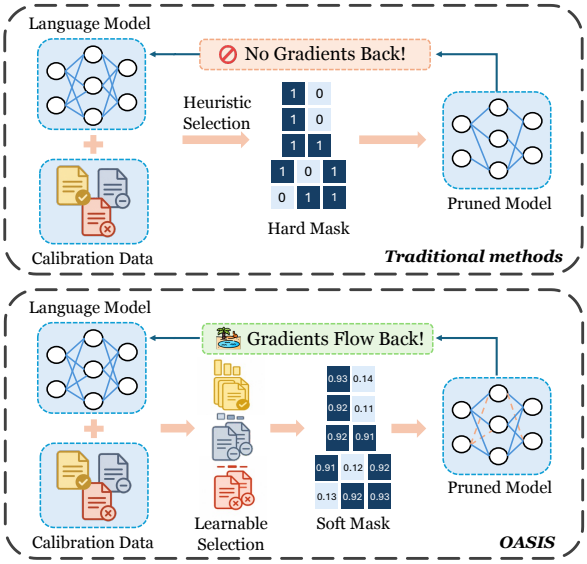


Figure 1: **Traditional vs. OASIS.** Traditional pruning blocks gradients, while **OASIS** enables feedback via soft masks and learnable calibration importance.

evidence shows that the performance of the final pruned model is highly sensitive to the choice of this calibration data (Ji et al., 2024; Kurz et al., 2024; Williams and Aletras, 2024), with factors such as data quality and diversity playing an outsized role (Bandari et al., 2024).

Despite its critical importance, the selection of calibration data is often treated as an afterthought. Common practice relies on simple heuristics, such as randomly sampling a few dozen examples from a large web corpus like C4 or Wikipedia (Sun et al., 2023; Frantar and Alistarh, 2023; Bandari et al., 2024). In this work, we first conduct a fine-grained analysis of why such heuristic approaches can be suboptimal. For each model–pruner pair, we operationally group calibration samples as "golden," "mediocre," or "detrimental" according to their individual post-pruning perplexity. In the evaluated settings, combining higher-ranked samples can be synergistic, whereas adding a lower-ranked sample can degrade the result; this poisoning effect is most pronounced for structured pruning. These labels describe observed outcomes rather than an

independent semantic taxonomy: the utility of a given sample is not a universal property. The set of golden samples varies with the specific pruning method and model being used. This dependency means that a simple, static approach to "mine" universally effective data is unlikely to generalize. It motivates a dynamic mechanism for selecting a calibration subset for each pruning scenario.

The goal, therefore, is to connect the choice of calibration data to the behavior of a pruned model. This connection is broken by the non-differentiable hard mask used in pruning, which prevents end-to-end gradient flow (Figure 1, top). To overcome this barrier, OASIS reframes selection as a learnable optimization task. By leveraging a differentiable soft-mask proxy, it propagates gradients from language-modeling loss on a separate proxy corpus to the calibration weights. The resulting ranking favors samples that improve the model–pruner-specific proxy objective and suppresses harmful ones before the original hard-pruning procedure is run.

In this work, we introduce **OASIS**, an **O**ptimized **A**pproach to **S**ystematic cal**I**bration data **S**election. Our primary contributions are:

- A thorough fine-grained analysis that systematically demonstrates the synergistic and degradative effects of individual calibration data points, exposing the critical flaws in common heuristic-based selection methods.
- A gradient-based calibration-selection framework that uses a soft-pruning proxy to optimize data weights, moving beyond entropy and random-sampling heuristics while retaining the original hard pruner for deployment.
- Comprehensive experiments showing that OASIS improves calibration robustness across multiple structured and unstructured pruning algorithms, providing a new adaptive alternative to fixed calibration-selection heuristics.

2 Related Work

LLM Pruning. Pruning aims to reduce LLM size and accelerate inference by removing redundant parameters, and can be categorized into three classes. *Unstructured pruning* (e.g., SparseGPT (Frantar and Alistarh, 2023), Wanda (Sun et al., 2023)) eliminates individual weights but yields irregular patterns with limited hardware benefit. *Semi-structured pruning* enforces patterns such as 2:4 sparsity (Fang et al., 2024; Zheng et al., 2024), offering hardware friendliness but requiring specialized support. *Structured pruning*, the most practical in deployment, removes entire neurons, heads, or layers for direct speedups. Early works (LLM-Pruner (Ma et al., 2023), Sheared Llama (Xia et al., 2024a), SlimGPT (Ling et al., 2024)) mainly used local saliency, while recent methods (Adapt-Pruner (Wang et al., 2025), FLAP (An et al., 2024), ShortGPT (Men et al., 2024), NIRVANA (Ai et al., 2025)) adopt global, layer-wise strategies. Other innovations include PCA-based designs (SliceGPT (Ashkboos et al., 2024), Olica (He and Lin, 2025)). Despite such progress, *all approaches remain highly sensitive to calibration data*, making data selection important for pruning effectiveness.

Calibration Data in Pruning. Most post-training pruning methods rely on a small set of *calibration data* to estimate activation or gradient statistics for importance scoring. A growing body of work demonstrates that pruning is highly sensitive to the properties of this data, with factors like quality, diversity significantly influencing which components are preserved and the final model performance (Williams and Aletras, 2024; JAISWAL et al., 2024; Bandari et al., 2024). To address this sensitivity, several research directions have emerged. One approach generates calibration data with the language model itself, aiming to reflect its internal knowledge distribution (Ji et al., 2024). This family is closest to our Synthetic baseline; OASIS instead ranks a fixed candidate pool with a differentiable model–pruner-specific objective. The mechanisms by which particular data distributions affect pruning nevertheless remain underexplored. Reflecting this challenge, specialized research has focused on curating calibration data for specific goals, such as language-specific pruning (Kurz et al., 2024; Zeng et al., 2024) and task-specific customization (He et al., 2026; Zhao et al., 2025), highlighting the need for data that aligns closely with the desired post-pruning model behavior.

Data Selection. Data selection is critical across all aspects of large language model training (Albalak et al., 2024), including pre-training (Gu et al., 2024), fine-tuning (Kang et al., 2024), instruction tuning (Xia et al., 2024b), and in-context learning (Zhang et al., 2022). Foundational data selection methods often employ a rank-and-select strategy, prioritizing the top-k data points based on various statistics such as training dynamics (Paul et al., 2021), entropy-based confidence and uncer-

tainty (Kremer et al., 2014), and marginal gain derived from submodular functions (Wei et al., 2015; Bhatt et al., 2024). Recent advancements have shifted towards model-aware selection beyond heuristics. These methods estimate the influence of data on the final model's performance by attributing a score to each training example. Prominent methods in this area include TracIn (Pruthi et al., 2020; Xia et al., 2024b), influence functions (Koh and Liang, 2017; Kwon et al., 2023), and training path unrolling (Bae et al., 2024).

3 Why Heuristics Fail: Multi-Scale Analysis of Calibration Data

In this section, we explore the nuances of calibration data selection to better understand the factors driving pruning performance. While existing heuristic approaches provide a valuable foundation, their analysis can be inconsistent under different pruning methods. For example, a zero-order method like Wanda (Sun et al., 2023) utilizes a saliency score that based on weight magnitudes and input activation norms, thus tends to favor data of high linguistic quality. In contrast, higher-order methods like SparseGPT (Frantar and Alistarh, 2023), leverage second-order Hessian information in the saliency score, and are sensitive to the complex curvature of the loss landscape. Calibration-data preferences can differ across pruning methods and model sizes, as shown in Section B.

Furthermore, these effects may differ substantially between structured and unstructured pruning methods, given the higher sensitivity in the structured pruning method (Ai et al., 2025). To investigate the source of this variance and build a more robust selection strategy, our analysis proceeds in two stages: we begin with a (1) macro-level investigation to identify which data properties (e.g., number of data points) have the most significant impact on performance, following previous work (Ji et al., 2024; Williams and Aletras, 2024; Bandari et al., 2024); we then conduct (2) micro-level analysis to uncover the interaction patterns behind these observations.

In the second stage, we define three operational tiers from each sample's individual post-pruning perplexity (PPL): **Golden** for the lowest-PPL group, **Mediocre** for the middle group, and **Detrimental** for the highest-PPL group. These outcome-based labels are specific to the evaluated model, pruner, candidate pool, and sparsity setting; they are used to study interactions and should not be interpreted as intrinsic semantic data categories.

3.1 Macro-Level Analysis: Data Source Outweighs Data Quantity

We begin by extending prior work examining general calibration properties such as quantity, sequence length, and source. While a common heuristic assumes that "more data is better," our results in Figure 2 challenge this notion. Performance often saturates after only 32 samples and can even degrade as sequence length increases. This suggests that simply increasing data volume is an ineffective strategy for improving pruning outcomes.

In contrast, our analysis reveals that the **data source** and specific random batch are the dominant performance factors, particularly for sensitive structured methods. The perplexity fluctuations observed across different sources and random seeds indicate that aggregate heuristics fail by failing to account for the immense variance in data quality. This high-level instability suggests that the true driver of pruning effectiveness resides at the individual sample level, necessitating the more fine-grained investigation that follows. This raises a more fundamental question: *what is happening at the individual sample level that causes these dramatic differences?*

3.2 Micro-Level Analysis: The Principles of Synergy and Degradation

To understand the mechanism behind the source-level effects observed previously, we conduct a fine-grained analysis of interactions between individual samples. For each model–pruner setting, we evaluate every candidate individually and assign its outcome-based tier from post-pruning perplexity. We then combine pairs to measure their joint effect (Figure 3). Because the tiers are defined by the same outcome being analyzed, they characterize interaction patterns rather than provide a causal or semantic taxonomy. Our analysis reveals two fundamental principles that govern these interactions:

Principle 1: Intra-Tier Synergy. Our first key finding is that combining multiple samples from the same quality tier often produces a synergistic effect, leading to a better outcome than using the constituent samples alone, as demonstrated in Figure 3 (top). The lines representing "Golden + Golden" and "Mediocre + Mediocre" combinations show a lower perplexity than the dashed lines, which indicate the best-performing individual sample within

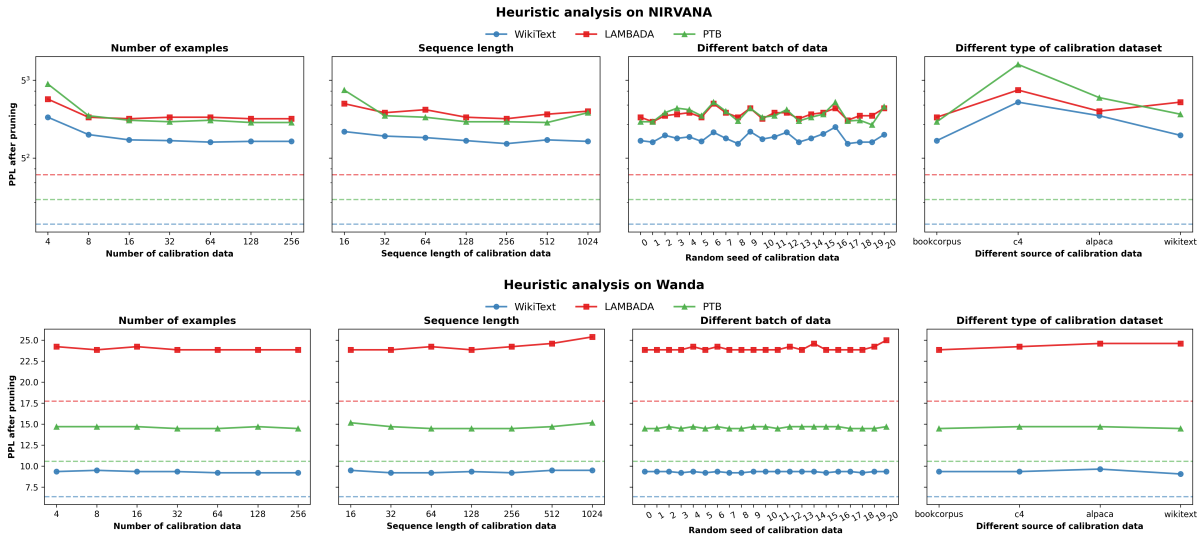


Figure 2: **Macro-level** analysis of how different heuristic choices for calibration data selection impact pruning performance. The top and bottom rows show results for the structured pruner (NIRVANA) and the unstructured pruner (Wanda), respectively. Each column investigates a different variable: (from left to right) the number of samples, sequence length, the random data batch, and the data source. The y-axis represents the post-pruning perplexity on three evaluation datasets: WikiText, PTB, and Lambada, indicated by separate colored lines. Dashed lines show the baseline perplexity of the unpruned model for comparison.

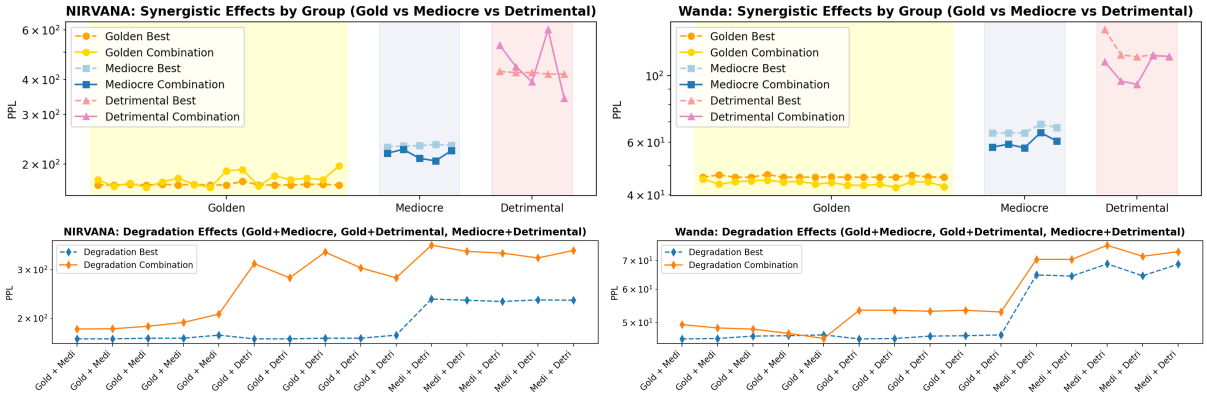


Figure 3: **Micro-level** analysis of calibration data interactions in NIRVANA (left) and Wanda (right). The top row investigates intra-tier synergy, comparing the performance of combining two samples from the same quality tier (e.g., "Golden Combination") against the performance of the better of the two individual samples ("Golden Best"). The bottom row demonstrates inter-tier degradation, showing the performance of combining samples from different tiers (e.g., "Golden + Detrimental") compared to the performance of the higher-quality sample in the pair ("Degradation Best").

each pair. This suggests that even high-quality samples contain diverse, complementary signals. Their combination creates a more robust and comprehensive calibration set, allowing the pruning algorithm to make more informed decisions. Even combining two "detrimental" samples often results in a "less detrimental" outcome, indicating that diversity can mitigate some harm from low-quality data.

Principle 2: Inter-Tier Degradation. Conversely, inter-tier combinations often degrade the stronger individual outcome in our evaluated settings. Figure 3 (bottom) illustrates this pattern. Golden+Detrimental pairs trigger a sharp perplexity spike, far exceeding the individual Golden baseline (dashed line). This degradation scales with the quality gap, as Golden+Mediocre combinations exhibit a more moderate yet distinct drop. Additionally, we observe a single-sample poisoning effect: injecting just one "detrimental" sample into an optimized set of "golden" samples can cause a disproportionate collapse in pruning quality (detailed in Section C.1).

These micro-level findings show why random sampling can be unreliable: selection must favor mutually compatible samples while avoiding candidates that distort the relevant saliency estimate. As Section B shows, the outcome-based tier membership varies with the pruning method and model

Takeaway: Why Heuristics Fail in Calibration Data Selection?

Our multi-scale analysis reveals that the failure of heuristic calibration data selection is **systemic, not incidental**: (1) simply adding more data or longer sequences yields no gain and can even degrade pruning quality; (2) calibration results are dominated by data **source**, showing that "random" heuristics are inherently unstable; (3) at the sample level, mixing operational quality tiers can trigger a **poisoning effect**, where one detrimental sample offsets much of the benefit of a selected set. Crucially, what counts as "golden" is **model- and method-dependent**, so a static ranking need not transfer across pruning settings. These observations motivate a dynamic, learnable mechanism for each model–pruner pair.

architecture. For second-order methods such as SparseGPT, some same-tier combinations can still degrade performance, suggesting a "less is more" effect where adding more calibration samples can distort curvature estimates. Rather than contradicting our motivation, this reinforces it: calibration utility is not compositional under simple heuristics, and the optimal subset must be discovered adaptively for each model-pruner pair.

4 Proposed Method

In this section, we provide the preliminary and the details of our proposed method, OASIS. An illustrative diagram is shown in Figure 4.

4.1 Preliminary and Notation

Let $\mathbf{w} = (w_1, \ldots, w_d)^\top \in \mathbb{R}^d$ denote the full set of model parameters. Consider a calibration data set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^N$ together with a weight vector

$$\mathbf{u} = (u_1, \ldots, u_N)^\top, \quad u_i \geq 0, \quad \sum_{i=1}^N u_i = 1,$$

where each u_i specifies the *importance* of sample (x_i, y_i) to the calibration procedure. The calibration loss is then defined as $L_{\text{cal}}(\mathbf{w}; \mathbf{u}) = \sum_{i=1}^N u_i \, \ell\big(f(x_i; \mathbf{w}), y_i\big)$, where $\ell(\cdot, \cdot)$ is the per-sample loss function and $f(\mathbf{x}; \mathbf{w})$ denotes the model output. For each parameter w_j, let

$$S(w_j) = \text{Saliency}\big(w_j \,;\, L_{\text{cal}}(w; u)\big), \quad j = 1, \ldots, d,$$

denote the saliency score assigned by a given pruning method, w.r.t. the calibration loss. A binary mask is obtained via top-k thresholding:

$$m_j = \begin{cases} 1, & S(w_j) \geq \text{Top}_k(S), \\ 0, & S(w_j) < \text{Top}_k(S), \end{cases}$$

The pruned parameters w'_j are then obtained by applying the hard mask $w'_j = m_j \, w_j$. Finally, given a separate proxy optimization corpus $\mathcal{D}_{\text{proxy}}$, the language-modeling loss of the soft-pruned model is

$$L_{\text{proxy}}(\mathbf{w}') = \sum_{(x,y) \in \mathcal{D}_{\text{proxy}}} \ell\big(f(x; \mathbf{w}'), \, y\big).$$

4.2 Gradient of the Task Loss w.r.t. Calibration Weights

We investigate whether this proxy loss yields a nonzero gradient with respect to the calibration weights u_i. By the chain rule,

$$\frac{\partial L_{\text{proxy}}}{\partial u_i} = \sum_{j=1}^d \frac{\partial L_{\text{proxy}}}{\partial w'_j} \underbrace{\frac{\partial w'_j}{\partial m_j}}_{=w_j} \underbrace{\frac{\partial m_j}{\partial S_j}}_{=0 \text{ (a.e.)}} \underbrace{\frac{\partial S_j}{\partial u_i}}_{\neq 0 \text{ in general}}.$$

Since $m_j = \mathbf{1}[S_j \geq \text{Top}_k(S)]$, the term $\frac{\partial m_j}{\partial S_j}$ vanishes almost everywhere. Thus $\frac{\partial L_{\text{proxy}}}{\partial u_i} = 0$. Hence, hard pruning renders the gradient with respect to $\mathbf{u}$ non-differentiable, precluding direct optimization.

4.3 OASIS: Pruning with Soft Mask

To address the non-differentiability of hard pruning, we adopt a *soft pruning* (Fang et al., 2024; Lin et al., 2024) approach, where each mask entry m_j is relaxed to lie in $[0, 1]$ rather than being restricted to $\{0, 1\}$: $m_j = \sigma(\alpha_t(S_j - \text{Top}_k(S)))$. We linearly anneal the sharpness from 50 to 10,000 over T optimization steps, $\alpha_t = 50 + (10{,}000 - 50)t/T$. This schedule supplies usable gradients early and a near-binary surrogate late in optimization, except at threshold ties. We detach the top-k selector from the computational graph; thus, gradients are propagated through the saliency scores $S(u)$ and the soft mask, while the rank threshold is treated as a fixed cut-off within each optimization step. Under this relaxation, the gradient of the proxy loss L_{proxy} w.r.t. the calibration weights $\mathbf{u}$ becomes

$$\frac{\partial L_{\text{proxy}}}{\partial u_i} = \sum_{j=1}^d \frac{\partial L_{\text{proxy}}}{\partial w'_j} \underbrace{\frac{\partial w'_j}{\partial m_j}}_{=w_j} \underbrace{\frac{\partial m_j}{\partial S_j}}_{\neq 0 \text{ (soft)}} \underbrace{\frac{\partial S_j}{\partial u_i}}_{\neq 0 \text{ in general}},$$

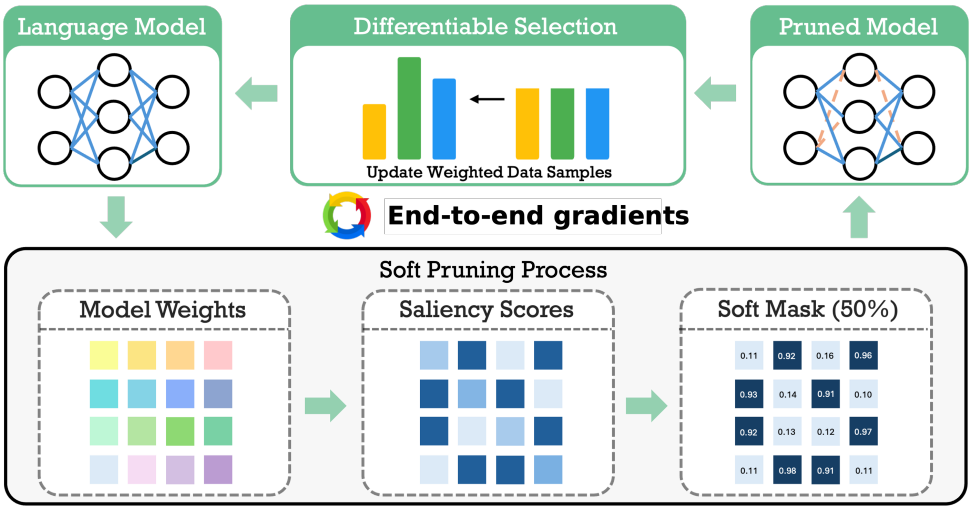


Figure 4: An overview of the OASIS framework for learning calibration data importance weights. OASIS reframes data selection as an end-to-end differentiable optimization loop.

which is nonzero and therefore permits gradient-based optimization of $\mathbf{u}$.

Stabilization via Perturbation. In practice, we observed that direct optimization of $\mathbf{u}$ using the proxy loss can be highly sensitive (Bishop, 1995; Park et al., 2022), often resulting in poor sample selection (see Section 5.3). To mitigate this instability, we introduce a perturbation at the input embedding level (Jiang et al., 2020; Zhu et al., 2020).

Let $\phi(x)$ denote the embedding of an input x, and let $\epsilon \in \mathbb{R}^{\dim(\phi(x))}$ be a random perturbation. We construct perturbed embeddings $\tilde{\phi}(x) = \phi(x) + \epsilon$, where ϵ is drawn from a uniform distribution following previous work (Aghajanyan et al., 2021). Abusing the notation f, the perturbed embeddings $\tilde{\phi}(x)$ are then passed through the model to compute the proxy loss L:

$$L = L_{\text{pert}}(\mathbf{w}', \mathbf{u}) = \sum_{(x,y)\in\mathcal{D}_{\text{task}}} \ell\big(f(\tilde{\phi}(x); \mathbf{w}'),\, y\big),$$

whose gradient with respect to u is used for updating the calibration weights. Mathematically, injecting noise during optimization is equivalent to *Tikhonov regularization* (Bishop, 1995). By Taylor expansion, the expected loss satisfies:

$$\mathbb{E}_\epsilon[L(f(\phi(x)+\epsilon))] \approx L(f(\phi(x))) + \frac{1}{2}\sigma^2 \mathrm{Tr}(\nabla^2_\phi L).$$

This implicit regularization smooths the proxy loss landscape (Jiang et al., 2020). Empirically, it stabilizes the learned ranking and improves its transfer to the post-pruning hard-mask regime (Section 5.3). The perturbed loss L is estimated on minibatches from the proxy optimization corpus.

Soft-to-hard connection. The soft mask is used only as a differentiable surrogate during calibration-data optimization. With $m_\alpha = \sigma(\alpha(S - \tau))$, increasing α makes the relaxed mask approach the corresponding hard threshold mask except at threshold ties. OASIS therefore follows a soft-training/hard-application protocol: it learns calibration weights through the relaxed proxy and converts them into a discrete top-k calibration set. Specifically, we rank the candidate samples by u_i and select the top-k samples. The selected subset is then passed to the corresponding pruning algorithm.

5 Experiment

We begin with the experimental setup, followed by evaluation results on both structured and unstructured pruning methods, and conclude with an ablation study. Additional experiments on Llama3.2-3B and 1B are provided in Section A.3, and the time cost is provided in Section A.4.

5.1 Experimental Setup

We evaluate OASIS across diverse model scales and architectures, including the **Llama3** family (Llama-3.1-8B, Llama-3.2-1B/3B) and **Qwen2.5-7B**. Our framework is applied to both **structured pruning** (NIRVANA (Ai et al., 2025), LLM-Pruner (Ma et al., 2023)) and **unstructured pruning** (Wanda (Sun et al., 2023), SparseGPT (Frantar and Alistarh, 2023)) at a target sparsity of 50%. We compare OASIS against random sampling, **Entropy-based** selection (Kremer et al., 2014), and **Synthetic** data generation (Ji et al., 2024); and TracIn (Pruthi et al., 2020) for influence-based ranking. Performance is assessed via zero-shot perplex-

Table 1: Zero-shot performance of Llama3.1-8B and Qwen2.5-7B after applying 50% structured/unstructured pruning. We compare baseline pruning methods against different calibration data selection strategies: entropy-based, synthetic data, and our proposed OASIS. Results are averaged through three runs. **Bold** and underline denote the best and second-best results per group, respectively. ↓: lower is better.

Method	PTB ↓	LambD ↓	ARC-e	WinoG	HellaS	BoolQ	PIQA	ARC-c	OBQA	COPA	Avg Acc.
Llama3.1-8B											
NIRVANA	82.48±11.61	106.92±13.96	32.97±1.10	51.78±1.48	36.91±0.79	61.52±0.48	61.37±1.22	24.23±0.84	28.20±0.35	62.00±2.00	44.87±0.58
+ Entropy	65.88±2.06	**73.91±3.50**	35.97±0.29	53.04±1.85	38.64±0.91	62.00±0.18	62.22±0.35	24.77±0.81	28.27±0.83	**64.67±1.15**	46.20±0.40
+ Synthetic	70.80±11.99	84.76±7.14	**37.77±0.57**	54.64±2.40	39.36±0.96	57.22±8.48	62.01±1.12	26.08±1.16	28.60±0.53	64.00±6.24	46.21±1.14
+ OASIS (ours)	**58.25±4.91**	80.49±7.06	37.60±0.86	**55.51±2.53**	**40.28±1.17**	**62.19±0.08**	**62.44±0.36**	**26.34±0.99**	**29.20±1.25**	64.67±3.06	**47.28±0.87**
LLM-Pruner	595.16± 61.23	353.06± 27.53	**28.54± 0.40**	50.22± 1.03	27.07± 0.39	37.94± 0.25	52.10± 0.33	23.52± 0.40	26.07± 0.81	58.33± 4.16	37.97± 0.74
+ Entropy	566.55± 128.49	**277.55± 15.25**	28.38± 0.77	50.62± 0.77	27.14± 0.28	38.13± 0.28	53.03± 0.35	22.78± 0.85	**28.33± 0.50**	57.33± 1.53	38.22± 0.28
+ Synthetic	671.06± 106.09	283.51± 18.69	28.20± 1.05	50.96± 0.32	26.62± 0.05	38.61± 1.33	52.65± 0.60	24.03± 0.21	27.00± 0.69	57.33± 2.31	38.18± 0.60
+ OASIS (ours)	**531.03± 57.74**	321.87± 31.45	28.06± 0.35	**51.12± 0.30**	**27.67± 0.20**	**39.63± 1.46**	**53.19± 0.03**	**24.09± 0.85**	27.47± 0.31	**61.33± 1.53**	**39.48± 0.24**
WANDA	14.69± 0.00	25.00± 0.00	**68.22± 0.55**	69.90± 0.46	69.08± 0.10	76.86± 0.41	76.51± 0.21	43.97± 0.26	40.53± 0.23	86.17± 0.52	66.37± 0.04
+ Entropy	14.54± 0.13	24.67± 0.00	67.63± 0.76	69.48± 0.48	69.00± 0.18	77.81± 0.49	76.73± 0.23	43.57± 0.49	40.13± 0.95	85.33± 0.58	66.21± 0.35
+ Synthetic	**14.47± 0.00**	25.13± 0.23	67.49± 0.46	69.77± 0.49	69.21± 0.25	78.00± 0.16	**77.11± 0.32**	43.20± 0.10	39.47± 1.17	87.13± 0.9	66.29± 0.19
+ OASIS (ours)	14.69± 0.00	**24.61± 0.00**	68.03± 0.91	**70.19± 0.41**	**69.22± 0.06**	**78.56± 0.36**	76.35± 0.08	**44.08± 0.38**	**40.67± 0.50**	**87.33± 0.58**	**66.68± 0.03**
SparseGPT	18.39± 0.33	31.93± 0.29	60.49± 0.20	66.14± 1.11	63.23± 0.64	71.65± 3.52	72.87± 0.91	35.86± 1.39	36.40± 1.11	81.67± 1.15	61.04± 0.82
+ Entropy	18.31± 0.57	29.84± 0.27	60.58± 0.74	**66.64± 0.91**	63.80± 0.60	**72.74± 1.88**	72.47± 0.66	35.86± 0.34	36.27± 0.83	81.33± 1.53	61.21± 0.10
+ Synthetic	18.97± 0.45	30.79± 0.28	**60.61± 3.09**	65.82± 0.65	63.36± 0.57	71.89± 1.34	73.16± 1.02	36.35± 0.81	35.60± 0.40	81.33± 3.21	61.02± 0.14
+ OASIS (ours)	**18.27± 0.62**	**29.44± 0.75**	60.30± 0.71	65.09± 1.00	**64.04± 0.07**	70.89± 1.01	**73.61± 0.05**	**37.68± 0.30**	**36.67± 1.86**	**84.00± 3.46**	**61.54± 0.32**
Qwen2.5-7B											
NIRVANA	94.39± 12.59	145.65± 11.53	36.11± 1.64	50.70± 0.73	35.72± 0.52	61.19± 1.46	58.76± 0.25	24.74± 0.56	**29.40± 0.72**	59.33± 2.31	44.49± 0.59
+ Entropy	77.22± 7.23	84.59± 2.64	37.29± 0.99	53.17± 0.09	38.12± 0.67	56.59± 2.96	59.96± 0.74	23.75± 0.38	27.73± 0.58	64.00± 1.00	45.08± 0.38
+ Synthetic	88.17± 1.60	112.06± 3.50	**43.88± 1.45**	52.25± 1.99	36.74± 0.55	60.75± 3.26	58.76± 0.63	**27.79± 1.15**	28.33± 0.31	61.00± 3.00	46.19± 0.77
+ OASIS (ours)	**60.60± 1.10**	**79.62± 6.44**	37.53± 0.80	**54.22± 0.44**	**38.39± 0.62**	**61.43± 0.54**	**60.01± 0.27**	25.60± 1.07	28.07± 0.95	**66.67± 3.51**	**46.49± 0.37**
LLM-Pruner†	323.29± 262.26	368.64± 179.37	38.51± 1.76	50.46± 1.83	33.07± 2.31	**53.40± 4.47**	57.85± 1.44	25.00± 1.01	27.20± 1.44	64.67± 5.51	43.77± 1.61
+ Entropy	248.19± 106.13	316.53± 150.38	40.54± 1.38	50.22± 2.47	35.75± 2.78	51.19± 6.30	58.20± 2.10	26.48± 0.92	28.47± 1.50	59.00± 2.00	43.73± 0.91
+ Synthetic	548.19± 480.26	473.16± 296.79	43.21± 3.08	49.93± 0.20	31.42± 0.88	49.40± 6.79	56.67± 0.54	27.82± 3.66	27.73± 0.81	56.33± 2.89	42.82± 0.79
+ OASIS (ours)	**115.31± 44.78**	**178.84± 91.82**	**47.15± 3.19**	**51.88± 1.57**	**40.52± 5.77**	52.70± 4.68	**62.44± 1.55**	**29.32± 2.10**	**29.60± 2.03**	**66.67± 1.53**	**47.54± 1.39**
WANDA	15.40± 0.00	24.23± 0.00	74.14± 2.02	69.51± 0.60	69.25± 0.02	82.46± 0.48	76.55± 0.14	46.10± 1.07	41.73± 0.31	83.33± 3.21	67.88± 0.36
+ Entropy	15.40± 0.00	**23.61± 0.21**	73.81± 2.35	68.82± 0.36	**69.79± 0.13**	82.46± 0.97	**77.58± 0.05**	45.90± 2.29	41.80± 0.35	83.33± 0.58	67.94± 0.71
+ Synthetic	15.64± 0.00	24.48± 0.22	73.72± 0.46	69.82± 0.57	69.25± 0.35	**82.59± 0.81**	76.08± 0.52	45.82± 0.70	40.93± 0.81	83.00± 1.00	67.65± 0.36
+ OASIS (ours)	**15.32± 0.14**	24.10± 0.22	**74.55± 0.54**	**70.27± 0.54**	69.52± 0.14	82.39± 0.49	76.37± 0.31	**46.70± 1.07**	**41.87± 0.81**	**84.00± 1.00**	**68.21± 0.10**

†LLM-Pruner on Qwen2.5-7B uses 40% sparsity

ity and accuracy on eight reasoning benchmarks using `lm-eval-harness` (Gao et al., 2024). The candidate and final calibration samples come from WikiText-2 (Merity et al., 2016). OASIS estimates its proxy loss on 2,048 Stanford Alpaca (Taori et al., 2023) sequences of length 128; these proxy examples are never included in the final calibration set. The evaluation benchmarks are disjoint and are not used to learn weights, tune hyperparameters, or select samples. Details are provided in Section A.1.

5.2 Main Results

Table 1 summarizes the performance of OASIS under structured and unstructured pruning. Importantly, the downstream benchmarks are not used to train or tune OASIS; therefore, the reported results reflect zero-shot post-pruning behavior rather than task-specific adaptation. We observe two distinct trends depending on pruning granularity. On Llama-3.1-8B, the comparison with TracIn preserves the same overall pattern (Table 4). A paired analysis across matched seeds gives an average-accuracy gain of $+1.26$ points for OASIS over Entropy with LLM-Pruner (95% CI $[+1.03, +1.78]$), but only $+0.31$ for Wanda (95% CI $[-0.06, +0.49]$). Thus, the structured-pruning improvement is consistent in these runs, whereas the unstructured gain is suggestive but not statistically conclusive.

Structured Pruning. Structured pruning removes entire model components, making the estimated saliency scores particularly vulnerable to biases or outliers in a small calibration set. OASIS consistently improves the overall post-pruning results: it reduces perplexity degradation in most structured settings and yields stronger average downstream performance compared with other calibration selection baselines. The comparison with the Synthetic baseline further illustrates this point: although model-generated data can appear fluent and high-confidence, it does not always provide the most informative signal for coarse-grained importance estimation. In several structured pruning settings, Synthetic selection exhibits high variance and can lead to sharp PPL increases, suggesting that fluency or model confidence alone is insufficient for selecting calibration data. By contrast, OASIS selects calibration subsets according to their effect on a soft-pruned proxy, which better aligns the selection process with the final post-

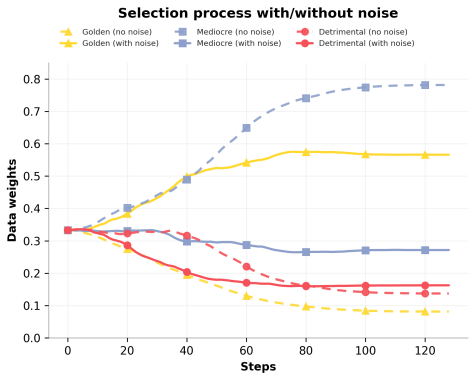


Figure 5: Evolution of importance weights during optimization. The dashed line (w/o noise) indicates overfitting to mediocre samples, while the full line (w/ noise) enables stable discovery of golden samples.

pruning objective. As a result, OASIS is less dependent on static data-quality heuristics and better preserves model behavior under aggressive structural removal.

Unstructured Pruning. For unstructured pruning, the gap between selection strategies is narrower. This is expected: weight-level unstructured pruning methods remove individual connections between neurons and are generally more stable with respect to the calibration set. In this regime, OASIS should therefore not be expected to produce the same level of improvement as in structured pruning. Still, it remains competitive with heuristic baselines and achieves modest average gains; the paired confidence interval above shows that the Wanda improvement is not conclusive with three runs. These results indicate that even when pruning is relatively stable, optimizing calibration selection can still provide a cleaner saliency signal than relying solely on uncertainty-based or generated-data heuristics.

Taken together, the results support OASIS as a principled adaptive calibration-selection method. Its main advantage is strongest in calibration-sensitive pruning regimes, where selecting the wrong data can substantially distort saliency estimation, while in more stable unstructured settings, it provides smaller but positive improvements.

5.3 Ablation Study

We ablate both embedding perturbation and optimization length. Figure 5 presents the evolution of data weights over training steps. Without noise (dashed lines), optimization is unstable and overfits to spurious artifacts, causing golden samples to be overshadowed by lower-quality data. Conversely, noise-induced smoothing (full lines) enables robust weighting: golden samples are consistently amplified while detrimental ones are suppressed. This demonstrates that embedding perturbation enables the stable amplification of golden samples while successfully suppressing lower-quality data. Quantitatively, Table 2 reports the post-pruning perplexity (PPL) with and without noise. The results with noise consistently surpass those without. For NIRVANA, reducing optimization from three epochs to one increases PTB/Lambada PPL from 58.25/80.49 to 73.69/91.72, indicating that the full optimization schedule materially contributes to selection quality.

Table 2: Embedding-noise and optimization-length ablations on Llama-3.1-8B. Lower perplexity is better.

Method (Llama-3.1-8B)	**PTB**	**LambD**
Structured Pruning: NIRVANA		
OASIS (3 epochs)	**58.25**	**80.49**
1 epoch	73.69	91.72
w/o noise	63.19	82.01
Structured Pruning: LLM-Pruner		
OASIS	**531.03**	**321.87**
w/o noise	542.01	324.72
Unstructured Pruning: Wanda		
OASIS	**14.69**	**24.61**
w/o noise	14.99	24.68

6 Conclusion

We study calibration data selection as an important factor in LLM pruning. Our analysis shows that calibration samples are not interchangeable: their effects can be synergistic or detrimental, and their utility varies across models and pruning algorithms. This context dependence explains why fixed heuristics such as random sampling, entropy ranking, or synthetic-data filtering can yield unstable pruning outcomes. To address this issue, we propose OASIS, an adaptive framework that learns per-sample importance through a differentiable soft-pruning proxy and applies the selected subset to the original hard-pruning pipeline.

Experiments across Llama and Qwen models show that OASIS improves calibration robustness in the evaluated settings. These results suggest that calibration selection should be treated as a model- and compressor-dependent optimization problem rather than a fixed preprocessing choice.

Limitations

OASIS adds a one-time optimization phase requiring forward and backward passes. On one

H100, selection takes 64–91 minutes for Llama-3.1-8B depending on the pruner, compared with 70–280 seconds for native pruning (Table 7). Whether this overhead is justified depends on the deployment scale and is clearest for the structured pruners, where the measured gains are larger.

Our candidate pool is fixed at 128 WikiText-2 sequences and the final budget at 32; we did not conduct a systematic pool-size sweep. We specify the 50-to-10,000 mask-sharpness schedule but do not sweep alternative temperature schedules. Selection is consequently bounded by the diversity and suitability of this initial pool. The tier labels are outcome-based and specific to a model, pruner, pool, sparsity level, and evaluation measure; they should not be interpreted as a general semantic characterization of text quality. Results use three seeds, so small unstructured-pruning differences remain uncertain. We evaluate English data, Llama/Qwen models, four pruners, and 50% sparsity (except the noted Qwen LLM-Pruner setting); broader languages, architectures, sparsity levels, and proxy corpora require further study. Finally, we study pruning rather than quantization or other calibration-dependent compression methods.

Ethical Considerations

Pruned models may require fewer deployment resources, but OASIS adds offline optimization compute. We did not measure energy use or carbon emissions, so we make no claim about net environmental benefit; any benefit depends on hardware, realized inference savings, and how many deployments amortize the selection cost.

We use publicly available pretrained models and established datasets and do not collect human-subject or private data. The compressed models retain biases, limitations, and potential harms inherited from their training data, and pruning or calibration selection may unevenly alter capabilities across groups or domains. OASIS is not a bias-mitigation method; fairness and safety should therefore be re-evaluated before deployment of each pruned model.

Acknowledgments

This work is supported by an Amazon Research Award. This work also used DeltaAI at NCSA through allocation CIS250202 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2021. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*.

Mengting Ai, Tianxin Wei, Sirui Chen, and Jingrui He. 2025. Nirvana: Structured pruning reimagined for large language models compression. *Preprint*, arXiv:2509.14230.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, and 1 others. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. 2024. Fluctuation-based adaptive structured pruning for large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):10865–10873.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. 2024. SliceGPT: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*.

Juhan Bae, Wu Lin, Jonathan Lorraine, and Roger B Grosse. 2024. Training data attribution via approximate unrolling. *Advances in Neural Information Processing Systems*, 37:66647–66686.

Abhinav Bandari, Lu Yin, Cheng-Yu Hsieh, Ajay Kumar Jaiswal, Tianlong Chen, Li Shen, Ranjay Krishna, and Shiwei Liu. 2024. Is c4 dataset optimal for pruning? an investigation of calibration data for llm pruning. *Preprint*, arXiv:2410.07461.

Gantavya Bhatt, Yifang Chen, Arnav M Das, Jifan Zhang, Sang T Truong, Stephen Mussmann, Yinglun Zhu, Jeffrey Bilmes, Simon S Du, Kevin Jamieson, and 1 others. 2024. An experimental design framework for label-efficient supervised finetuning of large language models. *arXiv preprint arXiv:2401.06692*.

Chris Bishop. 1995. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7:108–116.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical

commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and etc. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

An Yang et al. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Gongfan Fang, Hongxu Yin, Saurav Muralidharan, Greg Heinrich, Jeff Pool, Jan Kautz, Pavlo Molchanov, and Xinchao Wang. 2024. MaskLLM: Learnable semi-structured sparsity for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10323–10337. PMLR.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Yuxian Gu, Li Dong, Hongning Wang, Yaru Hao, Qingxiu Dong, Furu Wei, and Minlie Huang. 2024. Data selection via optimal control for language models. *arXiv preprint arXiv:2410.07064*.

Bowei He, Lihao Yin, Huiling Zhen, Shuqi LIU, Han Wu, Xiaokun Zhang, Mingxuan Yuan, and Chen Ma. 2026. Preserving LLM capabilities through calibration data curation: From analysis to optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jiujun He and Huazhen Lin. 2025. Olica: Efficient structured pruning of large language models without retraining. In *Forty-second International Conference on Machine Learning*.

AJAY KUMAR JAISWAL, Zhe Gan, Xianzhi Du, Bowen Zhang, Zhangyang Wang, and Yinfei Yang. 2024. Compressing LLMs: The truth is rarely pure and never simple. In *The Twelfth International Conference on Learning Representations*.

Yixin Ji, Yang Xiang, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, and Min Zhang. 2024. Beware of calibration data for pruning large language models. *Preprint*, arXiv:2410.17711.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.

Feiyang Kang, Hoang Anh Just, Yifan Sun, Himanshu Jahagirdar, Yuanzhi Zhang, Rongxing Du, Anit Kumar Sahu, and Ruoxi Jia. 2024. Get more for less: Principled data selection for warming up fine-tuning in llms. *arXiv preprint arXiv:2405.02774*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Jan Kremer, Kim Steenstrup Pedersen, and Christian Igel. 2014. Active learning with support vector machines. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(4):313–326.

Simon Kurz, Jian-Jia Chen, Lucie Flek, and Zhixue Zhao. 2024. Investigating language-specific calibration for pruning multilingual large language models. *Preprint*, arXiv:2408.14398.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. 2023. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. *arXiv preprint arXiv:2310.00902*.

Weihao Lin, Shengji Tang, Chong Yu, Peng Ye, and Tao Chen. 2024. S2HPruner: Soft-to-hard distillation bridges the discretization gap in pruning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

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*, volume 36, pages 21702–21720. Curran Associates, Inc.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2024. Shortgpt: Layers in large language models are more redundant than you expect. *Preprint*, arXiv:2403.03853.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Jungsoo Park, Gyuwan Kim, and Jaewoo Kang. 2022. Consistency training with virtual adversarial discrete perturbation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5646–5656, Seattle, United States. Association for Computational Linguistics.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.

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`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Boyao Wang, Rui Pan, Shizhe Diao, Xingyuan Pan, Jipeng Zhang, Renjie Pi, and Tong Zhang. 2025. Adapt-pruner: Adaptive structural pruning for efficient small language model training. *Preprint*, arXiv:2502.03460.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. 2015. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pages 1954–1963. PMLR.

Miles Williams and Nikolaos Aletras. 2024. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10100–10118, Bangkok, Thailand. Association for Computational Linguistics.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2024a. Sheared llama: Accelerating language model pre-training via structured pruning. *Preprint*, arXiv:2310.06694.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024b. Less: selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 54104–54132.

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

Hongchuan Zeng, Hongshen Xu, Lu Chen, and Kai Yu. 2024. Multilingual brain surgeon: Large language models can be compressed leaving no language behind. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11794–11812, Torino, Italia. ELRA and ICCL.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148.

Yirao Zhao, Guizhen Chen, Kenji Kawaguchi, Lidong Bing, and Wenxuan Zhang. 2025. Pruning general large language models into customized expert models. *Preprint*, arXiv:2506.02561.

Haizhong Zheng, Xiaoyan Bai, Xueshen Liu, Z. Morley Mao, Beidi Chen, Fan Lai, and Atul Prakash. 2024. Learn to be efficient: Build structured sparsity in large language models. *Preprint*, arXiv:2402.06126.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.

A More Experiments Details

A.1 Experimental Setup

Models. We conduct our experiments on the Llama3 (Dubey et al., 2024) family of models (specifically on Llama3.1-8B) and on Qwen2.5-7B (et al., 2025) to demonstrate the effectiveness of our method across different model scales. All models are evaluated in a post-training, zero-shot setting to isolate the impact of the pruning process itself. Additional experiment results on Llama3.2-3B and Llama3.2-1B can be found in the Appendix.

Pruning Baselines. To assess the versatility of OASIS, we apply our data selection method to four pruning algorithms: **Structured Pruning:** We use LLM-Pruner (Ma et al., 2023) and NIRVANA (Ai et al., 2025), two state-of-the-art methods that remove entire structural units (e.g., attention heads, FFN neurons). These methods are chosen for their practical relevance, as they can lead to direct computational speedups. **Unstructured Pruning:** We use Wanda (Sun et al., 2023), a magnitude-and-activation-based method, and SparseGPT (Frantar and Alistarh, 2023), a second-order method. Both remove individual weights and let us test fine-grained pruning techniques. We set a target sparsity of 50% for all methods. Specifically, we use per-sample calibration loss/gradient contribution weighted by softmaxed u for NIRVANA/LLM-Pruner, weighted activation statistics for Wanda, and weighted activation covariance/Hessian proxy for SparseGPT.

Calibration Data Selection Baselines. We compare OASIS against two data-selection baselines, in addition to the random sampling used by the original pruning methods: **Entropy-based Selection** (Kremer et al., 2014): A common heuristic that selects data samples with the highest token-level entropy, under the assumption that more complex and uncertain samples are more informative for calibration. **Synthetic Data Generation** (Ji et al., 2024): A recent approach where the LLM itself is prompted to generate its own calibration data. This strategy aims to create samples that are well-aligned with the model's internal knowledge distribution. For the additional Llama-3.1-8B comparison in Table 4, **TracIn** ranks WikiText-2 candidates by influence (Pruthi et al., 2020). Both TracIn and OASIS retain the same final budget of 32 sequences.

Evaluation tasks and datasets. We conduct experiments on two types of tasks. First, we follow (Ma et al., 2023) to evaluate zero-shot perplexity PTB (Marcus et al., 1993), and Lambada (Paperno et al., 2016). Second, we evaluate zero-shot accuracy on a suite of commonsense reasoning benchmarks, including ARC-easy (Clark et al., 2018), Winogrande (Sakaguchi et al., 2021), HellaSwag (Zellers et al., 2019), BoolQ (Clark et al., 2019) and PIQA (Bisk et al., 2020), ARC-challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), and COPA (Gordon et al., 2012). All evaluations use `lm-eval-harness=0.4.8` (Gao et al., 2024). Licenses are provided in Table 3.

Implementation Details. All our experiments were conducted on NVIDIA-H100-80GB GPUs. Each main selection method, including OASIS, selects a final subset from WikiText-2 (Merity et al., 2016) to guide pruning. All methods use the same final budget of 32 sequences. Random selection uniformly samples 32 sequences. Entropy-based selection ranks candidate sequences by token-level entropy and selects the top 32. For the Synthetic baseline, we follow the official implementation of Ji et al. and generate candidates from the same initial pool; we then apply the perplexity-based filtering criterion and retain the 32 selected samples. For OASIS, we uniformly initialize calibration weights over a pool of 128 WikiText-2 sequences. We optimize for three epochs with AdamW, learning rate 1×10^{-3}, weight decay 10^{-3}, and batch size 64. The sigmoid sharpness is linearly annealed from 50 to 10,000. Proxy language-modeling loss is computed on 2,048 Stanford Alpaca sequences of length 128. The proxy corpus is deliberately separate from the WikiText-2 candidate pool; proxy examples cannot enter the selected set. Evaluation datasets are used only after pruning and never for optimization or hyperparameter selection. After optimization, we rank candidate samples by their learned weights and select the top 32 samples as the final calibration set. The final pruning step uses the original hard-mask implementation of each pruning method, with only the calibration set changed.

A.2 Influence-Based Selection and Uncertainty

OASIS achieves higher mean average accuracy than TracIn for all four pruners, while metric-level differences remain method dependent. The main text reports paired confidence intervals for LLM-Pruner and Wanda; with only three matched runs, we avoid treating small unstructured gains as conclusive.

Table 3: Summary of Datasets and Their Licenses

Dataset / Benchmark	License / Copyright
Wikitext-2	CC BY-SA 4.0
Stanford Alpaca	CC BY-NC 4.0
PTB (Penn Treebank)	Non-commercial (Copyright 1995 UPenn / LDC)
LAMBADA	CC BY 4.0
ARC (Easy & Challenge)	CC BY-SA 4.0
Winogrande	Apache License 2.0
HellaSwag	Apache License 2.0
BoolQ	CC BY-SA 3.0
PIQA	AFL-3.0
OpenBookQA	Apache License 2.0
COPA	CC BY-NC-SA 3.0

Table 4: Comparison with influence-based calibration selection on Llama-3.1-8B. Entries are mean±standard deviation over three runs. TracIn ranks the WikiText-2 pool by influence.

Pruner	Selection	PTB ↓	LambD ↓	Avg Acc. ↑
NIRVANA	TracIn	60.64 ± 2.92	84.59 ± 2.64	46.74 ± 1.18
	OASIS	$\mathbf{58.25 \pm 4.91}$	$\mathbf{80.49 \pm 7.06}$	$\mathbf{47.28 \pm 0.87}$
LLM-Pruner	TracIn	$\mathbf{497.58 \pm 32.80}$	368.17 ± 31.02	38.23 ± 0.49
	OASIS	531.03 ± 57.74	$\mathbf{321.87 \pm 31.45}$	$\mathbf{39.48 \pm 0.24}$
Wanda	TracIn	$\mathbf{14.54 \pm 0.13}$	24.74 ± 0.02	66.33 ± 0.23
	OASIS	14.69 ± 0.00	$\mathbf{24.61 \pm 0.00}$	$\mathbf{66.68 \pm 0.03}$
SparseGPT	TracIn	18.67 ± 0.17	32.43 ± 0.29	60.21 ± 0.66
	OASIS	$\mathbf{18.27 \pm 0.62}$	$\mathbf{29.44 \pm 0.75}$	$\mathbf{61.54 \pm 0.32}$

A.3 Additional Results on Llama3.2-3B and Llama3.2-1B

In the main content, we provide the experiment results on Llama3.1-8B and Qwen2.5-7B. Here, we present additional results on Llama3.2-3B and Llama3.2-1B. The results presented in Table 5 and Table 6 on the Llama-3.2 3B and 1B models further validate the conclusions drawn in the main paper. These experiments demonstrate that the effectiveness of OASIS is consistent across different model sizes within the Llama family.

As shown in Table 5, the structured pruning results on the 3B model mirror the trends observed on the larger 8B model. (1) Consistent Superiority of OASIS: For both NIRVANA and LLM-Pruner, OASIS delivers the most noticeable improvements in language modeling performance, achieving the lowest (best) perplexity scores on nearly all benchmarks. This is a critical result, as perplexity is a strong indicator of a model's fundamental language understanding, which is often severely damaged by structured pruning. OASIS proves most effective at preserving this core capability. (2) Instability of Baseline Selection Methods: The limitations of heuristic and generative methods are again evident. The synthetic data approach, in particular, leads to a catastrophic degradation for both NIRVANA and LLM-Pruner, with perplexity scores increasing by a factor of 2-5x compared to the baseline. This reinforces our hypothesis that data which is merely "low perplexity" for the model is not necessarily informative for making coarse-grained pruning decisions and can be severely misleading. (3) Downstream Task Performance: While the baseline Nirvana model without any specialized selection performs surprisingly well on some downstream tasks, OASIS provides the most robust and balanced performance profile. It secures the top results on several tasks (ARC-e, HellaS, BoolQ) while remaining competitive on others, resulting in the highest average performance boost (Δ).

The results for Wanda on the Llama-3.2-1B

Table 5: Zero-shot performance of Llama3.2-3B after applying 50% structured pruning. We compare baseline pruning methods against versions enhanced with different calibration data selection strategies: entropy-based, synthetic data, and our proposed OASIS. **Bold** and underline denote the best and second-best results per group, respectively. ↓: lower is better.

Method	WikiT ↓	PTB ↓	LambD ↓	ARC-e	WinoG	HellaS	BoolQ	PIQA	Δ
Llama-3.2-3B	7.87	12.57	20.09	71.68	69.06	73.69	72.78	77.48	
Nirvana	61.87	112.03	105.24	33.80	51.78	34.75	57.28	**59.03**	
+ Entropy	61.87	115.58	105.24	32.45	51.38	34.27	47.71	57.94	
+ Synthetic	95.82	168.17	168.17	32.20	**54.30**	34.04	50.64	57.78	
+ OASIS (ours)	**58.12**	**108.58**	**102.00**	**34.13**	51.54	**35.12**	**59.11**	58.75	+0.41
LLM-Pruner	**196.62**	268.74	179.02	31.73	46.72	**28.92**	**60.37**	56.47	
+ Entropy	334.45	356.02	260.47	**32.66**	48.30	28.20	52.11	54.52	
+ Synthetic	938.00	605.62	486.63	30.89	48.54	28.09	48.72	54.30	
+ OASIS (ours)	**196.62**	**244.69**	**168.17**	31.81	**48.70**	28.64	59.14	**57.02**	+0.22

Table 6: Zero-shot performance of Llama-3.2-1B after applying 50% unstructured pruning. We compare baseline pruning methods against versions enhanced with different calibration data selection strategies: entropy-based, synthetic data, and our proposed OASIS. **Bold** denotes the best results per group. ↓: lower is better.

Method	WikiT ↓	PTB ↓	LambD ↓	ARC-e	WinoG	HellaS	BoolQ	PIQA	Δ
Llama-3.2-1B	9.64	16.65	23.48	60.27	59.98	63.66	63.88	74.27	
Wanda	21.05	38.72	50.50	49.92	55.49	45.01	62.02	66.10	
+ Entropy	21.05	38.72	52.10	50.21	55.72	**45.24**	61.37	66.10	
+ Synthetic	21.72	39.33	53.75	50.67	54.93	45.15	61.90	66.38	
+ OASIS(Ours)	21.05	**38.11**	50.50	**50.72**	**55.96**	44.98	**62.48**	**66.65**	+0.35

model, presented in Table 6, confirm that OASIS remains beneficial even for less sensitive, fine-grained pruning methods. (1) Modest but Consistent Gains: While the performance differences are much smaller compared to structured pruning, OASIS consistently outperforms the other selection strategies on the majority of downstream tasks. It achieves the highest scores on ARC-e, WinoG, and BoolQ, leading to the largest overall performance boost (Δ=+0.35). (2) Minimal Impact on Perplexity: For unstructured pruning at 50% sparsity, the impact on perplexity is generally low across all methods. However, OASIS still manages to achieve a slight edge, securing the best performance on PTB. This suggests that even when the overall degradation is minimal, a principled selection method can still find a better optimization path than heuristic approaches. (3) Overall, these additional experiments strengthen our central claim that OASIS is a versatile and robust tool for enhancing post-training pruning. It provides substantial benefits where they are needed most (structured pruning) while still offering consistent, marginal gains in less sensitive scenarios (unstructured pruning).

Table 7: Method-specific wall-clock time and peak GPU memory on Llama-3.1-8B using one H100. OASIS is the three-epoch selection stage.

Pruner	**Native pruning**		**OASIS**	
	Time	**Memory**	**Time**	**Memory**
NIRVANA	70 s	40 GB	91 min	75 GB
LLM-Pruner	118 s	42 GB	89 min	66 GB
Wanda	73 s	37 GB	64 min	59 GB
SparseGPT	280 s	38 GB	70 min	61 GB

A.4 Efficiency and Post-Pruning Recovery

We profile native pruning and the three-epoch OASIS selection stage on Llama-3.1-8B using one NVIDIA H100-80GB (Table 7). OASIS is a one-time offline step and introduces no inference-time module, but its 64–91 minute cost is substantial relative to the 70–280 second native pruning runs. This trade-off is most compelling for structured pruning, where the performance differences are larger.

After the same LoRA recovery procedure, the OASIS-calibrated model retains lower perplex-

Table 8: Post-pruning LoRA recovery on Stanford Alpaca for NIRVANA-pruned Llama-3.1-8B. Both runs use one H100 for approximately two hours.

Calibration	PTB ↓	LambD ↓	ARC-e	WinoG	HellaS	BoolQ	PIQA	ARC-c	OBQA	COPA	Avg Acc.
NIRVANA	44.15	50.22	48.48	57.30	43.88	61.99	63.74	28.58	27.40	67.00	49.80
NIRVANA + OASIS	**42.05**	**47.36**	**49.20**	**58.64**	**46.37**	**62.08**	**65.23**	**30.46**	**28.00**	**68.00**	**51.00**

Table 9: PPL transferability of the 50% OASIS calibration selection to 25% sparsity on Qwen2.5-7B with NIRVANA. The OASIS row at 25% reuses the calibration subset and weights selected at 50% sparsity.

Method	50% sparsity			25% sparsity		
	PTB ↓	LambD ↓	Avg ↓	PTB ↓	LambD ↓	Avg ↓
NIRVANA	108.58	157.98	133.28	22.41	41.86	32.14
+ Entropy	70.11	87.25	78.68	22.41	34.10	28.26
+ Synthetic	87.25	108.58	97.91	25.39	38.72	32.06
+ OASIS	**59.96**	**84.56**	**72.26**	**21.72**	**33.64**	27.68

ity and improves average accuracy from 49.80 to 51.00. This suggests that calibration selection and post-pruning adaptation are complementary in this setting rather than the selection advantage being erased by fine-tuning.

A.5 Transferability of the selected data

We evaluate transfer across sparsity levels in Table 9: the subset learned at 50% sparsity is reused without reselection at 25%. OASIS remains strongest in this single Qwen2.5-7B/NIRVANA comparison, providing evidence of transfer across this moderate sparsity shift rather than a general transfer guarantee.

B Context-Dependency of Calibration Data Preference

In the main paper, we argue that the quality of a calibration sample is not a universal property but is context-dependent on the pruning method and model. To provide evidence for this claim, we conduct an additional analysis on two widely-used unstructured pruning methods, Wanda and SparseGPT (this is to minimize the variance brought in by structured pruning VS unstructured pruning), across the Llama model family (1B, 3B, and 8B).

The results in Figure 6 highlight two key aspects of this context-dependency: (1) **Dependence on Pruning Method:** The "best" and "worst" calibration samples are clearly different for Wanda and SparseGPT. For example, sample #1 is consistently the most detrimental for Wanda across all model sizes, causing a massive spike in perplexity. However, for SparseGPT, sample #1 performs reasonably well. (2) **Dependence on Model Architecture/Size:** The consistency of data preference across a model family also varies by method. For Wanda (right column), the performance trend is remarkably consistent across the 1B, 3B, and 8B models; a sample that is good for the 1B model is generally also good for the 8B model. In contrast, for SparseGPT (left column), the trend is much less consistent. For instance, sample #7 yields one of the worst PPLs for the 1B model but delivers average performance on the 3B and 8B models. We hypothesize this difference in stability stems from the nature of their saliency scores. Wanda uses a zero-order metric, $S = \|\mathbf{W}\mathbf{X}^2\|$, which is based on parameter weights and input activations, making it relatively stable. In contrast, SparseGPT's score, $S = \|\mathbf{W}\|^2/\text{diag}(\mathbf{H})$, is a second-order metric that incorporates the Hessian matrix ($\mathbf{H}$). This inclusion of higher-order information, which captures the curvature of the loss function, makes its importance estimation more complex and sensitive to the specific model state and calibration data.

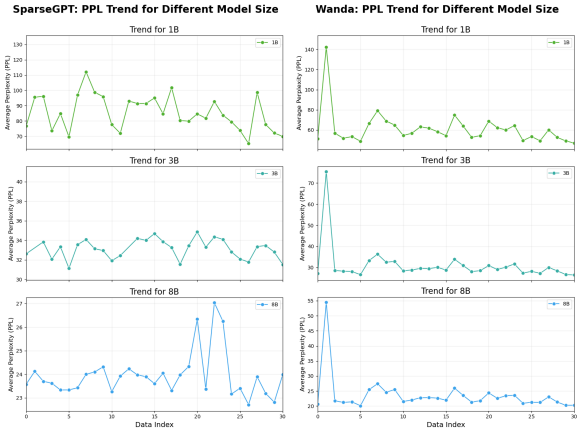


Figure 6: Perplexity (PPL) trends after pruning with individual calibration samples. (Left Column) SparseGPT results on Llama 1B, 3B, and 8B models. (Right Column) Wanda results on the same models. The x-axis indexes the different data samples used for calibration. The plots reveal that data preference is highly dependent on both the pruning method and, in some cases, the model size.

Together, these results strongly support our central claim: the notion of a "golden" calibration sample is not static. It is a function of the complex interaction between the data, the pruning algorithm, and the model architecture.

C Details on micro-level analysis

C.1 Impact of a Single Detrimental Sample

We test the operational "poisoning effect" by injecting one detrimental sample, identified in the micro-analysis, into a selected set of 32 samples. Table 10 shows a large perplexity increase for NIRVANA and a smaller increase for Wanda.

Table 10: Effect of adding one detrimental sample to an OASIS-selected set on Llama-3.2-1B. ↓ denotes lower is better.

Configuration	WikiT ↓	PTB ↓	LambD ↓	Avg ↓
Structured: NIRVANA				
Golden set (OASIS)	179.02	202.86	182.65	188.18
Golden set + 1 detrimental sample	**196.62**	**378.99**	**184.70**	**253.44**
Unstructured: Wanda				
Golden set (OASIS)	**21.05**	**38.11**	**50.50**	**36.76**
Golden set + 1 detrimental sample	22.24	39.21	52.27	37.91

As a control that does not condition on an OASIS-selected set, we also inject one detrimental sample into a random Llama-3.1-8B calibration set. For NIRVANA, PTB/Lambada PPL increases from 82.48/106.92 to 93.15/113.39; for Wanda, it changes only from 14.69/25.00 to 14.74/25.00. The effect therefore reflects broader structured-pruning sensitivity in this experiment and is negligible for Wanda, rather than being unique to OASIS.

C.2 Detailed Result on Sample Combination

In Section 3.2, we provide the micro analysis on calibration data. Here, we provide the details of that analysis in the following tables, with an additional result on SparseGPT.

D The Use of Large Language Models (LLMs)

We use LLMs in this work in order to revise the paper writing and check for grammar errors.

Table 11: Synergistic and Degradative Effects of Combining Calibration Samples from Different Quality Tiers for the **Wanda** Pruning Method. The change (Δ) is calculated as (Combination PPL - Best Individual PPL). Negative values (green) indicate synergistic improvement, while positive values (red) indicate performance degradation.

Category	Sample 1 (PPL)	Sample 2 (PPL)	Combination PPL	Change (Δ)	Observed Effect
Gold + Gold	77 (45.8)	111 (45.9)	**45.2**	-0.6	Synergy
	91 (46.8)	103 (46.6)	**43.4**	-3.2	Strong Synergy
	77 (45.8)	103 (46.6)	**44.2**	-1.6	Synergy
	77 (45.8)	91 (46.8)	**44.5**	-1.3	Synergy
	91 (46.8)	110 (46.8)	**44.8**	-2.0	Strong Synergy
Mediocre + Mediocre	9 (64.8)	23 (64.4)	**57.7**	-6.7	Strong Synergy
	23 (64.4)	54 (67.1)	**59.0**	-5.4	Strong Synergy
	23 (64.4)	33 (64.5)	**57.4**	-7.0	Strong Synergy
	8 (68.8)	16 (74.8)	**64.5**	-4.3	Synergy
	20 (68.7)	54 (67.1)	**60.5**	-6.6	Strong Synergy
Bad + Bad	1 (142.3)	66 (148.6)	**111.0**	-31.3	Improvement
	1 (142.3)	95 (117.3)	**95.9**	-21.4	Improvement
	1 (142.3)	86 (115.4)	**93.4**	-22.0	Improvement
	66 (148.6)	95 (117.3)	**116.5**	-0.8	Slight Improvement
	66 (148.6)	86 (115.4)	115.8	+0.4	Neutral / No Effect
Gold + Mediocre	77 (45.8)	58 (70.5)	49.5	+3.7	Degradation
	111 (45.9)	17 (63.9)	48.6	+2.7	Degradation
	115 (46.5)	33 (64.5)	48.3	+1.8	Degradation
	103 (46.6)	61 (64.8)	47.2	+0.6	Slight Degradation
	110 (46.8)	6 (66.4)	**45.9**	-0.9	Synergy (Outlier)
Gold + Bad	77 (45.8)	66 (148.6)	53.5	+7.7	Strong Degradation
	111 (45.9)	66 (148.6)	53.5	+7.6	Strong Degradation
	115 (46.5)	66 (148.6)	53.2	+6.7	Strong Degradation
	103 (46.6)	1 (142.3)	53.5	+6.9	Strong Degradation
	110 (46.8)	1 (142.3)	53.0	+6.2	Strong Degradation
Mediocre + Bad	9 (64.8)	1 (142.3)	70.5	+5.7	Degradation
	23 (64.4)	1 (142.3)	70.5	+6.1	Degradation
	8 (68.8)	66 (148.6)	76.1	+7.3	Strong Degradation
	33 (64.5)	66 (148.6)	71.7	+7.2	Strong Degradation
	20 (68.7)	95 (117.3)	73.5	+4.8	Degradation

Table 12: Synergistic Effects of Combining Multiple Golden Samples. The "Best Individual PPL" column serves as the baseline for each group. The Synergy (Δ) column, calculated as (Combination PPL - Best Individual PPL), quantifies the performance gain from the combination. All Δ values are negative, indicating a consistent synergistic effect. Random 10 samples avgppl is 43.3.

Category	# Samples (k)	Sample IDs in Combination	Best Individual PPL	Combination PPL	Synergy (Δ)
Golden Trio ($k = 3$)	3	77, 103, 111	45.8	**44.1**	-1.7
	3	77, 91, 111	45.8	**44.4**	-1.4
	3	77, 91, 103	45.8	**43.4**	-2.4
	3	91, 103, 111	45.9	**43.9**	-2.0
	3	77, 93, 110	45.8	**43.1**	-2.7
Golden Quad ($k = 4$)	4	77, 91, 103, 111	45.8	**43.0**	-2.8
	4	77, 93, 103, 110	45.8	**43.4**	-2.4
	4	77, 93, 107, 111	45.8	**42.3**	-3.5
	4	91, 93, 107, 115	46.5	**44.2**	-2.3
	4	91, 103, 111, 115	45.9	**44.2**	-1.7
Golden All-In	10	30, 77, 91, 93, 101, 103, 107, 110, 111, 115	45.8	**42.6**	-3.2

Table 13: Synergistic and Degradative Effects on the **SparseGPT** Method. Note the inconsistent effect in the "Gold + Gold" category and the synergistic effect in some "Mediocre + Bad" cases, which differ from the Wanda method's behavior.

Category	Sample 1 (PPL)	Sample 2 (PPL)	Combination PPL	Change (Δ)	Observed Effect
Gold + Gold	26 (65.2)	100 (65.4)	**62.8**	-2.4	Synergy
	26 (65.2)	94 (64.4)	**63.1**	-1.3	Synergy
	100 (65.4)	106 (65.1)	67.5	+2.4	Degradation
	26 (65.2)	106 (65.1)	66.6	+1.5	Degradation
	94 (64.4)	106 (65.1)	**63.8**	-0.6	Synergy
Mediocre + Mediocre	51 (84.3)	56 (84.1)	**79.9**	-4.2	Synergy
	18 (80.3)	97 (80.0)	**68.3**	-11.7	Strong Synergy
	18 (80.3)	31 (80.3)	**76.8**	-3.5	Synergy
	21 (81.7)	58 (81.8)	**73.9**	-7.8	Strong Synergy
	23 (83.7)	83 (82.5)	**80.6**	-1.9	Synergy
Bad + Bad	84 (120.6)	85 (123.2)	**114.6**	-6.0	Improvement
	67 (132.5)	84 (120.6)	**116.9**	-3.7	Improvement
	67 (132.5)	85 (123.2)	**109.8**	-13.4	Strong Improvement
	67 (132.5)	95 (119.3)	**106.1**	-13.2	Strong Improvement
	67 (132.5)	86 (119.2)	**106.7**	-12.5	Strong Improvement
Gold + Mediocre	26 (65.2)	58 (81.8)	70.2	+5.0	Degradation
	26 (65.2)	81 (84.0)	75.3	+10.1	Strong Degradation
	94 (64.4)	4 (84.9)	65.3	+0.9	Slight Degradation
	94 (64.4)	49 (85.4)	68.7	+4.3	Degradation
	100 (65.4)	103 (82.4)	**65.3**	-0.1	Neutral / No Effect
Gold + Bad	26 (65.2)	67 (132.5)	68.6	+3.4	Degradation
	94 (64.4)	67 (132.5)	67.7	+3.3	Degradation
	100 (65.4)	85 (123.2)	82.0	+16.6	Strong Degradation
	94 (64.4)	85 (123.2)	69.0	+4.6	Degradation
	106 (65.1)	84 (120.6)	71.0	+5.9	Degradation
Mediocre + Bad	31 (80.3)	67 (132.5)	81.3	+1.0	Slight Degradation
	21 (81.7)	67 (132.5)	**74.9**	-6.8	Synergy
	56 (84.1)	85 (123.2)	**82.0**	-2.1	Synergy
	18 (80.3)	84 (120.6)	**76.8**	-3.5	Synergy
	81 (84.0)	85 (123.2)	**81.5**	-2.5	Synergy

Table 14: Synergistic Effects of Combining Multiple Golden Samples for the **SparseGPT** Method. Note the "less is more" effect, where the best 4-sample combination outperforms the 10-sample combination. Random 10 samples avgppl is 61.0.

Category	# Samples (k)	Sample IDs in Combination	Best Individual PPL	Combination PPL	Synergy (Δ)
Golden Trio ($k = 3$)	3	26, 94, 100	64.4	**62.8**	-1.6
	3	26, 94, 106	64.4	**62.7**	-1.7
	3	94, 100, 106	64.4	**63.8**	-0.6
	3	26, 90, 106	65.1	**61.6**	-3.5
	3	90, 94, 106	64.4	65.2	+0.8
Golden Quad ($k = 4$)	4	26, 94, 100, 106	64.4	**63.0**	-1.4
	4	94, 100, 105, 106	64.4	**63.8**	-0.6
	4	26, 90, 94, 100	64.4	**61.3**	-3.1
	4	90, 94, 100, 106	64.4	**60.6**	-3.8
	4	94, 100, 105, 106	64.4	**63.8**	-0.6
Golden All-In	10	5, 26, 90, 91, 94, 100, 105, 106, 113, 114	64.4	**60.7**	-3.7

Table 15: Analysis of Data Combination Effects on NIRVANA. Note the inconsistent effects in "Gold + Gold" and the unstable behavior in "Bad + Bad" categories.

Category	Sample 1 (PPL)	Sample 2 (PPL)	Combination PPL	Change (Δ)	Observed Effect
Gold + Gold	111 (168.4)	110 (169.3)	175.6	+7.2	Degradation
	111 (168.4)	108 (173.7)	**166.6**	-1.8	Synergy
	110 (169.3)	108 (173.7)	170.9	+1.6	Degradation
	111 (168.4)	55 (174.4)	**165.0**	-3.4	Strong Synergy
	110 (169.3)	55 (174.4)	172.6	+3.3	Degradation
Mediocre + Mediocre	8 (230.1)	5 (232.0)	**218.5**	-11.6	Strong Synergy
	5 (232.0)	3 (233.2)	**225.6**	-6.4	Synergy
	10 (232.3)	114 (232.5)	**209.4**	-22.9	Strong Synergy
	15 (235.0)	78 (237.1)	**205.3**	-29.7	Strong Synergy
	3 (233.2)	61 (239.8)	**223.3**	-9.9	Strong Synergy
Bad + Bad	54 (473.6)	62 (426.9)	528.8	+101.9	Catastrophic Degradation
	102 (423.4)	54 (473.6)	444.6	+21.2	Strong Degradation
	62 (426.9)	102 (423.4)	**392.7**	-30.7	Strong Improvement
	43 (418.4)	54 (473.6)	602.3	+183.9	Catastrophic Degradation
	43 (418.4)	62 (426.9)	**342.4**	-76.0	Strong Improvement
Gold + Mediocre	111 (168.4)	8 (230.2)	182.9	+14.5	Strong Degradation
	111 (168.4)	115 (230.7)	183.2	+14.8	Strong Degradation
	110 (169.3)	78 (237.1)	187.0	+17.7	Strong Degradation
	110 (169.3)	97 (246.3)	193.2	+23.9	Strong Degradation
	108 (173.7)	8 (230.2)	207.1	+33.4	Strong Degradation
Gold + Bad	111 (168.4)	54 (473.6)	315.8	+147.4	Severe Degradation
	111 (168.4)	62 (426.9)	280.6	+112.2	Severe Degradation
	110 (169.3)	54 (473.6)	347.9	+178.6	Severe Degradation
	110 (169.3)	6 (325.2)	304.8	+135.5	Severe Degradation
	108 (173.7)	102 (423.4)	280.6	+106.9	Severe Degradation
Mediocre + Bad	15 (235.0)	54 (473.6)	368.5	+133.5	Severe Degradation
	114 (232.5)	54 (473.6)	349.7	+117.2	Severe Degradation
	8 (230.2)	54 (473.6)	345.0	+114.8	Severe Degradation
	3 (233.2)	62 (426.9)	331.2	+98.0	Severe Degradation
	71 (232.6)	62 (426.9)	352.8	+120.2	Severe Degradation

Table 16: Synergistic Effects of Combining Multiple Golden Samples for NIRVANA. The "less is more" effect is highly pronounced, with the 10-sample combination performing significantly worse than smaller, well-chosen sets.

Category	# Samples (k)	Sample IDs in Combination	Best Individual PPL	Combination PPL	Synergy (Δ)
Golden Trio ($k = 3$)	3	111, 110, 108	168.4	178.0	+9.6
	3	110, 108, 55	169.3	**168.9**	-0.4
	3	111, 110, 55	168.4	**165.2**	-3.2
	3	111, 110, 59	168.4	189.1	+20.7
	3	108, 55, 59	173.7	191.0	+17.3
Golden Quad ($k = 4$)	4	111, 110, 108, 55	168.4	**167.2**	-1.2
	4	111, 110, 108, 59	168.4	181.5	+13.1
	4	111, 110, 55, 59	168.4	175.9	+7.5
	4	110, 108, 55, 59	169.3	178.0	+8.7
	4	110, 108, 55, 59	169.3	175.9	+6.6
Golden All-In	10	111, 110, 108, 55, 59, 64, 31, 37, 56, 39	168.4	197.1	+28.7