Task-aware Block Pruning with Output Distribution Signals for Large Language Models

Anonymous ACL submission

Abstract

Large language models provide excellent performance, but their practical deployment is limited by significant inference costs. While block pruning effectively reduces latency with structural coherence, existing methods typically rely on representation similarity or costly sensitivity analyses, neglecting task-specific model behavior. This paper introduces an output-driven pruning method leveraging entropy-based estimations of output distributions to accurately identify less important model blocks. Extensive experiments validate the proposed method's effectiveness, demonstrating substantial efficiency gains without compromising downstream task performance.

1 Introduction

Recent advances in large language models (LLMs) have demonstrated remarkable performance across diverse natural language tasks (Dubey et al., 2024; Jiang et al., 2023). However, their growing size and inference cost have raised practical concerns, especially in deployment scenarios in resource-constrained environments. Model compression techniques, including pruning, quantization, and knowledge distillation (KD), have emerged as critical tools for addressing this efficiency-performance trade-off, and are often compatible with one another to yield additive benefits (Han et al., 2016; Mirashi et al., 2024; Zafrir et al., 2021; Kurtic et al., 2022; Zeng et al., 2024; Muralidharan et al., 2024; Song et al., 2024). Among them, pruning is especially attractive when combined with recovery aids such as fine-tuning or KD[1] as it inherit the strengths of larger models, avoiding costly neural architecture search (Frankle and Carbin, 2019; Chen et al., 2020; Zhang et al., 2021; Sarah et al., 2024; Bercovich et al., 2025). In particular, block pruning offers structurally coherent reductions that translate well to real-world latency improvement with stability (Zhong et al., 2025). However, prior methods often rely on representational similarity or performance sensitivity based on exhaustive removal experiemnts, which may overlook task-relevant internal behavior. This limitation motivates the explicit output-driven perspective which will be explored in this study.

While pruning has emerged as a powerful tool to improve efficiency, its practical effectiveness depends heavily on its pattern. Unstructured pruning is difficult to exploit efficiently without specialized hardware (Kim et al., 2018; Chen et al., 2019), while width and layer-level depth pruning often suffer from structural imbalance or instability, resulting in minimal efficiency gains or severe performance degradation (Kim et al., 2024; Lele et al., 2025; Xia et al., 2024; Zhang et al., 2024; He et al., 2024; Park et al., 2025). As a result, block pruning has been gaining growing interest, which achieves more proportional latency reductions with respect to compression ratio (Kim et al., 2024; Song et al., 2024; Zhong et al., 2025).

This paper investigates block pruning for downstream tasks, which are challenging and require explicit reasoning with distinct objectives, unlike general language modeling (Bachmann and Nagarajan, 2024). Most existing methods, however, focus on language modeling and rely solely on perplexity-based sensitivity measures for importance estimation (Kim et al., 2024; Song et al., 2024), which fails to reflect downstream performance (Liu et al., 2023; Hu et al., 2024; Zeng et al., 2025). Thus, the proposed method, inspired by the concept of entropy estimation (EE) (Liu et al., 2020; Hu et al., 2023) with promising performance in capturing internal model behavior, leverages output distributions to estimate block importance. The main contributions are as follows:

- We empirically demonstrate that analyzing

[1] E.g., Llama-3.2-1B (Dubey et al., 2024)

the output distribution, rather than relying on similarity-based measures at the representation level (Yang et al., 2025b) or computationally expensive sensitivity metrics, provides a more effective means of capturing internal model behavior, akin to EE and uncertainty quantification (Nostalgebraist, 2020; Kao et al., 2021; Chuang et al., 2024).

- Six effective criteria options and two metrics for characterizing output distributions are proposed through extensive experiments, enabling efficient pruning via constrained search within a limited candidate space.

- The necessity of observing model behavior on specific tasks is verified through experiments, highlighting the critical role of task-oriented importance estimation.

2 Related Work

Block pruning criteria include sensitivity on perplexity (PPL), representation entropy, and angular distance, with either one-shot or iterative removal (Kim et al., 2024; Song et al., 2024; Yang et al., 2025b; Gromov et al., 2025). To preserve performance, some approaches substitute or duplicate blocks supported by normalization, low-rank adaptation (LoRA), and approximation (Razzhigaev et al., 2024; Mikaelyan et al., 2025; Smith et al., 2025).

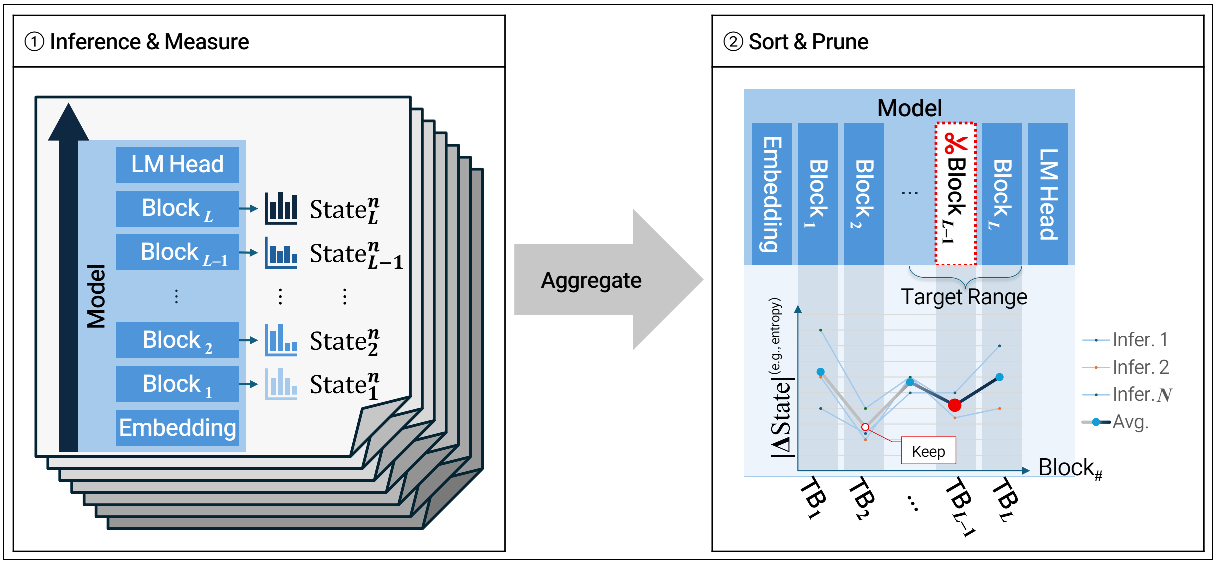

Figure 1: To estimate the importance of each block, states are calculated for N samples based on the token probability distribution output from every block, like EE and LogitLens (Kao et al., 2021; Nostalgebraist, 2020). Then, they are compared to the nearest previous blocks and aggregated block-wisely via penalty counting (L_0) and averaging (L_1). The graph on the right depicts L_1. By sorting blocks with the scores, relatively less important blocks are pruned.

3 Methodology

Inspired by EE, the proposing method hypothesizes that internal states such as entropy on the intermediate output distribution can reliably indicate block-wise capability, thereby avoiding costly ablation-based experiments. To identify a robust and effective criterion of block-level capability, comprehensive experiments explore a candidate space in Section 5.1, systematically evaluating various alternatives beyond entropy that is more suitable for capturing models' internal certainty or representational maturity at each block.

The numerical values derived from criterion are qualitative indicators that quantify the desirability of models' internal state or behavior via comparison. For example, entropy over the prediction score distribution is preferred to be low, meaning that model is more certain or decisive on its output. Building on this aspect, one strategy to identify the blocks to prune is to penalize those that negatively affect the reasoning process, by considering the direction of change in the value, *i.e.*, whether it increases or decreases compared to the previous block. This forms the basis of the first metric L_0, which assesses on whether an value suggests positive or negative contribution.

However, the reasoning for L_0 metric deviates substantially from conventional approaches that assess the absolute magnitude of contribution, irrespective of the direction of change. In other words,

Method	ARC-E	ARC-C	BoolQ	COPA	HeSw.	PIQA	Wino.	Avg.
Original Model	80.09	53.33	81.44	89.00	79.17	79.71	72.85	76.51
Prune 2 blocks (pruning ratio = 6.25%)								
SLEB	76.60	48.29	74.80	87.00	74.33	77.80	70.48	72.76
Short'dLLaMA	76.05	49.23	75.66	**90.00**	74.97	77.86	70.88	73.52
EntroDrop	67.55	44.45	63.09	84.00	68.58	75.57	71.43	67.81
JointLayerDrop	62.58	42.75	62.45	78.00	57.89	70.89	62.04	62.37
$L_0^{(\mathrm{ENT})}$	**78.07**	**49.91**	**78.20**	**90.00**	**77.21**	**78.78**	**73.09**	**75.04**
Prune 4 blocks (pruning ratio = 12.5%)								
SLEB	**72.47**	43.26	**66.12**	**87.00**	70.69	**76.12**	69.46	69.30
Short'dLLaMA	70.16	38.65	57.25	**87.00**	66.83	75.63	61.72	65.32
EntroDrop	40.78	33.53	58.59	67.00	37.43	61.86	58.48	51.10
JointLayerDrop	56.73	41.13	62.39	74.00	51.71	68.88	61.96	59.54
$L_0^{(\mathrm{ENT})}$	72.31	**45.22**	65.20	82.00	**73.48**	74.54	**73.01**	**69.39**
Prune 8 blocks (pruning ratio = 25%)								
SLEB	61.78	31.74	42.11	**77.00**	57.98	71.82	53.59	56.57
Short'dLLaMA	**61.83**	31.66	42.05	**77.00**	57.94	**71.87**	53.83	56.60
EntroDrop	33.33	29.78	62.11	66.00	26.97	57.67	56.59	47.49
JointLayerDrop	50.51	**37.97**	62.54	68.00	46.72	65.23	61.40	56.05
$L_0^{(\mathrm{ENT})}$	55.81	37.20	**76.48**	73.00	**58.58**	68.82	**70.17**	**62.87**

Table 1: Zero-shot accuracy (%) of LLaMA 3-8B on each task after pruning. Best and the second best results are indicated as bold and underline, respectively.

even an increase in entropy may imply that the block exerted influence within the broader context. Under this macro perspective, the alternative metric L_1 is derived to reflect the magnitude of deviation rather than the sign of change. For each block, it computes the average absolute change across all N samples, as depicted in Figure 1. Blocks with smaller L_1 scores are considered to contribute minimally to overall variation and are thus selected for pruning. Both norm-based metrics are used to aggregate and compare states across blocks to determine pruning candidates that contribute negatively or minimally onto the process. Table 1 reports only the best results obtained with L_0, while secondary-best results are reported in Section 2.

To clearly capture models' behavior, multiple-choice question answering (MCQA) is adopted for a block importance estimation task. Comparing to the general text generation, MCQA differs significantly in their cognitive and behavioral demands placed on models. Particularly, MCQA is a setting focused on restricted options, while in open-ended text generation tasks, entropy is computed over the full vocabulary space[2], which introduces considerable noise and interpretability challenges, *i.e.*, high entropy may not reliably indicate genuine uncertainty, as it can be inflated by semantically insignificant tokens such as function words or punctuation. A more detailed comparison between MCQA and generation settings is discussed in Section 5.1.

By constraining models' output space to a small set of discrete given options, MCQA promotes interpretability and allows for clearer attribution of internal state changes to task-relevant decision points. This task-constrained setting not only improves signal clarity but also strengthens the validity of entropy observation regarding block-wise capability and redundancy. Mitigation of confounding factors in free-form generation is further supported by auxiliary heuristics, such as restricting token outputs to a fixed choice set (e.g., A, B, C, D) and preventing pruning of a few early blocks empirically shown to be important, as implemented in the other works (Kim et al., 2024; Song et al., 2024).

4 Experiments

We conducted experiments ARC-Easy and employed pretrained open-source LLMs with 32 transformer blocks, including LLaMA 3-8B and Mistral-7B (Clark et al., 2018; Grattafiori et al., 2024; Jiang et al., 2023). Section 5 reports results on LLaMA 3-8B, while Section C presents results on Mistral-7B. For MCQA tasks with 1024 samples from the ARC-Easy training set, inference was performed with logits processors from `transformers` library to strictly force model to decode only the provided answer key options (Hugging Face, 2025). Block-level pruning was applyed with two norm-based metrics (L_0, L_1), setting the number of blocks to prune at most 8 (*i.e.*, pruning ratio = 25%).

[2]For instance, LLaMA-3 8B has vocab size of 128,256.

Evaluation was implemented on the following datasets: ARC Easy & Challenge (ARC-E & ARC-C) (Clark et al., 2018), BoolQ (Clark et al., 2019), COPA (Gordon et al., 2012), HellaSwag (HeSw.) (Zellers et al., 2019), PIQA (Bisk et al., 2020), and WinoGrande (Wino.) (Sakaguchi et al., 2021). As baselines, ShortenedLLaMA, SLEB, and EntroDrop implement block pruning, whereas JointLayerDrop conducts layer-wise pruning (Kim et al., 2024; Song et al., 2024; Yang et al., 2025b; He et al., 2024).

5 Results

Three baselines and the proposed method L_0 with entropy are evaluated under equal pruning ratios across multiple benchmarks. Results for pruning 2, 4, and 8 blocks (corresponding to 6.25%, 12.5%, and 25% pruning ratios, respectively) are reported in Table 1. At 6.25% pruning ratio, $L_0^{(\mathrm{ENT})}$ consistently yields top scores across all downstream tasks, while JointLayerDrop underperforms all block-based methods on most tasks. At 12.5% pruning ratio, $L_0^{(\mathrm{ENT})}$ maintains its lead with an average score of 69.39%, slightly higher than SLEB and considerably above ShortenedLLaMA and JointLayerDrop, with notable strength on WinoGrande and HeSwag.

When pruning 8 blocks which is three forth of 32 blocks, $L_0^{(\mathrm{ENT})}$ exhibits strong robustness, achieving 62.87% on average. Particularly high accuracy is retained on BoolQ (76.48%) and WinoGrande (70.17%), while competing methods experience significant degradation. JointLayerDrop yields the lowest overall performance (37.82%), indicating instability under aggressive sparsity. These results demonstrate the effectiveness of output distribution-based pruning in preserving accuracy across diverse tasks, especially under high pruning ratios.

5.1 Ablation Study

Criteria candidate space. To identify a reliable criterion, various options in candidate space were considered, such as confidence score (*i.e.*, the maximum prob. in the distribution) (Valade, 2024; Yang et al., 2025a), the gap between the top-1 and top-2 probs. (Schuster et al., 2022; Valade, 2024), and the entropy over all token probs. in the dist. (Xin et al., 2020; Liu et al., 2020; Hu et al., 2023; Valade, 2024). These indicators differ in scope: a single, a pair, or a full set of probabilities in a distribution, or a pair of distributions. Details of the candidate space are summarized in Table 3, and Table 4 reports performance in terms of the average accuracy across tasks and the AUC of the accuracy curve over pruning ratios from 0% to 25%.

Method		Avg. Acc.@8	AUC@8
L_0	MCQA	**62.87**	**561.73**
	TG	58.02	<u>552.51</u>
L_1	MCQA	58.92	536.90
	TG	<u>59.59</u>	552.19
SLEB		56.57	536.24
Short'dLLaMA		56.60	530.61
EntroDrop		45.68	000.00
JointLayerDrop		56.05	000.00

Table 2: Average zero-shot accuracy (%) of LLaMA 3-8B after pruning eight blocks and AUC over pruning ratios from 0% to 25%, across different option settings. As a state criterion, entropy was most effective for MCQA, while gap performed best for text generation (TG).

Importance of task setting. To highlight the importance of restricting the lens scope, task-level comparison is conducted between multiple-choice QA (MCQA) and text generation (TG), which differ fundamentally in output characteristics: MCQA involves a constrained, discrete choice space with known answers, while TG requires open-ended generation with higher entropy and variance. As shown in Table 2, the best-performing state criterion varies by task, yet those listed outperform all baselines. This finding supports both the use of task-specific observation datasets for criterion search, rather than relying solely on general language modeling, and the importance of leveraging output distributions over internal representations, to avoid overlooking task-dependent redundancy or reasoning signals.

6 Conclusion

This work introduces a simple yet effective method for block pruning in LLMs using output distribution signals, such as entropy and probability gaps. The approach eliminates reliance on language modeling loss and exhaustive sensitivity analysis and enables pruning aligned with task-relevant internal behavior. Extensive experiments demonstrate improved performance preservation over baselines and robustness under high sparsity. The results highlight the importance of task-specific observation and output-level signals for effective and interpretable pruning.

Limitations

In this paper, output distributions at intermediate blocks are derived using the LM head attached to the final block. While this is a naïve practice for probing internal behavior like LogitLens (Nostalgebraist, 2020), a more precise analysis would be available by attaching new LM heads to each block and training them accordingly (Schuster et al., 2022; Chuang et al., 2024). Such block-specific supervision may yield more faithful representations of each block's behavior. Additionally, the generalizability of the proposed method across architectures and scales requires further validation.

References

Gregor Bachmann and Vaishnavh Nagarajan. 2024. The pitfalls of next-token prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 2296–2318. PMLR.

Akhiad Bercovich, Tomer Ronen, Talor Abramovich, Nir Ailon, Nave Assaf, Mohammad Dabbah, Ido Galil, Amnon Geifman, Yonatan Geifman, Izhak Golan, Netanel Haber, Ehud Karpas, Roi Koren, Itay Levy, Pavlo Molchanov, Shahar Mor, Zach Moshe, Najeeb Nabwani, Omri Puny, and 7 others. 2025. Puzzle: Distillation-based nas for inference-optimized llms. In *Proceedings of the 42nd International Conference on Machine Learning*, volume NNN of *Proceedings of Machine Learning Research*, pages NNN–NNN. PMLR.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pre-trained BERT networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yu-Hsin Chen, Tien-Ju Yang, Joel Emer, and Vivienne Sze. 2019. Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9:292–308.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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.

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

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Dan Roberts. 2025. The unreasonable ineffectiveness of the deeper layers. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Song Han, Huizi Mao, and William J. Dally. 2016. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Shwai He, Guoheng Sun, Zheyu Shen, and Ang Li. 2024. What matters in transformers? not all attention is needed. *Preprint*, arXiv:2406.15786.

Boren Hu, Yun Zhu, Jiacheng Li, and Siliang Tang. 2023. Smartbert: A promotion of dynamic early exiting mechanism for accelerating BERT inference. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 5067–5075. ijcai.org.

Yutong Hu, Quzhe Huang, Mingxu Tao, Chen Zhang, and Yansong Feng. 2024. Can perplexity reflect large language model's ability in long text understanding? In *The Second Tiny Papers Track at ICLR 2024*.

Hugging Face. 2025. Utilities for generation. `https://huggingface.co/docs/transformers/v4.53.2/en/internal/generation_utils#transformers.LogitsProcessor`. Accessed: 2025-07-16.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Wei-Tsung Kao, Tsung-Han Wu, Po-Han Chi, Chun-Cheng Hsieh, and Hung-Yi Lee. 2021. Bert's output layer recognizes all hidden layers? some intriguing phenomena and a simple way to boost bert. *Preprint*, arXiv:2001.09309.

Bo-Kyeong Kim, Geonmin Kim, Tae-Ho Kim, Thibault Castells, Shinkook Choi, Junho Shin, and Hyoung-Kyu Song. 2024. Shortened LLaMA: A simple depth pruning for large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Dongyoung Kim, Junwhan Ahn, and Sungjoo Yoo. 2018. Zena: Zero-aware neural network accelerator. *IEEE Design & Test*, 35:39–46.

Eldar Kurtic, Daniel Campos, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin, and Dan Alistarh. 2022. The optimal BERT surgeon: Scalable and accurate second-order pruning for large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4163–4181. Association for Computational Linguistics.

Nahush Lele, Arnav Chavan, Aryamaan Thakur, and Deepak Gupta. 2025. Rethinking the value of training-free structured pruning of LLMs. *Transactions on Machine Learning Research*.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. 2023. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22188–22214. PMLR.

Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. 2020. Fastbert: a self-distilling BERT with adaptive inference time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6035–6044. Association for Computational Linguistics.

Liana Mikaelyan, Ayyoob Imani, Mathew Salvaris, Parth Pathak, and Mohsen Fayyaz. 2025. Deltallm: Compress llms with low-rank deltas between shared weights. *Preprint*, arXiv:2501.18596.

Aishwarya Mirashi, Purva Lingayat, Srushti Sonavane, Tejas Padhiyar, Raviraj Joshi, and Geetanjali Kale. 2024. On importance of pruning and distillation for efficient low resource NLP. *CoRR*, abs/2409.14162.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. 2024. Compact language models via pruning and knowledge distillation. In *Advances in Neural Information Processing Systems*, volume 37, pages 41076–41102. Curran Associates, Inc.

Nostalgebraist. 2020. interpreting gpt: the logit lens. *LessWrong*.

Seungcheol Park, Sojin Lee, Jongjin Kim, Jinsik Lee, Hyunjik Jo, and U Kang. 2025. Accurate sublayer pruning for large language models by exploiting latency and tunability information. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages NNN5272–NNN5280. ijcai.org.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Nikolai Gerasimenko, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. 2024. Your transformer is secretly linear. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5376–5384, Bangkok, Thailand. Association for Computational Linguistics.

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

Anthony Sarah, Sharath Nittur Sridhar, Maciej Szankin, and Sairam Sundaresan. 2024. Llama-nas: Efficient neural architecture search for large language models. In *Computer Vision - ECCV 2024 Workshops - Milan, Italy, September 29-October 4, 2024, Proceedings, Part XI*, volume 15633 of *Lecture Notes in Computer Science*, pages 67–74. Springer.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

James Seale Smith, Chi-Heng Lin, Shikhar Tuli, Haris Jeelani, Shangqian Gao, Yilin Shen, Hongxia Jin, and Yen-Chang Hsu. 2025. FlexiGPT: Pruning and extending large language models with low-rank weight sharing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 718–730, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiwon Song, Kyungseok Oh, Taesu Kim, Hyungjun Kim, Yulhwa Kim, and Jae-Joon Kim. 2024. SLEB: Streamlining LLMs through redundancy verification and elimination of transformer blocks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46136–46155. PMLR.

Florian Valade. 2024. Accelerating large language model inference with self-supervised early exits. *Preprint*, arXiv:2407.21082.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2024. Sheared LLaMA: Accelerating language model pre-training via structured pruning. In *The Twelfth International Conference on Learning Representations*.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. 2020. Deebert: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2246–2251. Association for Computational Linguistics.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Zheng Lin, Li Cao, and Weiping Wang. 2025a. Dynamic early exit in reasoning models. *Preprint*, arXiv:2504.15895.

Liangwei Yang, Yuhui Xu, Juntao Tan, Doyen Sahoo, Silvio Savarese, Caiming Xiong, Huan Wang, and Shelby Heinecke. 2025b. Entropy-based block pruning for efficient large language models. *Preprint*, arXiv:2504.03794.

Ofir Zafrir, Ariel Larey, Guy Boudoukh, Haihao Shen, and Moshe Wasserblat. 2021. Prune once for all: Sparse pre-trained language models. In *NeurIPS 2021 Workshop on Efficient Natural Language and Speech Processing (ENLSP)*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Chao Zeng, Songwei Liu, Shu Yang, Fangmin Chen, Xing Mei, and Lean Fu. 2024. GQSA: group quantization and sparsity for accelerating large language model inference. *CoRR*, abs/2412.17560.

Hansi Zeng, Kai Hui, Honglei Zhuang, Zhen Qin, Zhenrui Yue, Hamed Zamani, and Dana Alon. 2025. Can pre-training indicators reliably predict fine-tuning outcomes of llms? *Preprint*, arXiv:2504.12491.

Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. 2021. Why lottery ticket wins? A theoretical perspective of sample complexity on sparse neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 2707–2720.

Yang Zhang, Yawei Li, Xinpeng Wang, Qianli Shen, Barbara Plank, Bernd Bischl, Mina Rezaei, and Kenji Kawaguchi. 2024. Finercut: Finer-grained interpretable layer pruning for large language models. In *Workshop on Machine Learning and Compression, NeurIPS 2024*.

Longguang Zhong, Fanqi Wan, Ruijun Chen, Xiaojun Quan, and Liangzhi Li. 2025. Blockpruner: Fine-grained pruning for large language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria and virtual meeting, July 27–August 1, 2025*, pages NNN1–NNN2. Association for Computational Linguistics.

A Experiment Details

All experiments are implemented in PyTorch using Huggingface Transformers on four Nvidia Titan Xp GPUs, each with 12 GB of memory. Evaluation was executed with `EleutherAI/lm-evaluation-harness` library.

For text generation in ablation study, models' behavior was observed using Wikitext-2. Since no answer key is specified in text generation, the next token was treated as the target and inference proceeded without logits processors. States were observed over the first 1024 tokens of Wikitext-2.

Although ShortenedLLaMA also introduced a Taylor-based metric, only PPL sensitivity was adopted in comparison experiments due to the high computational overhead of gradient calculation.

Abbreviation	Definition
CONF	Confidence score (top-1 probability)
K	Probability assigned to the answer key
G	Gap between top and second probs.
ENT	Entropy of the output token probability distribution
CE	Cross-entropy between output distributions of the final and intermediate blocks
KLD	KL divergence between output distributions of the final and intermediate blocks

Table 3: List of state criteria options.

	SLEB	**StLm.**	**EntDr.**	**JLD**	$L_0^{(\mathrm{CONF})}$	$L_0^{(\mathrm{K})}$	$L_0^{(\mathrm{G})}$	$L_0^{(\mathrm{ENT})}$	$L_0^{(\mathrm{CE})}$	$L_0^{(\mathrm{KLD})}$
Avg. Acc.@2	72.76	<u>73.52</u>	64.68	62.37	70.12	70.12	70.12	**75.04**	70.38	70.12
Avg. Acc.@4	<u>69.30</u>	65.32	49.71	59.54	68.40	68.40	67.72	**69.39**	69.11	68.48
Avg. Acc.@8	56.57	56.60	45.68	56.05	**63.94**	**63.94**	60.98	<u>62.87</u>	61.95	**63.94**
AUC@8	536.24	530.61	000.00	000.00	546.53	546.53	539.33	**561.73**	546.39	<u>546.61</u>

Table 4: Average of zero-shot accuracy (%) of LLaMA 3-8B after pruning eight blocks and AUC over pruning ratio from 0% to 25%. Best results are in bold; second-best results are underlined. StLm., EntDr., and JLD are the abbreviation of ShortenedLLaMA, EntroDrop, and JointLayerDrop, respectively. The L_0 metrics are defined in Section 3.

B State Criteria

Table 3 summarizes the candidate state criteria used to assess block-wise importance. Each criterion captures models' internal behavior based on token probability distributions and is applied via norm-based aggregation metrics, L_0 and L_1. Ablation results in Table 4 report average accuracy after pruning LLaMA 3-8B and the area under the curve (AUC) across pruning ratios from 0% to 25%. Among all candidates, entropy (ENT) consistently demonstrates strong performance across tasks, particularly in the MCQA setting, highlighting its reliability as a state indicator. Other criteria such as confidence score (CONF), gap (G), cross-entropy (CE), and KL divergence (KLD) yield comparable yet slightly lower accuracy. Based on these findings, entropy is adopted as the default criterion in the main experiments, while other options remain viable depending on the application context.

C Mistral-7B Results

Table 5 presents performance after pruning on Mistral-7B across multiple downstream tasks. Each method prunes the same number of blocks, enabling direct comparison under equal compression ratios. These results confirm the effectiveness of output-distribution-based pruning in retaining task-relevant capacity, especially under constrained pruning budgets.

Method	ARC-E	ARC-C	BoolQ	COPA	HeSw.	PIQA	Wino.	Avg.
Original Model								
Prune 2 blocks (pruning ratio = 6.25%)								
SLEB	76.22	46.67	77.71	88.00	77.30	80.03	67.96	73.41
Short'dLLaMA	74.87	44.62	74.53	88.00	73.88	42.20	78.62	72.02
EntroDrop								
JointLayerDrop								
Prune 4 blocks (pruning ratio = 12.5%)								
SLEB	71.68	41.98	74.22	87.00	72.83	77.58	64.96	70.04
Short'dLLaMA	66.84	34.30	59.51	88.00	60.68	75.41	56.43	63.02
EntroDrop	29.97	29.35	32.94	71.00	33.49	56.20	61.09	44.86
JointLayerDrop	49.83	36.09	62.51	66.00	46.18	64.47	62.75	55.40
Prune 8 blocks (pruning ratio = 25%)								
SLEB	63.59	35.75	59.54	79.00	62.95	72.47	60.85	62.02
Short'dLLaMA	51.73	26.28	54.89	70.00	48.28	69.26	52.17	53.23
EntroDrop	26.77	29.18	42.48	72.00	28.25	56.15	55.80	44.38
JointLayerDrop	49.20	36.77	62.45	66.00	45.37	63.93	62.51	55.18

Table 5: Zero-shot accuracy (%) of Mistral-7B on each task after pruning. Best results are in bold; second-best results are underlined.