Iterative Structured Pruning for Large Language Models with Multi-Domain Calibration

Guangxin Wu[1,2*], **Hao Zhang**[1,2,3*], **Zhibin Zhang**[1], **Jiafeng Guo**[1], **Xueqi Cheng**[1]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
(wuguangxin24, zhanghao233)@mails.ucas.ac.cn
(zhangzhibin, guojiafeng, cxq)@ict.ac.cn

Abstract

Large Language Models (LLMs) have achieved remarkable success across a wide spectrum of natural language processing tasks. However, their ever-growing scale introduces significant barriers to real-world deployment, including substantial computational overhead, memory footprint, and inference latency. While model pruning presents a viable solution to these challenges, existing unstructured pruning techniques often yield irregular sparsity patterns that necessitate specialized hardware or software support. In this work, we explore structured pruning, which eliminates entire architectural components and maintains compatibility with standard hardware accelerators. We introduce a novel structured pruning framework that leverages a hybrid multi-domain calibration set and an iterative calibration strategy to effectively identify and remove redundant channels. Extensive experiments on various models across diverse downstream tasks show that our approach achieves significant compression with minimal performance degradation.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language processing, enabling a wide range of applications such as question answering, summarization, and code generation (Ding et al., 2022; Qin et al., 2023; Zhu et al., 2023; Li et al., 2023a). Moreover, these models also demonstrate exceptional performance across a wide range of other domains, including medicine (Qi et al., 2025a; Luo et al., 2025; Cong et al., 2025; Qi et al., 2025b), security (Ma et al., 2025; Wu et al., 2025), and various social tasks (Zhang et al., 2025b,a; Zheng et al., 2025b,a). As model sizes continue to grow, LLMs exhibit emergent behaviors and enhanced reasoning abilities. However, the increasing scale and complexity of these models pose significant challenges for practical deployment. The substantial computational and memory requirements lead to high inference latency, elevated energy consumption, and strict hardware constraints, which limit their usability in resource-constrained or real-time settings (Zhang et al., 2023; Huang et al., 2023; Wang et al., 2023). These challenges highlight the urgent need for effective model compression and acceleration techniques that align with the unique characteristics of LLMs.

Among various solutions, model pruning (Ma et al., 2023; Ashkboos et al., 2024; Li et al., 2023b; Han et al., 2015) has emerged as a particularly promising direction. It can be broadly categorized into unstructured pruning and structured pruning. Unstructured pruning (Liao et al., 2023; Anonymous, 2024) removes individual weights from parameter matrices, but often results in irregular sparsity patterns that demand specialized hardware and software for efficient execution. This irregularity not only complicates storage and inference but also reduces portability and scalability. Common unstructured approaches evaluate the significance of individual parameters and eliminate those with minimal impact, followed by adjustments to the remaining weights. While effective in some cases, these methods disrupt the model's structural coherence.

Structured pruning (Ashkboos et al., 2024; Yang and Zhang, 2022) offers an alternative that addresses these limitations by removing entire architectural components such as neurons, channels, or layers. This type of pruning simplifies the model at a coarser granularity, making the resulting models more compatible with general-purpose hardware and standard deep learning frameworks. It reduces both computational overhead and memory usage while preserving the high-level structure of the original model.

In this work, we present a new structured pruning

*These authors contribute equally to this work.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 1–10
March 25-27, 2026

framework that integrates a hybrid calibration set drawn from multiple domains with an iterative calibration strategy. This design enables accurate identification of redundant channels with minimal loss in model performance. By combining diverse data representations with a progressive pruning process, our method achieves efficient model compression and strong generalization across downstream tasks. Extensive experiments on a variety of LLM architectures demonstrate that our approach outperforms existing structured pruning baselines in terms of both compression ratio and accuracy preservation. Our contributions are summarized as follows:

- **Multi-domain hybrid calibration set.** We design a diverse calibration dataset that spans multiple domains, including Wikipedia articles, Common Crawl data, code repositories, and mathematical texts. This diversity enables the pruning process to generalize more effectively across a wide range of linguistic and semantic patterns.
- **Iterative channel selection.** We propose an iterative calibration strategy that incrementally refines the choice of channels to prune. This progressive refinement improves both the accuracy of channel selection and the robustness of the pruned model.
- **Comprehensive evaluation.** We evaluate our approach on the Qwen2.5 families using a broad set of downstream tasks and datasets. Our method consistently achieves strong performance while delivering substantial model compression.

2 Related Work

2.1 Compression Techniques for Large Language Models

With the rapid growth of large language models (LLMs) containing billions of parameters, efficient and scalable compression has become increasingly essential. Knowledge distillation (Yang et al., 2021; Zhang et al., 2024), though effective, is often impractical at this scale due to the high cost of training student models. Quantization methods (Zhou et al., 2023; Cai et al., 2023; Zhou et al., 2024) reduce memory and computation by lowering numerical precision, but face challenges in LLMs such as activation outliers and sensitivity to precision errors that can significantly degrade performance.

2.2 Structured Pruning for Neural Networks

Network pruning is a long-standing approach for compressing neural networks by removing redundant parameters (Ma et al., 2023; Ashkboos et al., 2024; Li et al., 2023b; Han et al., 2015; Yang and Zhang, 2022). Early unstructured pruning methods eliminate individual weights based on magnitude or sensitivity, achieving high sparsity but poor hardware efficiency. In contrast, structured pruning removes entire channels, neurons, or attention heads, preserving layer regularity and enabling efficient parallel computation and memory access. Recent advances (Ma et al., 2023) extend structured pruning to transformer architectures, employing criteria such as ℓ_1 norms, gradient signals, and second-order approximations. Post-training structured pruning further enables compression without full retraining, though lightweight fine-tuning is often required to recover performance after aggressive pruning.

3 Methodology

In this section, we present a structured pruning framework for large language models that integrates a variance-based importance criterion from FLAP (An et al., 2024), a domain-diverse calibration dataset to enhance generalization across input distributions, and an iterative calibration strategy that refines pruning decisions by accounting for cumulative pruning effects, improving stability and final performance.

3.1 Preliminary

Recent studies introduce bias compensation to mitigate pruning-induced output shifts. In structured pruning, the output of an uncompressed layer can be expressed as follows:

$$W^\ell X^\ell = \underbrace{(M^\ell \odot W^\ell)X^\ell}_{\text{Retained Part}} + \underbrace{((1-M^\ell)\odot W^\ell)X^\ell}_{\text{Removed Part}} \quad (1)$$

where W^ℓ and X^ℓ denote the weights and inputs of the ℓ-th layer, and $M^\ell \in \{0,1\}^{\text{shape}(W^\ell)}$ is a binary mask indicating the retained structures. The goal is to minimize the influence of the removed part, $\Delta Y^\ell = ((1-M^\ell)\odot W^\ell)X^\ell$, on the output feature map. To compensate for this error, a bias term can be constructed from the mean input activations over tokens and samples for each channel as follows:

$$\overline{\mathbf{X}}^\ell_{:,j,:} = \frac{1}{NL}\sum_{n=1}^{N}\sum_{k=1}^{L}\mathbf{X}^\ell_{n,j,k} \quad (2)$$

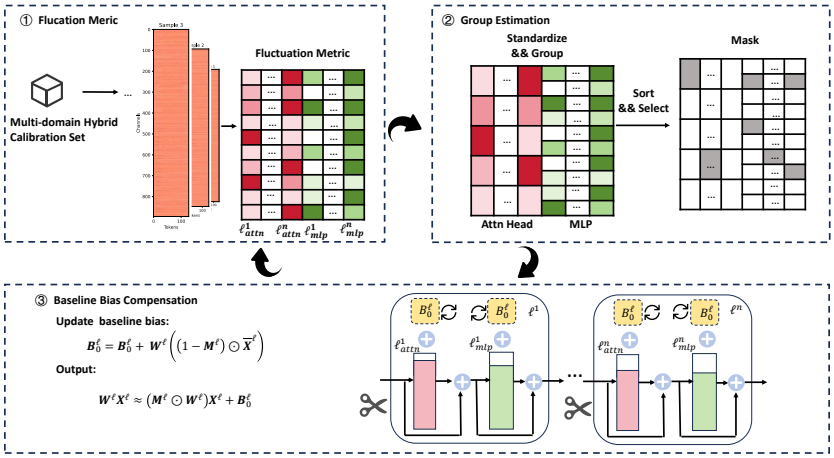


Figure 1: Overview of our proposed method.

After determining the pruning mask M_ℓ, the baseline activations of pruned channels are transformed into a bias vector as follows:

$$\mathbf{B}_0^\ell = \mathbf{W}^\ell\big((1-\mathbf{M}^\ell)\odot\overline{\mathbf{X}}^\ell\big) \tag{3}$$

$$\mathbf{W}^\ell\mathbf{X}^\ell \approx (\mathbf{M}^\ell\odot\mathbf{W}^\ell)\mathbf{X}^\ell + \mathbf{B}_0^\ell \tag{4}$$

where $\mathbf{B}_0^\ell \in \mathbb{R}^{C_{\text{out}}}$ approximates the output of the original layer. Channel importance depends on both input variance and weight magnitude. A fluctuation metric is defined as follows:

$$\mathbf{S}_{:,j}^\ell = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{X}_{n,j,:}^\ell - \overline{\mathbf{X}}_{:,j,:}^\ell)^2\cdot\|\mathbf{W}_{:,j}^\ell\|^2 \tag{5}$$

and channels with lower fluctuation scores are pruned, with the resulting error compensated by $\mathbf{B}_0^\ell$.

Compared to incremental pruning methods that analytically adjust weights after each removal step, this bias-based strategy prunes all target structures in one shot and compensates the output shift using the estimated bias term. It eliminates retraining and is computationally efficient, but its effectiveness depends on accurate activation statistics obtained from calibration data. To enhance robustness, we propose two extensions: (i) constructing a domain-diverse calibration dataset to better capture activation statistics, and (ii) introducing an iterative calibration strategy to mitigate cascading errors in one-shot pruning. These components are detailed below, and Figure 1 provides an overview of the method.

3.2 Multi-domain Hybrid Calibration Set

To enable structured pruning that generalizes across diverse real-world applications, we construct a domain-diverse calibration dataset. Prior pruning methods typically rely on calibration sets from a single or narrow domain, which biases importance estimation toward domain-specific features and reduces robustness in heterogeneous environments where input distributions vary widely.

Formally, consider K distinct domains $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$, each with input distribution $P_k(\mathbf{X})$. For the ℓ-th layer, the mean activation and variance in domain k are defined as follows:

$$\overline{\mathbf{X}}_k^\ell = \mathbb{E}_{\mathbf{X}\sim P_k}[\mathbf{X}^\ell], \quad \mathbf{V}_k^\ell = \mathbb{E}_{\mathbf{X}\sim P_k}[(\mathbf{X}^\ell - \overline{\mathbf{X}}_k^\ell)^2] \tag{6}$$

which capture domain-specific activation patterns shaped by linguistic or semantic properties. A single domain calibration dataset samples only from $P_k(\mathbf{X})$, yielding biased importance metrics that may degrade out-of-domain performance. To mitigate this, we construct a calibration dataset across diverse domains including natural language, source code and mathematical reasoning , ensuring broad coverage of linguistic and logical patterns. The combined calibration distribution is modeled as follows:

$$P_{\text{calib}}(\mathbf{X}) = \textstyle\sum_{k=1}^{K}\alpha_k P_k(\mathbf{X}), \alpha_k \geq 0,\ \sum_{k=1}^{K}\alpha_k = 1 \tag{7}$$

where α_k reflects each domain's relative importance. The overall statistics for pruning at layer ℓ are defined as follows:

$$\overline{\mathbf{X}}^\ell = \sum_{k=1}^{K}\alpha_k\overline{\mathbf{X}}_k^\ell, \quad \mathbf{V}^\ell = \sum_{k=1}^{K}\alpha_k\mathbf{V}_k^\ell \tag{8}$$

providing more representative importance estimates. Calibrating with this domain-diverse dataset enables the pruning algorithm to capture heterogeneous activation behaviors across linguistic and

reasoning tasks, yielding more robust and generalizable pruning decisions for large language models.

3.3 Iterative Calibration Strategy

During pruning, removing certain channels c_k in layer ℓ_i inevitably alters the activation statistics of downstream channels c_t in layers ℓ_j with $j > i$. Specifically, the baseline activation and variance are defined as follows:

$$b_t^{(j)} = \mathbb{E}[X_{c_t}^{(\ell_j)}], \quad v_t^{(j)} = \mathrm{Var}[X_{c_t}^{(\ell_j)}] \tag{9}$$

Single step calibration methods, such as FLAP, estimate these statistics only once before pruning. For instance, a channel c_k in ℓ_i may be pruned for low variance $v_k^{(i)}$, while a channel c_t in ℓ_j is retained for high variance $v_t^{(j)}$. However, pruning c_k and compensating it with a fixed bias replaces its activations with constants, shifting downstream distributions. Consequently, the variance of c_t may drop sharply as follows:

$$v_t^{(j)} \to v_t^{(j)'} \ll v_t^{(j)} \tag{10}$$

potentially making c_t redundant. This reveals a limitation of single-pass calibration: pruning decisions ignore cascading effects from earlier layers. If the pruning mask at step s is $M^{(s)}$, then the variance can be expressed as follows:

$$v_t^{(j,s)} = \mathrm{Var}\big[X_{c_t}^{(\ell_j)} \mid M^{(1)}, \ldots, M^{(s-1)}\big] \tag{11}$$

showing that channel variances depend on all prior pruning steps, while single-step methods assume $s = 1$.

To address this, we introduce an iterative calibration strategy that updates channel importance after each pruning step. At iteration s, recalibrated statistics are computed as follows:

$$b_t^{(j,s)} = \mathbb{E}[X_{c_t}^{(\ell_j)} \mid M^{(1)}, \ldots, M^{(s-1)}] \tag{12}$$

$$v_t^{(j,s)} = \mathrm{Var}[X_{c_t}^{(\ell_j)} \mid M^{(1)}, \ldots, M^{(s-1)}] \tag{13}$$

and pruning decisions are based on these refined estimates, allowing dynamically updated importance evaluation. The process continues until a target pruning ratio or convergence criterion is reached. By modeling cascading dependencies, this strategy yields more accurate importance estimation, better global optimization of pruning masks, and improved post-pruning accuracy. Its iterative nature also enables gradual adaptation, reducing reconstruction errors compared with one-shot pruning.

Overall, the iterative calibration can be formulated as minimizing reconstruction error over pruning masks M as follows:

$$\min_M \mathbb{E}_{\mathbf{X} \sim P_{\text{calib}}} \left[\|Y - \widehat{Y}(M; \mathbf{X})\|^2 \right] \tag{14}$$

where Y and $\widehat{Y}$ denote the outputs of the original and pruned models, respectively, and M is iteratively updated using refined activation statistics.

4 Experiments

4.1 Experimental Setup

Models and Datasets. To assess the effectiveness of our proposed method, we perform experiments on the Qwen2.5 model family, encompassing Qwen2.5-7B, Qwen2.5-14B, and Qwen2.5-32B variants (Yang et al., 2024). We evaluate zero-shot performance on six widely-used commonsense reasoning benchmarks: ARC-Challenge (Clark et al., 2018), ARC-Easy (Clark et al., 2018), HellaSwag (Zellers et al., 2019), OpenBookQA (OBQA) (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), and Winogrande (Sakaguchi et al., 2021).

Baselines. We benchmark our approach against two representative structured pruning methods: Wanda-sp (Sun et al., 2023) and FLAP (An et al., 2024). It is worth noting that Wanda-sp is an extension of the original Wanda method tailored for structured pruning.

Implementation Details. Our code is implemented using the PyTorch (Paszke et al., 2019) framework and Transformers (Wolf, 2020) libraries, with all experiments conducted on four NVIDIA A100 GPUs. For a fair and comprehensive comparison, all methods are evaluated under two pruning ratios: 25% and 50%. All evaluations are conducted using the LM-Harness (Gao et al., 2024).

4.2 Main Results

As shown in Tables 1 and 2, our method consistently surpasses existing structured pruning approaches across model scales and compression ratios. The performance gap over FLAP widens with larger models and higher pruning rates, highlighting the scalability and robustness of our approach. Specifically, on Qwen2.5-14B, the gain reaches 6% at 50% pruning; and on Qwen2.5-32B, it achieves 1.85% and 10.06% improvements at 25% and 50%, respectively. These results demonstrate that our iterative calibration effectively pre-

Method	Pruning Ratio	ARC-c	ARC-e	HellaSwag	OBQA	PIQA	Winogrande	Average
Qwen2.5-14B	0%	55.8	82.49	63.38	34.4	81.12	75.3	65.42
Wanda-sp(w_mix)	25%	37.12	63.59	**46.89**	**25.0**	**75.14**	58.25	51.0
FLAP(w_mix)		39.51	68.39	47.42	23.8	74.86	64.72	53.12
Ours(w_mix)		**39.76**	**68.77**	46.85	24.6	74.97	**68.67**	**53.94**
Wanda-sp(w_mix)	50%	21.5	27.23	25.73	14.6	54.08	49.41	32.09
FLAP(w_mix)		20.99	26.22	26.26	11.4	56.09	49.49	31.74
Ours(w_mix)		**21.42**	**39.52**	**30.49**	**16.4**	**62.62**	**53.67**	**37.35**

Table 1: Zero-shot performance of the compressed Qwen2.5-14B. Bold results highlight the best performance.

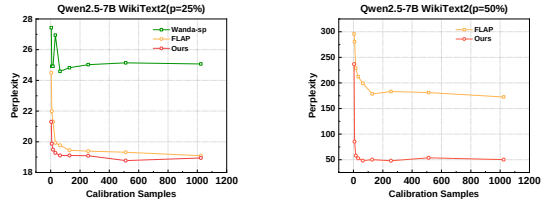
Method	Pruning Ratio	ARC-c	ARC-e	HellaSwag	OBQA	PIQA	Winogrande	Average
Qwen2.5-32 B	0%	53.41	80.51	64.91	34.2	81.88	75.3	65.04
Wanda-sp(w_mix)	25%	42.24	70.24	52.4	27.4	76.66	61.64	55.1
FLAP(w_mix)		42.24	72.85	55.02	28.6	78.02	72.53	58.21
Ours(w_mix)		**46.67**	**75.8**	**57.0**	**29.6**	**78.45**	**72.85**	**60.06**
Wanda-sp(w_mix)	50%	24.23	32.37	27.08	15.6	57.07	50.99	34.56
FLAP(w_mix)		22.7	36.36	29.43	15.6	64.36	51.07	36.59
Ours(w_mix)		**30.72**	**57.28**	**39.44**	**20.2**	**70.84**	**61.4**	**46.65**

Table 2: Zero-shot performance of the compressed Qwen2.5-32B. Bold results highlight the best performance.

serves task-relevant information and reasoning ability under aggressive compression.

4.3 Robustness to Calibration Samples

We assess the robustness of our method to the number of calibration samples on Qwen2.5-7B under 25% and 50% pruning using WikiText2. As shown in Figure 2a and Figure 2b, both FLAP and our method benefit from more calibration samples, as reflected in lower perplexity (PPL). Our method consistently outperforms FLAP, with the gap widening at higher pruning ratios. Notably, it achieves PPL $\approx$ 52 with only 32 samples and stabilizes near 50 with 128 or more, while FLAP remains above 170 at 50% pruning. These results show that our method better preserves model quality under high sparsity and is more robust to limited calibration data.

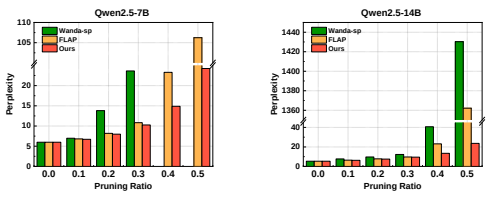


(a) Pruning ratio = 25% nsamples ablation study (b) Pruning ratio = 50% nsamples ablation study

Figure 2: Ablation study of nsamples on Qwen2.5-7B under different pruning ratios.

4.4 Different Pruning Ratios

We evaluate the robustness of our method across pruning ratios on Qwen2.5-7B and Qwen2.5-14B, comparing with Wanda-sp and FLAP. As shown in Figure 3a and Figure 3b, our method consistently outperforms both baselines, with the advantage increasing as pruning becomes more aggressive. On Qwen2.5-7B, at 50% pruning, Wanda-sp collapses (PPL $>$ 6800) and FLAP degrades severely (PPL $>$ 106), while our method maintains a low PPL of 24.2. A similar pattern appears on Qwen2.5-14B, where Wanda-sp and FLAP reach PPLs of 1430 and 1362, respectively, whereas our method achieves only 23.7. These results confirm that our iterative compensation strategy enables stable, high-quality performance even under extreme sparsity.

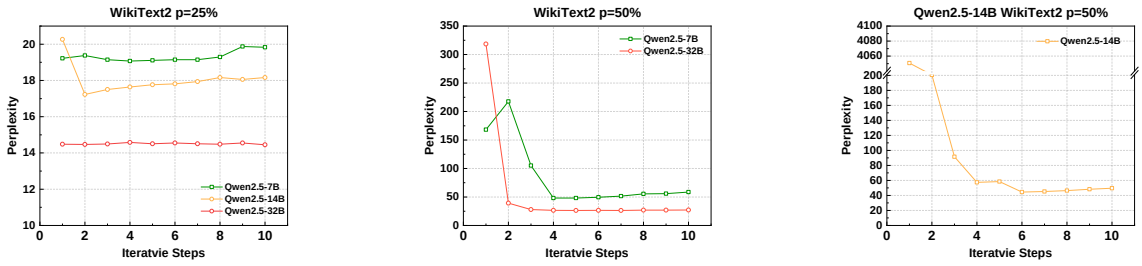


(a) Qwen2.5-7B ratios ablation study (b) Qwen2.5-14B ratios ablation study

Figure 3: Ablation studies on pruning ratios for Qwen2.5 models.

Method	Pruning Ratio	ARC-c	ARC-e	HellaSwag	OBQA	PIQA	Winogrande	Average
Qwen2.5-14B	0%	55.8	82.49	63.38	34.4	81.12	75.3	65.42
Ours	25%	**41.64**	**70.5**	44.73	**28.0**	71.16	67.72	**53.96**
Ours(w_mix)		39.76	68.77	**46.85**	24.6	**74.97**	**68.67**	53.94
Ours	50%	20.48	39.18	29.14	**16.8**	58.92	50.91	35.9
Ours(w_mix)		**21.42**	**39.52**	**30.49**	16.4	**62.62**	**53.67**	**37.35**

Table 3: Performance Comparsion of the compressed Qwen2.5-14B with and without multi-domain hybrid calibration set. Bold results highlight the best performance.



(a) Pruning ratio = 25% (b) Pruning ratio = 50% (c) Qwen2.5-14B

Figure 4: Ablation studies on iterative pruning steps across different pruning ratios and models.

4.5 Ablation Study

To comprehensively analyze the individual contribution of each component in our proposed framework, we conducted a series of ablation studies. These experiments specifically investigate the effectiveness of incorporating a multi-domain hybrid calibration set, as well as systematically assess the impact of the iterative pruning strategy.

Multi-domain Hybrid Calibration Set. Activation statistics (e.g., channel-wise mean and variance) vary across data domains, affecting pruning accuracy. To address this, we introduce a multi-domain hybrid calibration set to capture broader activation variations. We evaluate this design on Qwen2.5-14B under 25% and 50% pruning, comparing single-domain calibration with our hybrid approach. As shown in Tables 3, the hybrid setting consistently outperforms the single-domain variant, achieving higher zero-shot accuracy on average. These results confirm that multi-domain calibration provides more robust channel importance estimation and improves structured pruning performance.

Iterative Pruning. We study the effect of iterative pruning steps on model quality using Qwen2.5-7B, Qwen2.5-14B, and Qwen2.5-32B with WikiText2 calibration under 25% and 50% pruning. As shown in Figure 4, model perplexity remains stable across step counts at 25% pruning, indicating low sensitivity in this regime. In contrast, at 50% pruning, iterative pruning significantly improves performance: perplexity decreases with more steps, especially within the first three to four iterations. For instance, on Qwen2.5-14B, single-shot pruning causes severe degradation , while six iterative steps reduce it to about 44. These results clearly show that gradual, multi-step pruning is crucial for maintaining quality under high sparsity, and that four to six iterations are typically sufficient to achieve most of the gains, consistently across all evaluated datasets.

5 Conclusion

In this work, we introduce a novel structured pruning framework that synergistically integrates a multi-domain hybrid calibration set with an iterative, progressive pruning strategy. This design facilitates more precise identification of redundant channels while maintaining model performance across a wide spectrum of tasks. Comprehensive evaluations on multiple state-of-the-art large language models demonstrate that our approach consistently surpasses existing baselines, achieving substantial compression with minimal degradation in accuracy. These findings underscore the critical role of diverse calibration data and gradual pruning schedules in enabling efficient model compression.

Limitations

In this work, we conduct extensive experiments to evaluate the effectiveness of our pruning method. The results demonstrate that our approach achieves competitive performance compared to the baselines. However, due to computational constraints, we have not yet been able to evaluate it on larger scale models, such as those with 70 billion parameters. Exploring the scalability of our method to such large models constitutes an important direction for future work.

References

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

Anonymous. 2024. Unstructured pruning and low rank factorisation of self-supervised pre-trained speech models. *IEEE Transactions on Audio, Speech, and Language Processing*, pages 1046–1058.

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. *arXiv preprint arXiv:2401.15024*.

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.

Yuchen Cai, Zhen Wang, Yujun Li, Sheng Wang, Zhiyuan Liu, and Maosong Sun. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2302.06557*.

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. *arXiv preprint arXiv:1803.05457*.

Zhaoyang Cong, Ziyang Wang, Hao Zhang, Guowei Zheng, Keming Cao, Lina Zhao, Ruipeng Song, Jianqing Li, and Chengyu Liu. 2025. Hierarchical multiscale feature fusion network for multi-center major depressive disorder classification with t1-weighted mri. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, volume 2025, pages 1–4.

G. Ding and 1 others. 2022. Efficient fine-tuning for resource-constrained systems. *Proceedings of the Machine Learning Conference*.

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.

Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

E. Huang and 1 others. 2023. Evaluating large language models in complex scenarios. *Journal of Computational Linguistics*.

C. Li and 1 others. 2023a. Fine-tuning techniques for efficient model adaptation. *AI Research Journal*.

Yong Li, Wei Du, Liquan Han, Zhenjian Zhang, and Tongtong Liu. 2023b. A communication-efficient, privacy-preserving federated learning algorithm based on two-stage gradient pruning and differentiated differential privacy. *Sensors*, 23(23):9305.

Sheng Liao and 1 others. 2023. Can unstructured pruning reduce the depth in deep neural networks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Yang Luo, Shiru Wang, Jun Liu, Jiaxuan Xiao, Rundong Xue, Zeyu Zhang, Hao Zhang, Yu Lu, Yang Zhao, and Yutong Xie. 2025. Pathohr: Breast cancer survival prediction on high-resolution pathological images. *arXiv preprint arXiv:2503.17970*.

Chenrui Ma, Rongchang Zhao, Xi Xiao, Hongyang Xie, Tianyang Wang, Xiao Wang, Hao Zhang, and Yanning Shen. 2025. Cad-vae: Leveraging correlation-aware latents for comprehensive fair disentanglement. *arXiv preprint arXiv:2503.07938*.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720.

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. *arXiv preprint arXiv:1809.02789*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Xuyin Qi, Zeyu Zhang, Canxuan Gang, Hao Zhang, Lei Zhang, Zhiwei Zhang, and Yang Zhao. 2025a.

Mediaug: Exploring visual augmentation in medical imaging. In *Annual Conference on Medical Image Understanding and Analysis*, pages 218–232. Springer.

Xuyin Qi, Zeyu Zhang, Huazhan Zheng, Mingxi Chen, Numan Kutaiba, Ruth Lim, Cherie Chiang, Zi En Tham, Xuan Ren, Wenxin Zhang, and 1 others. 2025b. Medconv: Convolutions beat transformers on long-tailed bone density prediction. *IJCNN2025*.

A. Qin and 1 others. 2023. Advances in state-of-the-art natural language processing. *Journal of NLP Research*.

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*.

F. Wang and 1 others. 2023. Practical applications of llms in specialized domains. *Specialized AI Applications*.

Thomas Wolf. 2020. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yu-Hang Wu, Yu-Jie Xiong, Hao Zhang, Jia-Chen Zhang, and Zheng Zhou. 2025. Sugar-coated poison: Benign generation unlocks llm jailbreaking. *EMNLP 2025 Findings*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhen Yang, Zilun Zhang, Sheng Wang, Jie Li, Meishan Zhang, Zhiyuan Liu, and Maosong Sun. 2021. Knowledge distillation: A survey. *arXiv preprint arXiv:2106.05860*.

Zhengwu Yang and Han Zhang. 2022. Comparative analysis of structured pruning and unstructured pruning. In *Frontier Computing*, page 112. Springer.

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

D. Zhang and 1 others. 2023. Parameter-efficient fine-tuning methods for llms. *Journal of Machine Learning Research*.

Heng Zhang, Haichuan Hu, Yaomin Shen, Weihao Yu, Yilei Yuan, Haochen You, Guo Cheng, Zijian Zhang, Lubin Gan, Huihui Wei, and 1 others. 2025a. Asymoe: Leveraging modal asymmetry for enhanced expert specialization in large vision-language models. *arXiv preprint arXiv:2509.12715*.

Heng Zhang, Tianyi Zhang, Yuling Shi, Xiaodong Gu, Yaomin Shen, Zijian Zhang, Yilei Yuan, Hao Zhang, and Jin Huang. 2025b. Can representation gaps be the key to enhancing robustness in graph-text alignment? *arXiv preprint arXiv:2510.12087*.

Qifan Zhang, Yunhui Guo, and Yu Xiang. 2024. Continual distillation learning: Knowledge distillation in prompt-based continual learning. *Preprint*, arXiv:2407.13911.

Heng Zheng, Yuling Shi, Xiaodong Gu, Haochen You, Zijian Zhang, Lubin Gan, Hao Zhang, Wenjun Huang, and Jin Huang. 2025a. Graphgeo: Multi-agent debate framework for visual geo-localization with heterogeneous graph neural networks. *arXiv preprint arXiv:2511.00908*.

Heng Zheng, Haochen You, Zijun Liu, Zijian Zhang, Lubin Gan, Hao Zhang, Wenjun Huang, and Jin Huang. 2025b. G2rammar: Bilingual grammar modeling for enhanced text-attributed graph learning. *arXiv preprint arXiv:2511.00911*.

Yuxiao Zhou, Zhen Wang, Yujun Li, Sheng Wang, Zhiyuan Liu, and Maosong Sun. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. *arXiv preprint arXiv:2302.06557*.

Yuxiao Zhou, Zhen Wang, Yujun Li, Sheng Wang, Zhiyuan Liu, and Maosong Sun. 2024. Framequant: Flexible low-bit quantization for transformers. *arXiv preprint arXiv:2402.06557*.

B. Zhu and 1 others. 2023. Large language models: Progress and applications. *Advances in NLP*.

A Comparison Experiments on Qwen2.5-7B

We also conducted experiments on Qwen2.5-7B across multiple datasets. As shown in Table 5, our method consistently achieves strong performance, demonstrating the effectiveness and general applicability of our pruning approach.

B Ablation of Multi-Domain Calibration on Qwen2.5-32B

We evaluate multi domain calibration on Qwen2.5-32B under 25% and 50% pruning, comparing single-domain calibration with our hybrid approach. As shown in Tables 4, the hybrid setting consistently outperforms the single-domain variant, achieving higher zero-shot accuracy on average. These results confirm that multi-domain calibration provides more robust channel importance estimation and improves structured pruning performance.

Method	Pruning Ratio	ARC-c	ARC-e	HellaSwag	OBQA	PIQA	Winogrande	Average
Qwen2.5-32B	0%	53.41	80.51	64.91	34.2	81.88	75.3	65.04
Ours	25%	46.08	74.87	53.35	**30.6**	75.35	**73.32**	58.93
Ours(w_mix)		**46.67**	**75.8**	**57.0**	29.6	**78.45**	72.85	**60.06**
Ours	50%	29.01	57.28	36.89	**23.6**	65.18	58.88	45.14
Ours(w_mix)		**30.72**	57.28	**39.44**	20.2	**70.84**	**61.4**	**46.65**

Table 4: Performance Comparsion of the compressed Qwen2.5-32B with and without multi-domain hybrid calibration set. Bold results highlight the best performance.

Method	Pruning Ratio	ARC-c	ARC-e	HellaSwag	OBQA	PIQA	Winogrande	Average
Qwen2.5-7 B	0%	47.61	80.47	59.95	33.8	78.56	72.85	62.21
Wanda-sp(w_mix)	25%	33.62	63.22	**43.45**	23.8	**73.23**	54.06	48.56
FLAP(w_mix)		32.08	62.33	41.75	21.4	72.31	59.59	48.24
Ours(w_mix)		**34.04**	**65.45**	43.12	**24.6**	72.85	**60.54**	**50.1**
Wanda-sp(w_mix)	50%	**21.67**	25.59	25.64	**14.6**	51.85	**51.78**	31.85
FLAP(w_mix)		19.37	29.97	27.17	12.2	56.09	49.01	32.3
Our method(w_mix)		18.86	**35.4**	**29.35**	12.4	**60.77**	50.2	**34.49**

Table 5: Zero-shot performance of the compressed Qwen2.5-7B. Bold results highlight the best performance.