WINA: Weight Informed Neuron Activation for Accelerating Large Language Model Inference

Anonymous Authors[1]

Abstract

The growing computational demands of large language models (LLMs) make efficient inference and activation strategies increasingly critical. While recent approaches, such as Mixture-of-Experts (MoE), leverage selective activation but require specialized training, training-free sparse activation methods offer broader applicability and superior resource efficiency through their plug-and-play design. However, many existing methods rely solely on hidden state magnitudes to determine activation, resulting in high approximation errors and suboptimal inference accuracy. To address these limitations, we propose WINA (**W**eight **I**nformed **N**euron **A**ctivation), a novel, simple, and training-free sparse activation framework that jointly considers hidden state magnitudes and the column-wise ℓ_2-norms of weight matrices. We show that this leads to a sparsification strategy that obtains optimal approximation error bounds with theoretical guarantees tighter than existing techniques. Empirically, WINA also outperforms state-of-the-art methods (*e.g.*, TEAL) by up to 2.94% in average performance at the same sparsity levels, across a diverse set of LLM architectures and datasets. These results position WINA as a new performance frontier for training-free sparse activation in LLM inference, advancing training-free sparse activation methods and setting a robust baseline for efficient inference.

1. Introduction

While large language models (LLMs) have revolutionized the field of natural language processing, they often require substantial computational resources, particularly during inference, making reducing inference costs without degrading output quality a central challenge.

One strategy has been to activate only a sub-network of the full model (Jacobs et al., 1991) during inference using a Mixture of Experts (MoE) architecture, which has already seen adoption in popular and widely-used LLMs like GPT4 (Achiam et al., 2023) and Mistral (Jiang et al., 2023). Other methods include model distillation, where a smaller model is trained using knowledge distilled from a larger teacher model to route inference requests more efficiently. However, these approaches require a considerable amount of training, which can also be computationally costly.

An alternative is training-free sparse activation, which retains the original dense model while selectively deactivating components during inference, using criteria like hidden-state magnitudes, weight importance, or validation data to determine deactivation targets.

However, current training-free methods exhibit critical limitations. Most notably, these approaches fail to account for how interactions between inputs and the weight matrix during forward propagation affect model outputs, leading to accumulated approximation errors in sparse activation.

Contributions. In this paper, we propose WINA: a simple, easy-to-use, training-free framework that performs sparse activation based on the magnitude of hidden states and the column-wise ℓ_2-norm of the weight matrix. In summary, our detailed contributions include:

- **Weighted-informed Activation:** we introduce a novel sparse activation method that jointly considers hidden state magnitudes and the column-wise ℓ_2-norms of weight matrices, leading to a more informed construction of a sub-network during inference.
- **Theoretically Tighter Approximation Error:** we conduct a analysis to demonstrate that our weight-informed activation mechanism yields a lower expected output error compared to prior methods under mild assumptions.
- **Numerical Experiments:** The extensive evaluations performed on multiple LLMs varying from 7B to 14B demonstrate our method's superior accuracy.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Table 1. Feature comparison between WINA and existing methods (Liu et al., 2024; Lee et al., 2024).

	WINA	TEAL	CATS
Tight Approx Error	✓	✗	✗
Layer Generality	✓	✓	✗
Hetero Sparsity	✓	✓	✗

2. Related Work

Modern dynamic expert selection approaches fall into two principal paradigms: training-based methods and training-free methods. Training-based methods typically employ a trainable router to dynamically select activated experts for each token, with the Mixture-of-Experts (MoE) architecture (Jacobs et al., 1991) serving as the foundational framework. In this framework, each expert operates an individual component of the model, as only the relevant experts are activated for each input during inference.

Training-free methods, in contrast, do not rely on a learnable router, instead using predefined or calculated criteria to perform sparse activation. Q-Sparse (Wang et al., 2024) produces sparsity as a function of input magnitudes, achieving sparsity rates of 60% with negligible performance degradation. CATS (Lee et al., 2024) applies sparse activation on SwiGLU outputs within gated MLP layers, achieving performance comparable to the original dense model while achieving 25% model sparsity. In contrast, TEAL (Liu et al., 2024) extends magnitude-based activation sparsity to all network layers, achieving 40-50% model-wide sparsity across architectures with minimal performance impact.

However, current sparse activation methods suffer from fundamental limitations. Most notably, they determine activation elements solely based on the magnitude of hidden states, neglecting the crucial influence of the weight matrix, which results in suboptimal error control.

3. Methodology

As illustrated in Figure 1, WINA jointly considers both the input tensor and the associated weight matrix, rather than relying solely on input magnitudes. During inference, it activates only the most influential neurons, effectively constructing a sparse sub-network that maintains the expressive power of the original model.

3.1. Problem Statement

Main Problem. Consider a deep neural network (DNN) $\mathcal{M}$ consisting of L layers. We denote the weight matrix of the l-th layer as $W_l \in \mathbb{R}^{m_l \times n_l}$ and the corresponding input as an arbitrary tensor $X \in \mathbb{R}^{n_l \times s_l}$ for $l \in \{1, \ldots, L\}$, representing

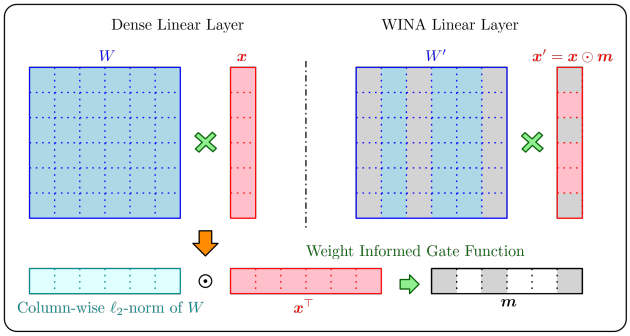


Figure 1. **Overview of WINA.** WINA performs training-free sparse activation by adaptively selecting the most influential input dimensions through joint consideration of both hidden state magnitudes and the corresponding column-wise ℓ_2-norms of weight.

the full information content. Our goal is to identify a set of binary activation gates $\mathcal{G} = \{\boldsymbol{g}_1, \cdots, \boldsymbol{g}_L\}$, where each $\boldsymbol{g}_l \in \{0,1\}^{n_l}$, such that the deviation between the model's original output and the gated output is minimized:

$$\underset{\boldsymbol{g}_1, \cdots, \boldsymbol{g}_L}{\text{minimize}} \quad \|\mathcal{M}(X) - \mathcal{M}(X \mid \mathcal{G})\|_2 . \tag{1}$$

Since obtaining the complete set of possible inputs X is generally infeasible, we instead use a sampled subset $\tilde{X}$ to approximate it. The activation gating operates in the input vector space to reduce output deviation. With this observation, we can reformulate the original problem into a per-layer version to make the problem more tractable.

Refined Problem. Given a weight matrix $W \in \mathbb{R}^{m \times n}$ and a sampled input vector $\boldsymbol{x} \in \mathbb{R}^n$, the standard linear transformation is $\boldsymbol{y} \leftarrow W\boldsymbol{x}$. Our objective then becomes identifying an activation gate $\boldsymbol{g} \in \{0,1\}^n$ such that the gated output $\boldsymbol{y}_{\boldsymbol{g}} \leftarrow W(\boldsymbol{g} \odot \boldsymbol{x})$ approximates the original by solving:

$$\underset{\boldsymbol{g} \in \{0,1\}^n}{\text{minimize}} \quad \|W\boldsymbol{x} - W(\boldsymbol{g} \odot \boldsymbol{x})\|_2 . \tag{2}$$

3.2. Weight Informed Gate Function

Motivation. Many current sparse activation methods (*e.g.*, Q-sparse (Wang et al., 2024), CATS (Lee et al., 2024), TEAL (Liu et al., 2024)) operate via a top-K gating mechanism governed by the absolute values of the hidden states:

$$\boldsymbol{g}_i = \begin{cases} 1 & \text{if } |x_i| \text{ is among the top-}K \text{ values in } |\boldsymbol{x}|, \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

However, this approach ignores the critical role that weight matrices play.

Formalization. In our proposed WINA framework, we systematically construct binary activation gates by selecting

the top-K components according to specific criteria:

$$\boldsymbol{g}_i = \begin{cases} 1 & \text{if } |x_i c_i| \text{ is among the top-}K \text{ values in } |\boldsymbol{x} \odot \boldsymbol{c}|, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\boldsymbol{c} \in \mathbb{R}^n$ represents the column-wise ℓ_2 norm of W and $\odot$ denotes the Hadamard or element-wise product.

The choice of K can be adapted to different use cases, ranging from (1) a coarse-grained universal criterion where a shared K is applied across all layers to (2) a fine-grained layer-specific strategy that assigns K individually to better minimize approximation error.

3.3. Theoretical Analysis

WINA also offers theoretical advantages, capable of achieving a more optimal bound on the approximation error than TEAL. To demonstrate this, our formal analysis begins with a fundamental **Lemma** C.1 that rigorously establishes WINA's advantage in single-layer networks under column-wise orthogonality constraints (see Appendix C for details).

Using our single-layer Building upon our single-layer Lemma C.1, we now generalize the theoretical framework to deep networks with L consecutive linear layers. As stated in Theorem 3.1 below, we see that WINA still achieves smaller approximation error than TEAL in the L layer case.

Theorem 3.1 (Optimal approximation error over consecutive L layer)**.** *Let $\boldsymbol{x} \in \mathbb{R}^{d_0}$ be an input vector and $\{W^{(\ell)}\}_{\ell=1}^N$ denote the weight matrices of an N-layer neural network, where each $W^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ is column-wise orthogonal, i.e., $(W^{(\ell)})^\top W^{(\ell)} = diag((c_1^{(\ell)})^2, (c_2^{(\ell)})^2, \dots (c_n^{(\ell)})^2)$ where $c_i^{(\ell)} = \left\| W_{:,i}^{(\ell)} \right\|$. For any target sparsity level $k \in \mathbb{N}^+$ with $k < \min_{\ell \in \{1,\dots,N\}} d_\ell$, the expected deviation satisfies:*

$$\mathbb{E}\left[\|\boldsymbol{y}_{WINA} - \boldsymbol{y}\|_2^2 \right] \le \mathbb{E}\left[\|\boldsymbol{y}_{TEAL} - \boldsymbol{y}\|_2^2 \right], \tag{5}$$

where $\boldsymbol{y}_{WINA}$ denotes the output produced by WINA; $\boldsymbol{y}_{TEAL}$ is the output of TEAL; and $\boldsymbol{y}$ is the original dense network output without any sparsification.

Proof. See Appendix D.

Using these, we now consider realistic deep neural networks equipped with various activation functions. Following our established methodology, we first present the fundamental **Lemma** E.1, which rigorously demonstrates that WINA maintains its tighter error approximation guarantee for single-layer networks equipped with common activation functions (*e.g.*, ReLU and its variants, sigmoidal functions, and softmax) under standard architectural constraints (complete analysis in Appendix E).

Finally, we extend this theorem to the case of a multi-layer network with activation activations.

Theorem 3.2 (Optimal approximation error over consecutive L layer with MIF)**.** *Let $\boldsymbol{x} \in \mathbb{R}^{d_0}$ be an input vector that follow a zero-mean symmetric distribution and $\{W^{(\ell)}\}_{\ell=1}^N$ denote the weight matrices of an N-layer neural network, where each $W^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ is column-wise orthogonal, i.e., $(W^{(\ell)})^\top W^{(\ell)} = diag((c_1^{(\ell)})^2, (c_2^{(\ell)})^2, \dots (c_n^{(\ell)})^2)$ where $c_i^{(\ell)} = \left\| W_{:,i}^{(\ell)} \right\|$. Let $f : \mathbb{R} \to \mathbb{R}$ be an activation function. For any target sparsity level $k \in \mathbb{N}^+$ with $k < \min_{\ell \in \{1,\dots,N\}} d_\ell$, the expected deviation satisfies:*

$$\mathbb{E}\left[\|\boldsymbol{y}_{WINA} - \boldsymbol{y}\|_2^2 \right] \le \mathbb{E}\left[\|\boldsymbol{y}_{TEAL} - \boldsymbol{y}\|_2^2 \right], \tag{6}$$

where $\boldsymbol{y}_{WINA}$ denotes the output produced by WINA; $\boldsymbol{y}_{TEAL}$ is the output of TEAL; and $\boldsymbol{y}$ is the original dense network output without any sparsification.

Proof. See Appendix F.

3.4. From Theory to Practice

Motivation. In Section 3.3, our theoretical analysis assumes zero-mean symmetric inputs (have been observed (Liu et al., 2024)) and the column-wise orthogonality of the weight matrices, which LLMs often violate in reality. To bridge this gap between theory and practice, we propose a tensor transformation framework that enforces column-wise orthogonality on some weights.

Transformation Protocol. We perform Singular Value Decomposition (SVD) on W:

$$W = U\Sigma V^\top$$

where U and V are orthogonal matrices, and Σ is a diagonal matrix containing the singular values of W. Then we transform W to $\hat{W}$ as follows:

$$\hat{W} = WV$$

This transformation guarantees that the resulting matrix W' satisfies the column-orthogonality:

$$(\hat{W})^\top \hat{W} = \Sigma^\top U^\top U \Sigma = \Sigma^2 \tag{7}$$

To maintain the model's final output unchanged after this transformation, we propagate these transformations through adjacent layers using computational invariance (Ashkboos et al., 2024), as formally derived in Appendix B).

4. Experiments

4.1. Experimental Setup

Models. We present comprehensive experimental results across four models: Qwen-2.5-7B (Dong et al., 2024),

Llama-2-7B (Touvron et al., 2023), Llama-3-8B (Dubey et al., 2024), and Phi-4-14B (Abdin et al., 2024).

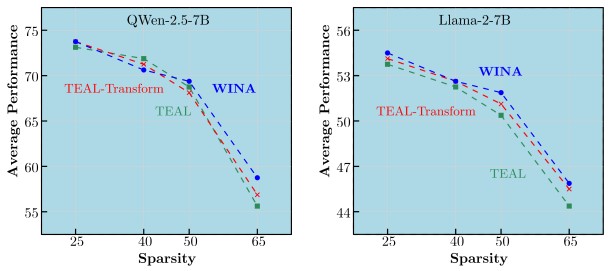
Evaluation. We follow TEAL's use of the lm-evaluation-harness pipeline (Gao et al., 2023) for our evaluations on an extensive suite of downstream tasks, including PIQA (Bisk et al., 2020), WinoGrande (Sakaguchi et al., 2019), HellaSwag (Zellers et al., 2019), Arc Challenge (Clark et al., 2018), MMLU (Hendrycks et al., 2020), and GSM8K (Cobbe et al., 2021).

Baselines. To eliminate the potential effect introduced by the transformation process, we introduce an additional baseline, TEAL-Transform. In this variant, the TEAL approach is applied to the transformed model, retaining the k elements with the largest absolute values $|x|$.

To further improve performance, we assign layer-specific sparsity ratios instead of a uniform sparsity across the model through the greedy algorithm proposed in TEAL.

4.2. Results.

Here, we provide an empirical comparison of WINA against TEAL-based baselines (e.g., TEAL and TEAL-transform) across different sparsity levels (25%, 40%, 50% and 65%) to demonstrate the effectiveness of our proposed algorithm under various experimental settings.



(a) QWen-2.5-7B (b) Llama-2-7B

(c) Llama-3-8B (d) Phi-4-14B

Figure 2. **Sparsity-performance frontiers.** Sparsity-performance across Qwen-2.5-7B, Llama-2-7B, Llama-3-8B, and Phi-4-14B.

As demonstrated in Figure 2 and Table 2, WINA consistently outperforms both TEAL and TEAL-transform baselines across multiple model architectures. Specifically, it achieves superior performance on Qwen-2.5-7B and Llama-2-7B for most sparsity levels, and maintains this advantage across all sparsity levels for Llama-3-8B and Phi-4-14B. Notably, as sparsity increases, the performance gap between WINA and the baselines becomes more pronounced. For instance, at 65% sparsity, WINA achieves improvements of 2.94% (vs TEAL) and 1.41% (vs TEAL-transform) on Qwen-2.5-7B, 1.61% (vs TEAL) and 0.31% (vs TEAL-transform) on Llama-2-7B, 2.41% (vs TEAL) and 0.49% (vs TEAL-transform) on Llama-3-8B, and 2.01% (vs TEAL) and 0.86% (vs TEAL-transform) on Phi-4-14B. This scaling behavior demonstrates WINA's superior robustness.

Table 2. Results over different sparsity levels across Qwen-2.5-7B, Llama-2-7B, Llama-3-8B and Phi-4-14B.

Model	Method	Sparsity			
		0.25	0.4	0.5	0.65
Qwen-2.5 -7B	TEAL	72.83	**71.66**	68.93	55.40
	TEAL-transform	72.99	71.52	68.62	56.93
	WINA	**73.02**	71.38	**69.26**	**58.34**
Llama-2 -7B	TEAL	54.11	52.49	50.51	44.21
	TEAL-transform	54.19	**52.84**	51.40	45.51
	WINA	**54.42**	52.83	**51.76**	**45.82**
Llama-3 -8b	TEAL	65.58	62.65	58.51	45.36
	TEAL-transform	65.99	62.94	59.19	47.28
	WINA	**66.00**	**63.20**	**59.57**	**47.77**
Phi-4 -14B	TEAL	76.86	75.98	74.36	68.71
	TEAL-transform	76.56	76.50	74.74	69.86
	WINA	**77.57**	**76.71**	**75.91**	**70.72**

Furthermore, our comprehensive experimental results in Appendix A demonstrate that WINA achieves superior performance particularly strong performance such as GSM8K, ARC Challenge, and HellaSwag. Notably, the method maintains robust accuracy even under aggressive sparsity level, substantially outperforming baseline approaches.

5. Conclusion

In this paper, we introduce WINA, a training-free sparse activation framework that selects active neurons based on both hidden state magnitudes and the column-wise ℓ_2-norms of subsequent weight matrices.

Our theoretical analysis demonstrates that WINA achieves a tighter bound on approximation error compared to existing approaches, under mild assumptions. To bridge the gap between theoretical guarantees and practical deployment in pre-trained LLMs, we further adopted a tensor transformation protocol that enforces column-orthogonality in weight matrices without altering model output. Our extensive experiments across multiple LLM architectures and benchmarks also validate WINA's superior performance under controlled sparsity settings, establishing it as a new state-of-the-art in the domain of training-free sparse activation.

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ashkboos, S., Croci, M. L., do Nascimento, M. G., Hoefler, T., and Hensman, J. Slicegpt: Compress large language models by deleting rows and columns, 2024. URL https://arxiv.org/abs/2401.15024.

Bisk, Y., Zellers, R., Gao, J., Choi, Y., et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Dong, Y., Liu, Z., Xu, Y., Cui, Y., Che, W., Sun, T., and Liu, T. Qwen2: Scaling up language models with data mixture of expert quality, 2024. URL https://huggingface.co/Qwen/Qwen2-7B.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Lee, D., Lee, J.-Y., Zhang, G., Tiwari, M., and Mirhoseini, A. Cats: Contextually-aware thresholding for sparsity in large language models, 2024. URL https://arxiv.org/abs/2404.08763.

Liu, J., Ponnusamy, P., Cai, T., Guo, H., Kim, Y., and Athiwaratkun, B. Training-free activation sparsity in large language models, 2024. URL https://arxiv.org/abs/2408.14690.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

Touvron, H., Lavril, T., Izacard, G., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama 2: Open foundation and fine-tuned chat models, 2023.

Wang, H., Ma, S., Wang, R., and Wei, F. Q-sparse: All large language models can be fully sparsely-activated. *arXiv preprint arXiv:2407.10969*, 2024.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Table 3. Results of controlled sparsity experiments over Qwen-2.5-7B

Method	Sparsity	PiQA	WinoGrande	HellaSwag	Arc-c	MMLU	GSM8K	Avg
Baseline	-	79.71	72.85	78.93	51.11	71.93	83.32	72.98
TEAL (Liu et al., 2024)	25%	79.27	78.56	72.77	51.19	71.30	82.87	72.83
	40%	78.40	77.28	73.09	52.65	70.20	78.32	**71.66**
	50%	78.62	75.02	69.77	51.02	67.72	71.42	68.93
	65%	73.72	63.35	62.67	42.75	54.95	34.95	55.40
TEAL-transform	25%	80.09	72.77	78.65	51.79	71.56	83.09	72.99
	40%	79.71	72.30	77.73	51.28	69.93	77.18	71.52
	50%	78.56	68.67	75.74	50.00	67.28	71.49	68.62
	65%	76.06	61.33	67.30	44.20	56.06	32.60	56.93
WINA	25%	80.05	72.69	78.58	51.37	71.51	83.93	**73.02**
	40%	78.40	70.56	78.02	50.94	70.54	79.83	71.38
	50%	78.67	69.30	76.48	50.85	67.99	72.25	**69.26**
	65%	76.17	61.01	70.09	42.92	59.48	38.36	**58.34**

A. Results

Qwen-2.5-7B. We evaluate WINA on Qwen2.5-7B (Yang et al., 2024) across various sparsity levels (i.e, 25% – 65%) under the controlled sparsity setting. As shown in Table 3, WINA consistently matches or outperforms both TEAL and TEAL -transform across all sparsity levels. Notably, as sparsity increases, the performance gap between WINA and the baselines becomes more pronounced. For instance, at 65% sparsity, WINA outperforms TEAL by 2.94% and TEAL-transform by 1.41% on average. This trend indicates that WINA is more robust under high sparsity, likely due to its ability to retain the most influential activations by jointly considering hidden state magnitudes and weight norms. Particularly on harder tasks such as GSM8K and HellaSwag, WINA maintains relatively strong performance even when aggressive sparsification is applied.

Llama-2-7B. On Llama-2-7B, WINA again shows strong performance under various sparsity constraints. As shown in Table 4, WINA achieves the highest average accuracy at 25% sparsity, outperforming both TEAL-based baselines and the full model. While performance naturally degrades at the extreme 65% sparsity level, WINA still offers the best accuracy, suggesting its robustness under aggressive pruning.

Llama-3-8B. The results on Llama-3-8B further emphasize WINA's resilience to pruning, as summarized in Table 5. While TEAL slightly outperforms at the 25% level, WINA leads in all remaining sparsity configurations, culminating in +1.06% and +2.41% over TEAL at 50% sparsity and 65% sparsity, respectively. Notably, WINA sustains particularly strong performance on reasoning-intensive tasks like GSM8K and ARC Challenge, where other methods show significant drops under compression. These patterns suggest that WINA is not only compression-friendly but also capable of preserving complex decision-making abilities under tight computational budgets.

Phi-4-14B. WINA also delivers robust performance on Phi-4-14B across all tested sparsity levels, as detailed in Table 6. It consistently either matches or exceeds the accuracy of both TEAL and TEAL-transform, and achieves the top average score at every sparsity setting. At the highest sparsity of 65%, for instance, WINA improves upon TEAL and TEAL-transform by +2.01% and +0.86%, respectively. Its ability to retain high performance on complex benchmarks such as GSM8K and MMLU, even under severe pruning, highlights its stability. These outcomes demonstrate that WINA can effectively preserve key reasoning mechanisms in large-scale models, making it well-suited for sparsity-constrained deployments.

Table 4. Results of controlled sparsity experiments over Llama-2-7B

Method	Sparsity	PiQA	Arc-c	WinoGrande	HellaSwag	MMLU	GSM8K	Avg
Baseline	-	79.05	46.33	68.98	76.00	41.82	13.87	54.34
TEAL (Liu et al., 2024)	25%	78.18	45.99	69.85	76.01	41.30	13.34	54.11
	40%	77.53	44.45	67.88	75.32	38.66	11.07	52.49
	50%	77.53	41.21	67.25	73.57	34.71	8.79	50.51
	65%	74.43	33.87	62.12	64.20	27.05	3.56	44.21
TEAL-transform	25%	78.45	46.42	69.14	75.93	41.75	13.42	54.19
	40%	77.69	45.48	68.43	75.18	39.22	11.05	**52.84**
	50%	78.07	43.77	66.54	73.48	36.28	10.24	51.40
	65%	74.32	37.71	63.77	66.49	29.11	3.64	45.51
WINA	25%	78.45	46.16	69.69	75.95	42.14	14.10	**54.42**
	40%	77.91	45.56	67.32	75.52	39.58	11.07	52.83
	50%	78.35	44.45	67.96	73.65	36.55	9.63	**51.76**
	65%	74.59	37.88	63.93	66.55	28.81	3.18	**45.82**

Table 7. (G)FLOPs over different sparsity across diffrent model architecture.

Sparsity	QWen2.5-7B	Llama-2-7B	Llama-3-8B	Phi-4
Baseline	7.07	6.61	7.50	14.15
0.25	5.44 (↓ 23.1%)	4.99 (↓ 24.5%)	5.76 (↓ 23.2%)	10.74(↓ 24.1%)
0.4	4.46 (↓ 36.9%)	4.02 (↓ 39.2%)	4.71 (↓ 37.2%)	8.69 (↓ 38.6%)
0.5	3.81 (↓ 46.1%)	3.37 (↓ 49.0%)	4.01 (↓ 46.5%)	7.33 (↓ 48.2%)
0.65	2.83 (↓ 60.0%)	2.40 (↓ 63.7%)	2.97 (↓ 60.4%)	5.28 (↓ 62.7%)

Acceleration. In addition to performance gains, WINA yields substantial computational acceleration across all evaluated LLMs. As shown in Table 7, WINA reduces the overall (G)FLOPs by up to 60.0% on Qwen-2.5-7B, 63.7% on Llama-2-7B, 60.4% on Llama-3-8B, and 62.7% on Phi-4-14B at the 65% sparsity level. These consistent reductions in floating point operations could translate to faster inference speeds and lower computational costs, validating WINA's effectiveness as a practical solution for deployment under tight resource constraints.

B. Orthogonal Tensor Transformation

Without loss of generality, we present the pseudocode for the orthogonal tensor transformation applied to a transformer-based model $\mathcal{M}$ comprising L decoder layers. Each layer includes the following weight matrices: $\left\{W_k^{(\ell)}, W_q^{(\ell)}, W_v^{(\ell)}, W_o^{(\ell)}, W_{\text{gate}}^{(\ell)}, W_{\text{up}}^{(\ell)}, W_{\text{down}}^{(\ell)}\right\}$ for $\ell = 1, \ldots, L$, along with the output projection matrix W_{head} of the final head layer. While we focus on this specific setup for clarity, the transformation is readily applicable to other transformer architectures, including those with encoder-decoder structures or alternative feedforward configurations.

C. Lemma regarding the optimal approximation error over single layer

Lemma C.1 (Optimal approximation error over single layer)**.** *Let $\boldsymbol{x} \in \mathbb{R}^n$ be an input vector and $W \in \mathbb{R}^{m\times n}$ be a matrix satisfying column-wise orthogonality: $W^\top W = diag(c_1^2, c_2^2, \ldots c_n^2)$ where $c_i = \|W_{:,i}\|$. For any target sparsity level $k \in \mathbb{N}^+$ satisfying $k < n$, the expected deviation between the original network output and the gated output via WINA is less or equal to that of TEAL's. Formally:*

$$\mathbb{E}\left[\|W\boldsymbol{x}_{WINA} - W\boldsymbol{x}\|_2^2\right] \le \mathbb{E}\left[\|W\boldsymbol{x}_{TEAL} - Wx\|_2^2\right],$$

where $\boldsymbol{x}_{WINA}$ is the sparse input via WINA, retaining the k elements activated with the largest $|x_j \cdot \|W_{\cdot,j}\|_2|$, and $\boldsymbol{x}_{TEAL}$ is the

Table 5. Results of controlled sparsity experiments over Llama-3-8B

Method	Sparsity	PiQA	Arc-c	WinoGrande	HellaSwag	MMLU	GSM8K	Avg
Baseline	-	80.79	53.33	72.61	79.17	62.20	50.19	66.38
TEAL (Liu et al., 2024)	25%	80.25	53.16	73.32	78.85	61.85	48.07	65.58
	40%	79.11	48.98	71.82	77.43	59.26	39.27	62.65
	50%	78.24	48.12	70.01	74.83	54.50	27.37	58.51
	65%	73.34	37.37	63.46	61.76	32.07	4.17	45.36
TEAL-transform	25%	80.85	53.50	73.16	78.85	61.57	47.99	65.99
	40%	79.43	50.60	70.88	77.36	59.23	40.11	62.94
	50%	77.69	48.38	69.06	75.70	54.82	29.49	59.19
	65%	73.23	39.51	61.96	65.25	38.66	5.08	47.28
WINA	25%	80.79	53.16	73.24	78.96	61.54	48.29	**66.00**
	40%	79.60	50.09	71.27	77.54	58.82	41.85	**63.20**
	50%	78.35	49.06	70.32	75.12	55.26	29.34	**59.57**
	65%	73.45	40.10	62.67	64.89	38.48	7.05	**47.77**

Table 6. Results of controlled sparsity experiments over Phi-4-14B

Method	Sparsity	PiQA	WinoGrande	HellaSwag	Arc-c	MMLU	GSM8K	Avg
Baseline	-	81.28	76.80	81.93	55.97	77.06	90.22	77.21
TEAL (Liu et al., 2024)	25%	81.07	75.45	81.92	56.23	76.63	89.84	76.86
	40%	80.79	73.80	81.21	54.95	75.10	88.02	75.98
	50%	80.63	71.98	80.06	53.84	73.52	86.13	74.36
	65%	77.64	66.06	74.26	50.77	65.17	74.37	68.71
TEAL-transform	25%	80.96	74.59	81.60	55.63	76.68	89.92	76.56
	40%	81.18	74.19	80.94	54.61	75.99	90.07	76.50
	50%	79.82	72.38	79.79	53.92	74.51	88.02	74.74
	65%	77.64	68.51	74.72	52.47	66.64	77.18	69.86
WINA	25%	81.01	75.37	81.91	56.31	76.60	90.22	**77.57**
	40%	81.18	72.45	81.44	56.06	76.44	90.67	**76.71**
	50%	81.39	73.95	81.75	54.95	75.83	87.57	**75.91**
	65%	78.24	70.72	77.10	51.11	70.05	77.10	**70.72**

Algorithm 1 Orthogonal Tensor Transformation

Input: Model $\mathcal{M}$ with matrix W_{emb} of embedding layer, L decoder layers with matrices $\{W_k^{(\ell)}, W_q^{(\ell)}, W_v^{(\ell)}, W_o^{(\ell)}, W_{gate}^{(\ell)}, W_{up}^{(\ell)}, W_{down}^{(\ell)}\}$, and matrix W_{head} of head layer.
Output:
  Transformed model M'
$W_k^{(0)} = U\Sigma V^\top$ (▷) Perform SVD on $W_k^{(0)}$
$Q_k^{(0)} \leftarrow V$
$\hat{W}_{emb} \leftarrow W_{emb} Q_k^{(0)}$
for $\ell = 1, 2, \ldots, L$ **do**
  $\hat{W}_k^{(\ell)} \leftarrow W_k^{(\ell)} Q_k^{(\ell)}, \quad \hat{W}_q^{(\ell)} \leftarrow W_q^{(\ell)} Q_k^{(\ell)}, \quad \hat{W}_v^{(\ell)} \leftarrow W_v^{(\ell)} Q_k^{(\ell)}$
  $W_{gate}^{(\ell)} = U\Sigma V^\top$ (▷) Perform SVD on $W_{gate}^{(\ell)}$
  $Q_{gate}^{(\ell)} \leftarrow V$
  $\hat{W}_o^{(\ell)} \leftarrow (Q_{gate}^{(\ell)})^\top W_o^{(\ell)}$
  $\hat{W}_{gate}^{(\ell)} \leftarrow W_{gate}^{(\ell)} Q_{gate}^{(\ell)}, \quad \hat{W}_{up}^{(\ell)} \leftarrow W_{up}^{(\ell)} Q_{gate}^{(\ell)}$
  if $\ell < L$ **then**
    $W_k^{(\ell+1)} = U\Sigma V^\top$ (▷) Perform SVD on $W_k^{(\ell+1)}$
    $Q_k^{(\ell+1)} \leftarrow V$
    $\hat{W}_{down}^{(\ell)} \leftarrow (Q_k^{(\ell)})^\top W_{down}^{(\ell)}$
  end if
end for
return M'

sparse input via TEAL, retaining the k elements with the largest $|x_j|$.

Proof. Let $\mathcal{I}^{=0}(\boldsymbol{x}) := \{i | \boldsymbol{x}_i = 0\}$ be the set of indices of zero elements at $\boldsymbol{x}$. The output deviation between the original network output and the gated output via a general-format sparsification is:

$$\begin{aligned}
\|W(\boldsymbol{x}_{\mathcal{I}^{=0}} - \boldsymbol{x})\|^2 &= \left\| \sum_{i \in \mathcal{I}^{=0}} \boldsymbol{x}_i W_{:,i} \right\|_2^2 \\
&= \left(\sum_{i \in \mathcal{I}^{=0}} x_i W_{:,i} \right)^\top \left(\sum_{i \in \mathcal{I}^{=0}} \boldsymbol{x}_i W_{:,i} \right) \\
&= \sum_{j \in \mathcal{I}^{=0}} \sum_{i \in \mathcal{I}^{=0}} \boldsymbol{x}_j \boldsymbol{x}_i W_{:,j}^\top W_{:,i} \\
&= \sum_{i \in \mathcal{I}^{=0}} \boldsymbol{x}_i^2 \|W_{:,i}\|_2^2 + \sum_{i \neq j \in \mathcal{I}^{=0}} \boldsymbol{x}_j \boldsymbol{x}_i W_{:,j}^\top W_{:,i}
\end{aligned}$$

The expected output deviation for WINA is:

$$\begin{aligned}
e_{\text{WINA}} &= \left\| W \boldsymbol{x}_{\mathcal{I}^{=0}_{\text{WINA}}} - W\boldsymbol{x} \right\|^2 \\
&= \sum_{i \in \mathcal{I}^{=0}_{\text{WINA}}} \boldsymbol{x}_i^2 \|W_{:,i}\|_2^2 + \sum_{i \neq j \in \mathcal{I}^{=0}_{\text{WINA}}} \boldsymbol{x}_j \boldsymbol{x}_i W_{:,j}^\top W_{:,i}.
\end{aligned}$$

Since W is assumed to be column orthogonal, the cross-term expectations vanish, and the expected output error is determined solely by the main term:

$$e_{\text{WINA}} = \sum_{i \in \mathcal{I}^{=0}_{\text{WINA}}} \boldsymbol{x}_i^2 \|W_{:,i}\|_2^2.$$

Because WINA sparsification sets the k smallest $|\boldsymbol{x}_i \boldsymbol{c}_i|$ terms to zero, we have the mask of WINA reaches out the lower bound of approximation error for a single layer network, *i.e.*,

$$\boldsymbol{g}_{\text{WINA}}(\boldsymbol{x}) = \underset{\boldsymbol{g}\in\{0,1\}^n}{\operatorname{argmin}} \quad \|W(\boldsymbol{x}\odot\boldsymbol{g}-\boldsymbol{x})\|^2. \tag{8}$$

Thus, the above indicates that WINA sparsification achieves the tight lower bound of the approximation error, including those of TEAL and CATS.

D. Proof of Theorem 3.2

Proof. We prove this by mathematical induction.

Step 1: Base case $N = 2$. The output error for sparse activation parameterized via mask $\boldsymbol{g}$ is:

$$\begin{aligned}
&\left\|\boldsymbol{y}_{\boldsymbol{g}}^{(2)} - \boldsymbol{y}^{(2)}\right\| \\
&= \left\|W^{(2)}(\boldsymbol{y}_{\boldsymbol{g}}^{(1)} \odot \boldsymbol{g}^{(2)}) - W^{(2)}\boldsymbol{y}^{(1)}\right\| \\
&= \left\|W^{(2)}(\boldsymbol{y}_{\boldsymbol{g}}^{(1)} \odot \boldsymbol{g}^{(2)}) - W^{(2)}\boldsymbol{y}_{\boldsymbol{g}}^{(1)} + W^{(2)}\boldsymbol{y}_{\boldsymbol{g}}^{(1)} - W^{(2)}\boldsymbol{y}^{(1)}\right\| \\
&= \left\|W^{(2)}(\boldsymbol{y}_{\boldsymbol{g}}^{(1)} \odot \boldsymbol{g}^{(2)} - \boldsymbol{y}_{\boldsymbol{g}}^{(1)}) + W^{(2)}(\boldsymbol{y}_{\boldsymbol{g}}^{(1)} - \boldsymbol{y}^{(1)})\right\| \\
&= \left\|W^{(2)}(\boldsymbol{y}_{\boldsymbol{g}}^{(1)} \odot \boldsymbol{g}^{(2)} - \boldsymbol{y}_{\boldsymbol{g}}^{(1)}) + W^{(2)}(W^{(1)}\boldsymbol{x} \odot \boldsymbol{g}^{(1)} - W^{(1)}\boldsymbol{x})\right\|
\end{aligned}$$

Let:

$$\Delta^{(1)} = \operatorname{diag}(\boldsymbol{g}^{(1)} - 1), \quad \Delta^{(2)} = \operatorname{diag}(\boldsymbol{g}^{(2)} - 1), \quad M^{(1)} = \operatorname{diag}(\boldsymbol{g}^{(1)}).$$

Then, let $\boldsymbol{v}$ and $\boldsymbol{u}$ be

$$\begin{aligned}
\boldsymbol{v} &= W^{(2)}(\boldsymbol{y}_{\boldsymbol{g}}^{(1)} \odot (\boldsymbol{g}^{(2)} - 1)) \\
&= W^{(2)}(W^{(1)}(\boldsymbol{x} \odot \boldsymbol{g}^{(1)}) \odot (\boldsymbol{g}^{(2)} - 1)) \\
&= W^{(2)}\Delta^{(2)}W^{(1)}M^{(1)}\boldsymbol{x} \\
\boldsymbol{u} &= W^{(2)}W^{(1)}(\boldsymbol{x} \odot (\boldsymbol{g}^{(1)} - 1)) \\
&= W^{(2)}W^{(1)}\Delta^{(1)}\boldsymbol{x}.
\end{aligned}$$

Since $\mathbb{E}\|\boldsymbol{u}+\boldsymbol{v}\|^2 = \mathbb{E}\|\boldsymbol{u}\|^2 + \mathbb{E}\|\boldsymbol{v}\|^2 + 2\mathbb{E}(\boldsymbol{u}^\top\boldsymbol{v})$, the expected value of the cross-term is:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{u}^\top\boldsymbol{v}] &= \mathbb{E}[\boldsymbol{x}^\top\Delta^{(1)}(W^{(1)})^\top(W^{(2)})^\top(W^{(2)})\Delta^{(2)}W^{(1)}M^{(1)}\boldsymbol{x}] \\
&= \mathbb{E}[\operatorname{tr}(W^{(2)}\Delta^{(2)}W^{(1)}M^{(1)}\boldsymbol{x}\boldsymbol{x}^\top\Delta^{(1)}(W^{(1)})^\top(W^{(2)})^\top)] \\
&= \operatorname{tr}\left(W^{(2)}\mathbb{E}[\Delta^{(2)}]W^{(1)}\mathbb{E}[M^{(1)}\Delta^{(1)}]\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top](W^{(1)})^\top(W^{(2)})^\top\right)
\end{aligned}$$

Since $\mathbb{E}[M_1\Delta_1] = \mathbb{E}[\boldsymbol{g}^{(1)} \odot (\boldsymbol{g}^{(1)} - 1)] = 0$, the cross-term expectation $\mathbb{E}[\boldsymbol{u}^\top\boldsymbol{v}]$ is zero. Thus, the expected output deviation via sparse activation $\boldsymbol{g}$,

$$\begin{aligned}
e^{(2)} &= \mathbb{E}[\|\boldsymbol{u}+\boldsymbol{v}\|^2] \\
&= \mathbb{E}[\left\|W^{(2)}(\boldsymbol{y}_{\boldsymbol{g}}^{(1)} \odot \boldsymbol{g}^{(2)} - \boldsymbol{y}_{\boldsymbol{g}}^{(1)})\right\|^2] + \mathbb{E}[\left\|W^{(2)}(W^{(1)}\boldsymbol{x} \odot \boldsymbol{g}^{(1)} - W^{(1)}\boldsymbol{x})\right\|^2]
\end{aligned} \tag{9}$$

Upon Lemma C.1, we have that

$$\mathbb{E}[\|W^{(2)}(W^{(1)}\boldsymbol{x} \odot \boldsymbol{g}_{\text{WINA}}^{(1)} - W^{(1)}\boldsymbol{x})\|^2] \le \mathbb{E}[\|W^{(2)}(W^{(1)}\boldsymbol{x} \odot \boldsymbol{g}_{\text{TEAL}}^{(1)} - W^{(1)}\boldsymbol{x})\|^2]$$

Next, we compare $\mathbb{E}[\|W^{(2)}(\boldsymbol{y}_{\boldsymbol{g}}^{(1)} \odot \boldsymbol{g}^{(2)} - \boldsymbol{y}_{\boldsymbol{g}}^{(1)})\|^2]$ given $\boldsymbol{g}_{\text{WINA}}^{(2)}$ and $\boldsymbol{g}_{\text{TEAL}}^{(2)}$.

$$
\begin{aligned}
&\mathbb{E}[\|W^{(2)}(\boldsymbol{y}_{\boldsymbol{g}}^{(1)} \odot \boldsymbol{g}^{(2)} - \boldsymbol{y}_{\boldsymbol{g}}^{(1)})\|^2] \\
=\ &\mathbb{E}[\sum_{j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(2)})}\ \sum_{i\in\mathcal{I}^{=0}(\boldsymbol{g}^{(2)})} \boldsymbol{y}_j^{(1)}\boldsymbol{y}_i^{(1)}\,(W_{:,j}^{(2)})^\top W_{:,i}^{(2)}] \\
=\ &\mathbb{E}[\sum_{j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(2)})} (\boldsymbol{y}_j^{(1)})^2\|W_{:,j}^{(2)}\|^2 + \sum_{\substack{i,j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(2)})\\ i\neq j}} \boldsymbol{y}_j^{(1)}\boldsymbol{y}_i^{(1)}\,(W_{:,j}^{(2)})^T W_{:,i}^{(2)}] \\
=\ &\sum_{j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(2)})} (\boldsymbol{c}_j^{(2)})^2\mathbb{E}(\boldsymbol{y}_j^{(1)})^2 + \sum_{\substack{i,j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(2)})\\ i\neq j}} (W_{:,j}^{(2)})^\top W_{:,i}^{(2)}\mathbb{E}[\boldsymbol{y}_j^{(1)}\boldsymbol{y}_i^{(1)}] \\
=\ &\sum_{j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(2)})} (\boldsymbol{c}_j^{(2)})^2\mathbb{E}(\boldsymbol{y}_j^{(1)})^2,
\end{aligned}
$$

where the last line is due to $W^{(2)}$ is column-orthogonal, the cross-term's expectation is zero.

Because WINA sparsification sets the k smallest $(\boldsymbol{y}_j^{(1)}c_j^{(2)})^2$ terms to zero, we have:

$$
\mathbb{E}[\|W^{(2)}(\boldsymbol{y}_{\boldsymbol{g}_{\text{WINA}}}^{(1)} \odot \boldsymbol{g}^{(2)} - \boldsymbol{y}_{\boldsymbol{g}_{\text{WINA}}}^{(1)})\|^2] \le \mathbb{E}[\|W^{(2)}(\boldsymbol{y}_{\boldsymbol{g}_{\text{TEAL}}}^{(1)} \odot \boldsymbol{g}^{(2)} - \boldsymbol{y}_{\boldsymbol{g}_{\text{TEAL}}}^{(1)})\|^2]
$$

Therefore, we have that

$$
e_{\text{WINA}}^{(2)} \le e_{\text{TEAL}}^{(2)}.
$$

Step 2: Inductive proof for $N > 2$. Assume for some $N \ge 2$ that

$$
e_{\text{WINA}}^{(N)} \le e_{\text{TEAL}}^{(N)}.
$$

Define the exact output of $(N+1)$ layer network:

$$
\boldsymbol{y} = W^{(N+1)}\boldsymbol{y}^{(N)}, \quad \boldsymbol{y}^{(N)} = W^{(N)}\cdots W^{(1)}\boldsymbol{x}
$$

The output via mask $\boldsymbol{g}^{(N+1)}$ is that

$$
\begin{aligned}
&\boldsymbol{y}_{\boldsymbol{g}}^{(N+1)} - \boldsymbol{y} \\
=\ &W^{(N+1)}(\boldsymbol{y}_{\boldsymbol{g}}^{(N)} \odot \boldsymbol{g}^{(N+1)}) - W^{(N+1)}\boldsymbol{y}^{(N)} \\
=\ &W^{(N+1)}((\boldsymbol{y}_{\boldsymbol{g}}^{(N)} \odot \boldsymbol{g}^{(N+1)}) - \boldsymbol{y}_{\boldsymbol{g}}^{(N)}) + W^{(N+1)}(\boldsymbol{y}_{\boldsymbol{g}}^{(N)} - \boldsymbol{y}^{(N)})
\end{aligned}
$$

The expected output deviation is:

$$
e_{\boldsymbol{g}}^{N+1} = \mathbb{E}\|W^{(N+1)}(\boldsymbol{y}_{\boldsymbol{g}}^{(N)} \odot \boldsymbol{g}^{(N+1)} - \boldsymbol{y}_{\boldsymbol{g}}^{(N)})\|^2 + \mathbb{E}\|W^{(N+1)}(\boldsymbol{y}_{\boldsymbol{g}}^{(N)} - \boldsymbol{y}^{(N)})\|^2, \tag{10}
$$

the cross-term zeros out because of the assumption.

Upon induction assumption, for the second term, we have that

$$
\mathbb{E}\|W^{(N+1)}(\boldsymbol{y}_{\boldsymbol{g}_{\text{WINA}}}^{(N)} - \boldsymbol{y}^{(N)})\|^2 \le \mathbb{E}\|W^{(N+1)}(\boldsymbol{y}_{\boldsymbol{g}_{\text{TEAL}}}^{(N)} - \boldsymbol{y}^{(N)})\|^2. \tag{11}
$$

For the first term, we have that

$$
\begin{aligned}
&\mathbb{E}[\|W^{(N+1)}(\boldsymbol{y}_{\boldsymbol{g}}^{(N)} \odot \boldsymbol{g}^{(N+1)} - \boldsymbol{y}_{\boldsymbol{g}}^{(N)})\|^2] \\
=\ &\mathbb{E}[\sum_{j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(N+1)})}\ \sum_{i\in\mathcal{I}^{=0}(\boldsymbol{g}^{(N+1)})} \boldsymbol{y}_j^{(N)}\boldsymbol{y}_i^{(N)}\,(W_{:,j}^{(N+1)})^\top W_{:,i}^{(N+1)}] \\
=\ &\sum_{j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(N+1)})} (\boldsymbol{c}_j^{(N+1)})^2\mathbb{E}(\boldsymbol{y}_j^{(N)})^2 + \sum_{\substack{i,j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(N+1)})\\ i\neq j}} (W_{:,j}^{(N+1)})^\top W_{:,i}^{(N+1)}\mathbb{E}[\boldsymbol{y}_j^{(N)}\boldsymbol{y}_i^{(N)}] \\
=\ &\sum_{j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(N+1)})} (\boldsymbol{c}_j^{(N+1)})^2\mathbb{E}(\boldsymbol{y}_j^{(N)})^2,
\end{aligned}
$$

where the last line is due to $W^{(N+1)}$ is column-orthogonal, the cross-term's expectation is zero.

Since WINA retains the k largest $|\boldsymbol{y}_j^{(N)} \boldsymbol{c}_j^{N+1}|$, thus:

$$\mathbb{E}[\|W^{(N+1)}(\boldsymbol{y}_{\boldsymbol{g}}^{(N)} \odot \boldsymbol{g}_{\text{WINA}}^{(N+1)} - \boldsymbol{y}_{\boldsymbol{g}}^{(N)})\|^2] \leq \mathbb{E}[\|W^{(N+1)}(\boldsymbol{y}_{\boldsymbol{g}}^{(N)} \odot \boldsymbol{g}_{\text{TEAL}}^{(N+1)} - \boldsymbol{y}_{\boldsymbol{g}}^{(N)})\|^2]. \tag{12}$$

Consequently, we reach the conclusion that

$$e_{\text{WINA}}^{(N+1)} \leq e_{\text{TEAL}}^{(N+1)}. \tag{13}$$

E. Lemma regarding the optimal approximation error over a single layer with activation function

Lemma E.1 (Optimal approximation error over a single layer with activation function)**.** *Let $\boldsymbol{x} \in \mathbb{R}^n$ be a zero-mean input vector, $W \in \mathbb{R}^{m \times n}$ be a matrix that satisfies column-wise orthogonality: $W^\top W = diag(c_1^2, c_2^2, \dots c_n^2)$ where $c_i = \|W_{:,i}\|$ and $f : \mathbb{R} \to \mathbb{R}$ be an activation function. For any target sparsity level $k \in \mathbb{N}^+$ satisfying $k < n$, the expected deviation between the original output and the gated output via WINA gating mechanism is less than or equal to that of TEAL gating mechanism. Formally:*

$$\mathbb{E}\left[\|f(W\boldsymbol{x}_{WINA}) - f(W\boldsymbol{x})\|_2^2\right] \leq \mathbb{E}\left[\|f(W\boldsymbol{x}_{TEAL}) - f(W\boldsymbol{x})\|_2^2\right],$$

where $\boldsymbol{x}_{WINA}$ is the sparse input via WINA, retaining the k elements with the largest $|x_j \cdot \|W_{\cdot,j}\|_2|$, and $\boldsymbol{x}_{TEAL}$ is the sparse input via TEAL, retaining the k elements with the largest $|x_j|$.

Proof. Let Δ be the error term of the output via sparse activation parameterized with $\boldsymbol{g}$,

$$\Delta_{\boldsymbol{g}} = W(\boldsymbol{x} \odot (1 - \boldsymbol{g})) = \sum_{i=1}^{d} W_{i,:}\boldsymbol{x}_i \odot (1 - \boldsymbol{g}_i).$$

Using a Taylor expansion and ignoring higher-order terms (assuming Δ_i is small), the output deviation given an activation function f is:

$$f(W_{i,:}\boldsymbol{x} + \Delta_{\boldsymbol{g},i}) - f(W_{i,:}\boldsymbol{x}) \approx \nabla^\top f(W_{:,i}\boldsymbol{x})\Delta_i.$$

Thus, the expected squared output deviation between the original output and the gated output approximates to:

$$\begin{aligned}
e_{\boldsymbol{g}} &= \mathbb{E}\|f(W\boldsymbol{x} + \Delta_{\boldsymbol{g}}) - f(W\boldsymbol{x})\|^2 \\
&= \mathbb{E}\|\sum_{i=1}^{d} f(W_{i,:}\boldsymbol{x} + \Delta_{\boldsymbol{g},i}) - f(W_{i,:}\boldsymbol{x})\|^2 \\
&\approx \mathbb{E}\left[\|\sum_{i=1}^{d} \nabla f(W_{i,:}\boldsymbol{x})\Delta_{\boldsymbol{g},i}\|^2\right] \\
&= \sum_{i=1}^{d} \mathbb{E}[\nabla^2 f(W_{i,:}\boldsymbol{x})\Delta_{\boldsymbol{g},i}^2] \\
&= \sum_{i=1}^{d} \mathbb{E}[\nabla^2 f(W_{i,:}\boldsymbol{x})(W_{i,:}\boldsymbol{x}_i \odot (1 - \boldsymbol{g}_i))^2] \\
&= \sum_{i=1}^{d} \mathbb{E}[\nabla^2 f(W_{i,:}\boldsymbol{x})] \sum_{i=1}^{d} \mathbb{E}(W_{i,:}\boldsymbol{x}_i \odot (1 - \boldsymbol{g}_i))^2 \\
&= \sum_{i=1}^{d} \mathbb{E}[\nabla^2 f(W_{i,:}\boldsymbol{x})] \sum_{i \in \mathcal{I}^{=0}(\boldsymbol{g})}^{d} \mathbb{E}[\boldsymbol{c}_i^2 \boldsymbol{x}_i^2]
\end{aligned}$$

Because WINA sparsification select the k smallest $\boldsymbol{x}_j^2 \boldsymbol{c}_j^2$ terms to zero, we have that

$$e_{\text{WINA}} \leq e_{\text{TEAL}}. \tag{14}$$

F. Proof of Theorem 3.5

Proof. We prove this by mathematical induction.

Step 1: Base case $N = 2$**.** The output error for sparse activation via $\boldsymbol{g}^{(2)}$ is:

$$\begin{aligned}
&\left\|\boldsymbol{y}_{\boldsymbol{g}}^{(2)} - \boldsymbol{y}^{(2)}\right\| \\
&= \left\|W^{(2)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(1)}) \odot \boldsymbol{g}^{(2)}) - W^{(2)} f(\boldsymbol{y}^{(1)})\right\| \\
&= \left\|W^{(2)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(1)}) \odot \boldsymbol{g}^{(2)}) - W^{(2)} f(y_{\boldsymbol{g}}^{(1)}) + W^{(2)} f(y_{\boldsymbol{g}}^{(1)}) - W^{(2)} f(\boldsymbol{y}^{(1)})\right\| \\
&= \left\|W^{(2)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(1)}) \odot \boldsymbol{g}^{(2)}) - f(y_{\boldsymbol{g}}^{(1)})) + W^{(2)}(f(y_{\boldsymbol{g}}^{(1)}) - f(\boldsymbol{y}^{(1)}))\right\|.
\end{aligned}$$

Let:

$$M^{(1)} = \mathrm{diag}(\boldsymbol{g}^{(1)}), \quad M^{(2)} = \mathrm{diag}(\boldsymbol{g}^{(2)} - 1).$$

Then, let $\boldsymbol{v}$ and $\boldsymbol{u}$ be

$$\begin{aligned}
\boldsymbol{v} &= W^{(2)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(1)}) \odot (\boldsymbol{g}^{(2)} - 1)) \\
&= W^{(2)} M^{(2)} f(W^{(1)} M^{(1)} \boldsymbol{x}) \\
\boldsymbol{u} &= W^{(2)}[f(W^{(1)} M^{(1)} \boldsymbol{x}) - f(W^{(1)} \boldsymbol{x})].
\end{aligned}$$

Let $D = {W^{(2)}}^{\top} W^{(2)}$, then the expected value of the cross-term becomes:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{u}^{\top}\boldsymbol{v}] &= \mathbb{E}[f(W^{(1)}M^{(1)}x) - f(W^{(1)}x)]^{\top} {W^{(2)}}^{\top} W^{(2)} M^{(2)} f(W^{(1)}M^{(1)}x) \\
&= \mathbb{E}\left[f(W^{(1)}M^{(1)}x) - f(W^{(1)}x)\right]^{T} D M^{(2)} f(W^{(1)}M^{(1)}x) \\
&= \mathbb{E}\sum_{i} D_{ii} \cdot (M^{(2)})_{ii} \cdot \left(f(W^{(1)}M^{(1)}\boldsymbol{x})_i - f(W^{(1)}\boldsymbol{x})_i\right) \cdot f(W^{(1)}M^{(1)}x)_i
\end{aligned}$$

When $\boldsymbol{g}_i^{(2)} = 1$, $(M^{(2)})_{ii} = 0$, and the corresponding terms disappear. When $\boldsymbol{g}_i^{(2)} = 0$, $(M^{(2)})_{ii} = 1$. Therefore:

$$E\left[\boldsymbol{u}^{\top}\boldsymbol{v}\right] = E\left[\sum_{i:\boldsymbol{g}_i^{(2)}=0} D_{ii} \cdot \left(f(W^{(1)}M^{(1)}\boldsymbol{x})_i - f(W^{(1)}\boldsymbol{x})_i\right) \cdot f(W^{(1)}M^{(1)}\boldsymbol{x})_i\right]$$

Since $\boldsymbol{x}$ follows a symmetric distribution with mean 0, and $W^{(1)}$ has orthogonal columns, the distributions of $W^{(1)}M^{(1)}\boldsymbol{x}$ and $W^{(1)}x$ are symmetric. For any activation function f, the cross-term cancels out under the symmetric distribution. Thus, the expected output deviation becomes

$$\begin{aligned}
e_{\boldsymbol{g}}^{(2)} &= \mathbb{E}[\|\boldsymbol{u} + \boldsymbol{v}\|^2] = \mathbb{E}[\|\boldsymbol{u}\|^2] + \mathbb{E}[\|\boldsymbol{v}\|^2] \\
&= \mathbb{E}[\|W^{(2)}M^{(2)}f(W^{(1)}M^{(1)}\boldsymbol{x})\|^2] + \mathbb{E}[\|W^{(2)}[f(W^{(1)}M^{(1)}\boldsymbol{x}) - f(W^{(1)}\boldsymbol{x})]\|_2^2]
\end{aligned}$$

Here, the latter one yields the below due to Lemma **??**.

$$\mathbb{E}[\|W^{(2)}[f(W^{(1)}M^{(1)}_{\text{WINA}}\boldsymbol{x}) - f(W^{(1)}\boldsymbol{x})]\|^2] \le \mathbb{E}[\|W^{(2)}[f(W^{(1)}M^{(1)}_{\text{TEAL}}\boldsymbol{x}) - f(W^{(1)}\boldsymbol{x})]\|^2].$$

Next, we compare the former term. We have that:

$$\begin{aligned}
&\mathbb{E}[\|W^{(2)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(1)}) \odot \boldsymbol{g}^{(2)}) - W^{(2)} f(\boldsymbol{y}_{\boldsymbol{g}}^{(1)})\|^2] \\
=&\mathbb{E}\sum_{j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(1)})}\sum_{i\in\mathcal{I}^{=0}(\boldsymbol{g}^{(1)})} f(\boldsymbol{y}_j^{(1)})\, f(\boldsymbol{y}_i^{(1)})\, (W_{:,j}^{(2)})^T W_{:,i}^{(2)} \\
=&\mathbb{E}\sum_{j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(1)})} f(\boldsymbol{y}_j^{(1)})^2 \|W_{:,j}^{(2)}\|^2 + \sum_{\substack{i,j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(1)}) \\ i\neq j}} f(\boldsymbol{y}_j^{(1)})\, f(\boldsymbol{y}_i^{(1)}) (W_{:,j}^{(2)})^{\top} W_{:,i}^{(2)} \\
=&\sum_{j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(1)})} (\boldsymbol{c}_j^{(2)})^2 \mathbb{E} f(\boldsymbol{y}_{\boldsymbol{g},j}^{(1)})^2 + \sum_{\substack{i,j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(1)}) \\ i\neq j}} (W_{:,j}^{(2)})^T W_{:,i}^{(2)} \mathbb{E} f(\boldsymbol{y}_j^{(1)}) f(\boldsymbol{y}_i^{(1)}) \\
=&\sum_{j\in\mathcal{I}^{=0}(\boldsymbol{g}^{(1)})} (\boldsymbol{c}_j^{(2)})^2 \mathbb{E} f(\boldsymbol{y}_{\boldsymbol{g},j}^{(1)})^2,
\end{aligned}$$

where the last line is due to $W^{(2)}$ being column-orthogonal, thereby the cross-term's expectation is zero.

Because WINA sparsification sets the k smallest $(f(\boldsymbol{y}_{\boldsymbol{g},j}^{(1)})\boldsymbol{c}_j^{(2)})^2$ terms to zero, we have that

$$\mathbb{E}[\|W^{(2)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(1)}) \odot \boldsymbol{g}_{\text{WINA}}^{(2)}) - W^{(2)} f(\boldsymbol{y}_{\boldsymbol{g}}^{(1)})\|^2] \leq \mathbb{E}[\|W^{(2)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(1)}) \odot \boldsymbol{g}_{\text{TEAL}}^{(2)}) - W^{(2)} f(\boldsymbol{y}_{\boldsymbol{g}}^{(1)})\|^2]$$

Thus, we have that

$$e_{\text{WINA}}^{(2)} \leq e_{\text{TEAL}}^{(2)}.$$

Step 2: Inductive proof for $N > 2$. Assume for $N \geq 2$, the below holds

$$e_{\text{WINA}}^{(N)} \leq e_{\text{TEAL}}^{(N)}.$$

Consider the output of $(N+1)$ layers network, *i.e.*, $\boldsymbol{y}^{(N+1)} = W^{(N+1)} f(\boldsymbol{y}^{(N)})$.

The output deviation via sparse activation of $\boldsymbol{g}$ is:

$$\begin{aligned}
&\boldsymbol{y}_{\boldsymbol{g}}^{(N+1)} - \boldsymbol{y}^{(N+1)} \\
=&W^{(N+1)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}) \odot \boldsymbol{g}^{(N+1)}) - W^{(N+1)} f(\boldsymbol{y}^{(N)}) \\
=&W^{(N+1)}((f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}) \odot \boldsymbol{g}^{(N+1)}) - f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)})) + W^{(N+1)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}) - f(\boldsymbol{y}^{(N)}))
\end{aligned}$$

The expected output deviation is:

$$e_{\boldsymbol{g}}^{N+1} = \mathbb{E}\|W^{(N+1)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}) \odot \boldsymbol{g}^{(N+1)} - f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}))\|^2 + \mathbb{E}\|W^{(N+1)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}) - f(\boldsymbol{y}^{(N)}))\|^2, \tag{15}$$

the cross-term zeros out because of the assumption.

Upon the induction assumption, the second term yields that

$$\mathbb{E}\|W^{(N+1)}(f(\boldsymbol{y}_{\boldsymbol{g}_{\text{WINA}}}^{(N)}) - f(\boldsymbol{y}^{(N)}))\|^2 \leq \mathbb{E}\|W^{(N+1)}(f(\boldsymbol{y}_{\boldsymbol{g}_{\text{TEAL}}}^{(N)}) - f(\boldsymbol{y}^{(N)}))\|^2. \tag{16}$$

For the first term, we have that

$$\begin{aligned}
&\mathbb{E}\|W^{(N+1)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}) \odot \boldsymbol{g}^{(N+1)} - f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}))\|^2 \\
=&\,\mathbb{E} \sum_{j \in \mathcal{I}^{=0}(\boldsymbol{g}^{(N)})} f(\boldsymbol{y}_{\boldsymbol{g},j}^{(N)})^2 \|W_{:,j}^{(N+1)}\|_2^2 + \sum_{\substack{i,j \in \mathcal{I}^{=0}(\boldsymbol{g}^{(N)}) \\ i \neq j}} f(\boldsymbol{y}_{\boldsymbol{g},i}^{(N)}) f(\boldsymbol{y}_{\boldsymbol{g},j}^{(N)}) (W_{:,j}^{(N+1)})^\top W_{:,i}^{(N+1)} \\
=&\,\mathbb{E} \sum_{j \in \mathcal{I}^{=0}(\boldsymbol{g}^{(N)})} f(\boldsymbol{y}_{\boldsymbol{g},j}^{(N)})^2 \|W_{:,j}^{(N+1)}\|_2^2 \\
=&\,\mathbb{E} \sum_{j \in \mathcal{I}^{=0}(\boldsymbol{g}^{(N)})} f(\boldsymbol{y}_{\boldsymbol{g},j}^{(N)})^2 \boldsymbol{c}_j^2.
\end{aligned}$$

Since WINA retains the k largest $\mathbb{E}[f(\boldsymbol{y}_{\boldsymbol{g},j}^{(N)})^2 \boldsymbol{c}_j^2]$, therefore:

$$\mathbb{E}\|W^{(N+1)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}) \odot \boldsymbol{g}_{\text{WINA}}^{(N+1)} - f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}))\|^2 \leq \mathbb{E}\|W^{(N+1)}(f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}) \odot \boldsymbol{g}_{\text{TEAL}}^{(N+1)} - f(\boldsymbol{y}_{\boldsymbol{g}}^{(N)}))\|^2.$$

Consequently, we conclude that

$$e_{\text{WINA}}^{(N+1)} \leq e_{\text{TEAL}}^{(N+1)}.$$

G. Resources Used & Limitations

The total run time of our experiments were run using two A100 80GB GPUs for a couple of days.

In terms of limitations, we focus the comparisons of our approach with current leading methodologies for sparse activation (i.e., TEAL (Liu et al., 2024) and CATS (Lee et al., 2024)). Naturally, we are unable to compare with all existing sparse activation methodologies and prior works, but, instead, we use these TEAL and CATS as they currently represent the current upper bound of optimal performance-efficiency trade-offs; as such, we use these approaches to compare against in order to ensure our performance tests and comparisons are robust and fair.