This AI Paper from Microsoft Introduces WINA: A Training-Free Sparse Activation Framework for Efficient Large Language Model Inference

By **Sana Hassan** - May 31, 2025

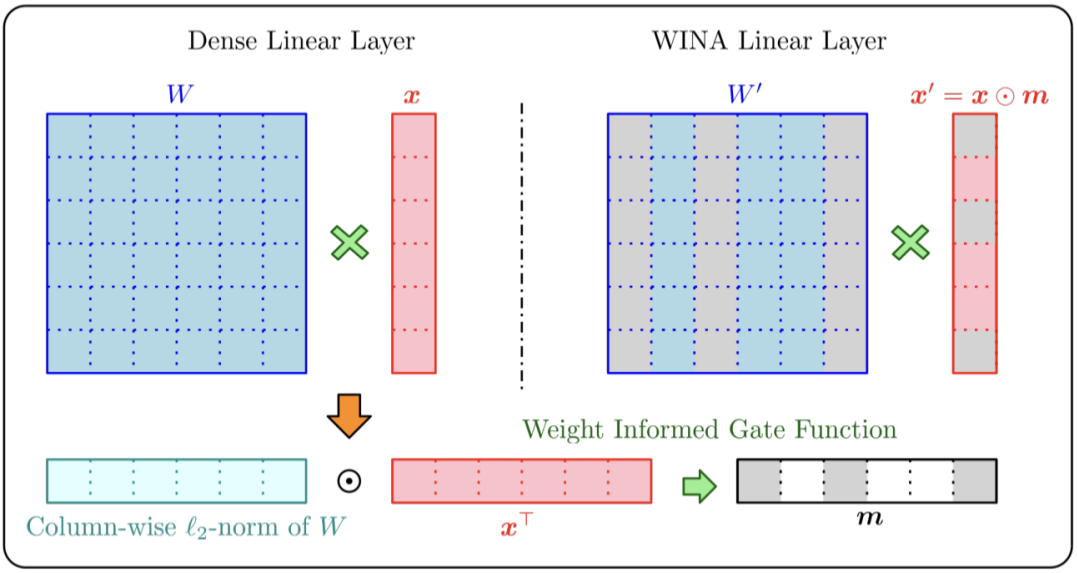


Figure 1: **Overview of WINA.** WINA performs training-free sparse activation by selecting the most influential input dimensions based on both hidden state magnitudes and the column-wise ℓ_2-norms of weight matrices. This joint criterion ensures accurate sub-network activation at each layer during inference, preserving model performance while reducing computational overhead.

Large language models (LLMs), with billions of parameters, power many AI-driven services across industries. However, their massive size and complex architectures make their computational costs during inference a significant challenge. As these models evolve, optimizing the balance between computational efficiency and output quality has become a crucial area of research.

The core challenge lies in how LLMs handle inference. Every time an input is processed, the entire model is activated, which consumes extensive computational resources. This full activation is unnecessary for most tasks, as only a small subset of neurons contribute meaningfully to the final output. Existing sparse activation methods attempt to address this by selectively deactivating less important neurons. However, these approaches often focus only on the magnitude of hidden states while ignoring the critical role of weight matrices in propagating errors through the network. This oversight leads to high approximation errors and deteriorates model performance, particularly at higher sparsity levels.

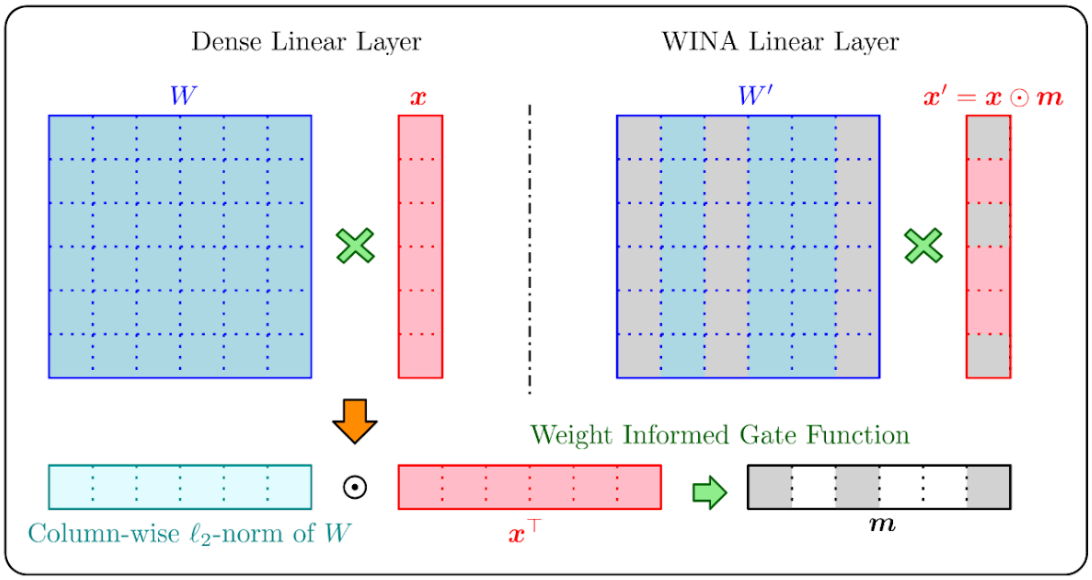


Figure 1: **Overview of WINA.** WINA performs training-free sparse activation by selecting the most influential input dimensions based on both hidden state magnitudes and the column-wise ℓ_2-norms of weight matrices. This joint criterion ensures accurate sub-network activation at each layer during inference, preserving model performance while reducing computational overhead.

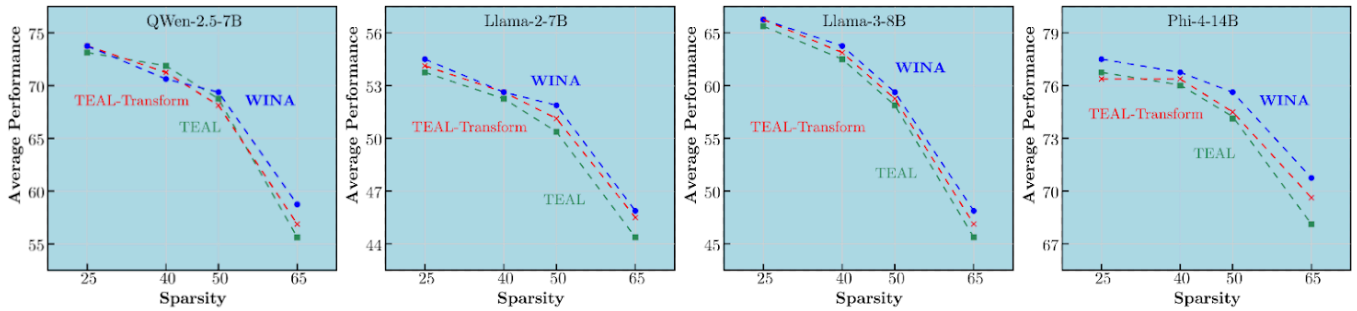
🎤 Talks • Demos • Networking • Certificate

FREE REGISTRATION

Sparse activation techniques have included methods like Mixture-of-Experts (MoE) used in models such as GPT-4 and Mistral, which rely on additional training to learn which experts to activate for each input. Other approaches, such as TEAL and CATS, aim to reduce computation by using the size of hidden activations to prune neurons, but they still leave room for improvement. These methods often struggle with balancing sparsity and accuracy, as they can

mistakenly deactivate important neurons or retain those with minimal influence. Moreover, they require model-specific threshold tuning, making them less flexible across different architectures.

Researchers from Microsoft, Renmin University of China, New York University, and the South China University of Technology proposed a new method called WINA (Weight Informed Neuron Activation) to address these issues. WINA introduces a training-free sparse activation technique that uses both hidden state magnitudes and column-wise ℓ2 norms of weight matrices to determine which neurons to activate during inference. By considering the combined impact of input magnitudes and weight importance, WINA creates a more effective sparsification strategy that adapts to different layers of the model without requiring retraining or fine-tuning.



(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.

The WINA method is built on a simple yet powerful idea: neurons that have strong activations and large weight magnitudes are more likely to influence downstream computations. To operationalize this, WINA calculates the element-wise product of hidden states and weight norms, selecting the top-K components based on this combined metric. This strategy allows WINA to construct a sparse sub-network that preserves the most important signals while ignoring redundant activations. The method also includes a tensor transformation step that enforces column-wise orthogonality in weight matrices, ensuring theoretical error bounds translate effectively to real-world performance. By combining these steps, WINA maintains a tight approximation error while delivering significant computational savings.

The research team evaluated WINA on several large language models, including Qwen-2.5-7B, LLaMA-2-7B, LLaMA-3-8B, and Phi-4-14B, across various tasks and sparsity levels. WINA outperformed TEAL and CATS across all tested models and sparsity settings. For example, on Qwen-2.5-7B at 65% sparsity, WINA achieved up to 2.94% higher average performance than TEAL and 1.41% better than TEAL-Transform. On LLaMA-3-8B, WINA delivered gains of 1.06% at 50% sparsity and 2.41% at 65% sparsity. Even at high sparsity levels, WINA retained stronger performance on reasoning-intensive tasks like GSM8K and ARC Challenge. WINA also delivered consistent computational savings, reducing floating-point operations by up to 63.7% on LLaMA-2-7B and 62.7% on Phi-4-14B.

Table 1: 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**

Llama-2-7B. On Llama-2-7B, WINA again shows strong performance under various sparsity constraints. As shown in Table 2, 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.

In summary, WINA offers a robust, training-free solution for sparse activation in large language models by combining hidden state magnitudes with weight matrix norms. This approach addresses the limitations of prior methods, such as TEAL, resulting in lower approximation errors, improved accuracy, and significant computational savings. The research team's work represents an important step forward in developing more efficient LLM inference methods that can adapt to diverse models without requiring additional training.

Check out the Paper and GitHub Page . All credit for this research goes to the researchers of this project. Also, feel free to follow us on **Twitter** and don't forget to join our **95k+ ML SubReddit** and Subscribe to **our Newsletter**.

Sana Hassan

+ posts

Sana Hassan, a consulting intern at Marktechpost and dual-degree student at IIT Madras, is passionate about applying technology and AI to address real-world challenges. With a keen interest in solving practical problems, he brings a fresh perspective to the intersection of AI and real-life solutions.

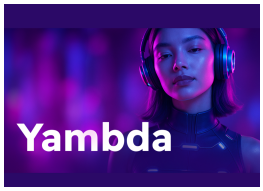
Previous article

Cisco's Latest AI Agents Report Details the Transformative Impact of Agentic AI on Customer Experience

Next article

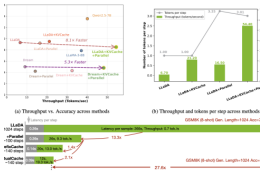
Meet NovelSeek: A Unified Multi-Agent Framework for Autonomous Scientific Research from Hypothesis Generation to Experimental Validation

RELATED ARTICLES



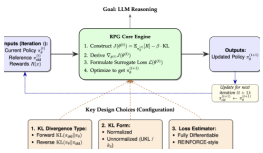
Meet Yambda: The World's Largest Event Dataset to Accelerate Recommender Systems

Asif Razzaq - June 2, 2025



NVIDIA AI Introduces Fast-dLLM: A Training-Free Framework That Brings KV Caching...

Nikhil - June 1, 2025



Off-Policy Reinforcement Learning RL with KL Divergence Yields Superior Reasoning in...

Sana Hassan - June 1, 2025

A Coding Implementation of an Intelligent AI Assistant with Jina Search,...

Asif Razzaq - June 1, 2025

Guide to Using the Desktop Commander MCP Server

Arham Islam - June 1, 2025

The Legal Accountability of AI-Generated Deepfakes in Election Misinformation

Aabis Islam - June 1, 2025

Marktechpost AI

Watch on

TECH POST

ABOUT US

Marktechpost is a California-based AI News Platform providing easy-to-consume, byte size updates in machine learning, deep learning, and data science research.

Contact us: Asif@marktechpost.com

FOLLOW US