MASSIVE ACTIVATION GATING CHANNEL IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Massive activations, a phenomenon in which a small number of hidden channels exhibit exceptionally large magnitudes, are pervasive in large language models (LLMs). However, the mechanism by which a token develops massive activations as it propagates through a pretrained LLM remains poorly understood. In this paper, we find that the emergence of massive activations is controlled by a single channel in the input embedding to a spike feed-forward network (FFN). The position of this channel is fixed for a particular LLM. We name this channel the massive activation gating channel (MAGC). When the value of the MAGC is sufficiently large (or small, depending on the LLM), the output of the spike FFN exhibits massive activations. Examining six LLMs across four model families and different model sizes, we verify the existence and effect of MAGC. We further provide a theoretical explanation of the mechanism by which MAGC induces massive activations. When the value of MAGC is sufficiently large (or small), the output of a spike FFN asymptotically reduces to a quadratic form that mixes a few columns of the down-projection matrix of the FFN. Since these columns exhibit the shape of massive activations, the output therefore exhibits massive activations.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated strong capabilities on various tasks (Bubeck et al., 2023). While the majority of existing work focuses on external behaviors of LLMs, it is also important to understand the internal mechanisms of LLMs. Among them, attention sink (Xiao et al., 2023) and massive activations (Sun et al., 2024) have been found in modern LLMs across different model families. Specifically, attention sink refers to the phenomenon in which the first token (or other tokens with little semantic importance) in a sequence receives a disproportionately large amount of attention from subsequent tokens (Yu et al., 2024; Barbero et al., 2025), typically in the middle transformer blocks. The cause of attention sink is usually attributed to massive activations (Sun et al., 2024; Yan et al., 2024), where a few channels (dimensions) in the hidden representation of the first token (or other tokens with little semantic importance) exhibit large activation magnitudes before attention computation. It is important to unveil the mechanism underlying massive activations and attention sink, as these phenomena have been widely employed in various critical tasks, including KV-cache optimization (Ge et al., 2023), inference (Xiao et al., 2023), quantization (Son et al., 2024), and hallucination mitigation (Liu et al., 2026).

In this study, we focus on massive activations. Although existing research has extensively demonstrated the phenomenon of massive activations, research on the mechanism underlying massive activations remains limited. We address this problem by asking: *how does a token develop massive activations as it propagates a pretrained transformer?* We refer to tokens that exhibit massive activations as spike tokens (Sun et al., 2026).

As LLMs are powered by billions of parameters, our analysis draws attention to the computation of each sub-module (attention and feed-forward networks) in transformers. Existing work (Zhang et al., 2026; Sun et al., 2026) has demonstrated that massive activations emerge after certain feed-forward networks (FFNs) and are maintained in the middle transformer blocks through the residual layer (see Section 2 for more details). For example, in LLaMA-2-7B, massive activations of a spike token emerge after the FFN in the second transformer block (Sun et al., 2024; 2026) and are maintained until reaching the last FFN. We refer to the FFN that produces massive activations as the spike FFN.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The magnitudes of massive activations remain similar in the middle transformer blocks, indicating that only the residual layer maintains massive activations and that no further massive activations are triggered during the middle computation.

Sun et al. (2026) assume that the SiLU activation function in the spike FFN behaves as a near-identity function for spike tokens. Under this assumption, each output channel k of the spike FFN can be expressed as a quadratic form characterized by a channel-specific coefficient matrix S_k. The magnitude of output channel k is determined by the dominant eigenvalue of S_k. Since the dominant eigenvalues associated with a few channels are substantially larger than those of other channels, massive activation emerges.

However, our empirical results show that the near-identity assumption on which Sun et al. (2026) rely to explain massive activations is only valid for a small number of models. Among the six models examined in this study, only one model approximately satisfies the assumption (see Appendix A for more details). Moreover, their explanation overlooks the properties of input embeddings to the spike FFN. Instead, we find that the emergence of massive activations is controlled by a single channel in the input embeddings of a spike token to the spike FFN: massive activations appear once the value of this channel is sufficiently large (or small, depending on the LLM). For a given LLM, the position of this channel is fixed. We refer to this channel as the massive activation gating channel (MAGC). Unlike massive activations, whose magnitudes can reach approximately 400 in LLaMA-3.2-1B, the MAGC in the embedding of a spike token has an average value of only around 8.8. Although this value is substantially larger than those of the other channels in the embedding, its magnitude remains moderate because the embedding is layer-normalized before being fed into the FFN.

We verify this finding across multiple decoder-only LLM families, including LLaMA, Qwen, DeepSeek, and Mistral, with model sizes ranging from 1B to 32B parameters. Moreover, through theoretical analysis, we show that massive activations can be attributed to the alignment between MAGC and certain rows and columns of the spike FFN weight matrices. Similarly, we empirically demonstrate that this alignment occurs in different LLMs. We summarize our contributions as follows.

1. The emergence of massive activations is controlled by a single channel in the input embedding to a spike FFN, and the position of this channel is fixed for a particular LLM. We name this channel the massive activation gating channel (MAGC). When the value of the MAGC is sufficiently large (or small, depending on the LLM), the output of the spike FFN exhibits massive activations. Examining six LLMs across four model families and different model sizes, we verify the existence and effect of MAGC.
2. We provide a theoretical explanation of the mechanism by which MAGC induces massive activations. When the value of MAGC is sufficiently large (or small), the output of a spike FFN asymptotically reduces to a quadratic form that mixes a few columns of the down-projection matrix of the FFN. Since these columns exhibit the shape of massive activations, the output therefore exhibits massive activations.

2 Transformer Architecture

A dense transformer model consists of L transformer blocks, each of which can be further decomposed into modules. Taking LLaMA as an example (Touvron et al., 2023), denote the input to the l-th transformer block as $X^l = \langle \boldsymbol{x}_0^l, \boldsymbol{x}_1^l, \cdots, \boldsymbol{x}_T^l \rangle$. Without further specification, $\cdot_t^l$ denotes an embedding of the t-th token in the l-th transformer block. Since our following analysis focuses on the single spike FFN, we omit the superscript l for notational simplicity. Each block of LLaMA consists of an attention sub-module and an FFN sub-module, each of which uses pre-normalization (RMSNorm (Zhang & Sennrich, 2019)) and includes a residual layer after its computation.

The input embedding is first normalized by an RMSNorm, referred to as input layer norm. Formally, if $\boldsymbol{x}_t = \langle x_t(0), x_t(1), \cdots, x_t(d-1) \rangle$, where d is the dimension of the embedding, then

$$\boldsymbol{n}_t = \text{RMSNorm}(\boldsymbol{x}_t) : n_t(j) = \frac{x_t(j) \cdot w_j}{\text{RMS}(\boldsymbol{x}_t)}, \text{RMS}(\boldsymbol{x}_t) = \sqrt{\frac{\sum_j x_t(j)^2}{d} + \epsilon}, \tag{1}$$

where w_j are trainable parameters and ϵ is a predefined small number. We denote $\cdot(j)$ as the j-th channel of the embedding. Next, the normalized embedding $\boldsymbol{n}_t$ is transformed by multi-head attention

(MHA, or more specifically, grouped-query attention (Ainslie et al., 2023)):

$$(\boldsymbol{u}_0, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_T) = \text{MHA}(\boldsymbol{n}_0, \boldsymbol{n}_1, \cdots, \boldsymbol{n}_T),$$

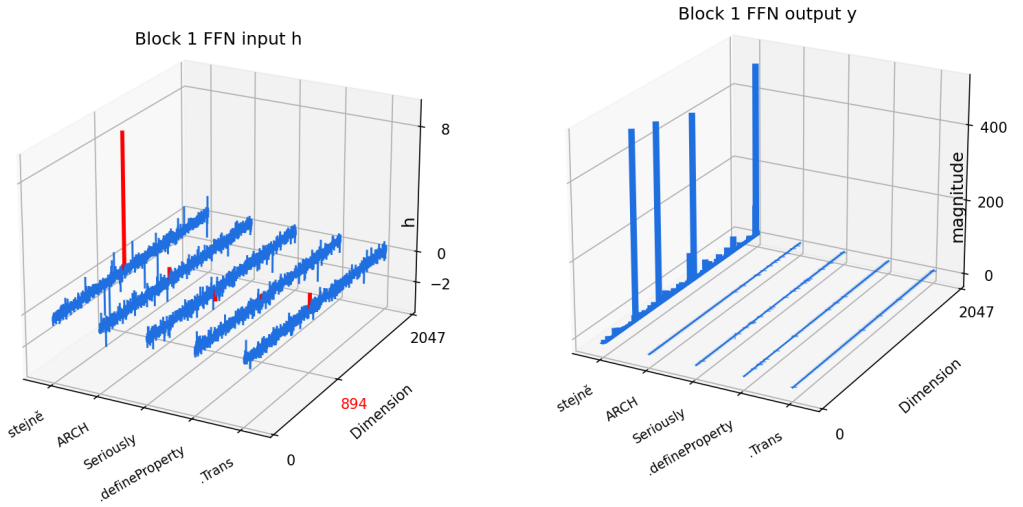
where $\boldsymbol{u}_t$ is the embedding output of MHA. Usually, a transformer includes a residual layer after the multi-head attention (Vaswani et al., 2017). Next, the FFN sub-module follows a similar configuration with another RMSNorm of different parameters.

$$\boldsymbol{h}_t = \text{RMSNorm}(\boldsymbol{x}_t + \boldsymbol{u}_t).$$

A feed-forward network (FFN) is then applied to $\boldsymbol{h}_t$. In LLaMA, the FFN consists of a SwiGLU gated activation function (Shazeer, 2020) with three weight matrices:

$$\boldsymbol{y}_t = \text{FFN}(\boldsymbol{h}_t), \boldsymbol{y}_t = \boldsymbol{W}_{\text{down}} \cdot (\text{SiLU}\,(\boldsymbol{W}_{\text{gate}}\boldsymbol{h}_t) \odot (\boldsymbol{W}_{\text{up}}\boldsymbol{h}_t))\,, \tag{2}$$

where $\odot$ is the element-wise product and $\boldsymbol{y}$ is the output of FFN. $\boldsymbol{W}_{\text{gate}}, \boldsymbol{W}_{\text{up}} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{W}_{\text{down}} \in \mathbb{R}^{d \times m}$ are pretrained weight matrices, where $m > d$ is the FFN intermediate dimension. Finally, another residual layer follows the FFN. That is, the output of a transformer block is $\boldsymbol{x}_t + \boldsymbol{u}_t + \boldsymbol{y}_t$.



(a) Value of input $\boldsymbol{h}$ to the spike FFN. We highlight the massive activation gating channel (894) in red.

(b) Magnitude of output $\boldsymbol{y}$ of the spike FFN. Token 0 exhibits massive activations. We depict magnitudes instead of actual values for better visualization.

Figure 1: Inputs and Outputs of the Spike FFN.

3 Massive Activation Gating Channel

Experimental Setup. In this section, we utilize a small but advanced LLM, LLaMA-3.2-1B (Dubey et al., 2024). We verify the findings discovered in this section on other model families with larger model sizes in Section 5.1. In LLaMA-3.2-1B, $d = 2,048$ and $m = 8,192$. We generate token sequences (without the BOS token) completely randomly to avoid the influence of semantics. Specifically, we randomly sample five tokens in the vocabulary set $\mathcal{V}$ and construct an input token sequence to the LLM. Note that the input tokens can be multilingual as well as completely meaningless. We conduct this sampling 5 times.

3.1 Existence of MAGC in the Input to Spike FFN

Following Sun et al. (2026), we demonstrate the top-3 channel magnitudes of the transformer block outputs $\boldsymbol{x} + \boldsymbol{u} + \boldsymbol{y}$ in Figure A1 in Appendix B. Similarly, we show that massive activations are first triggered by the FFN in an early layer, and are then preserved and propagated along the residual stream across layers. As depicted, we identify that in LLaMA-3.2-1B, the FFN in the second transformer

block (i.e., block 1) produces massive activations. We name this FFN the spike FFN and focus our analysis on it.

A straightforward method for analyzing the causes of massive activations is to trace their inputs. To this end, we directly visualize the embeddings of the inputs and outputs of the spike FFN for different tokens in Figure 1, using a token sequence example. We demonstrate other examples in Figure A2 in Appendix B.2. As shown in Figure 1b, the first token ("stejně") exhibits massive activations, whereas other tokens do not. Comparing their inputs to the FFN, we find that the input h for the first token has a single outlier value, as shown in Figure 1a. Specifically, the 894th channel of the first token has a value of 8.807, while the absolute values of the other channels are all less than 2. The 894th channel of the other tokens has values of 0.814, -0.763, 0.335, and 1.029, respectively. In the next subsection, we will show that this single channel controls the emergence of massive activations, and we name this channel the massive activation gating channel (MAGC).

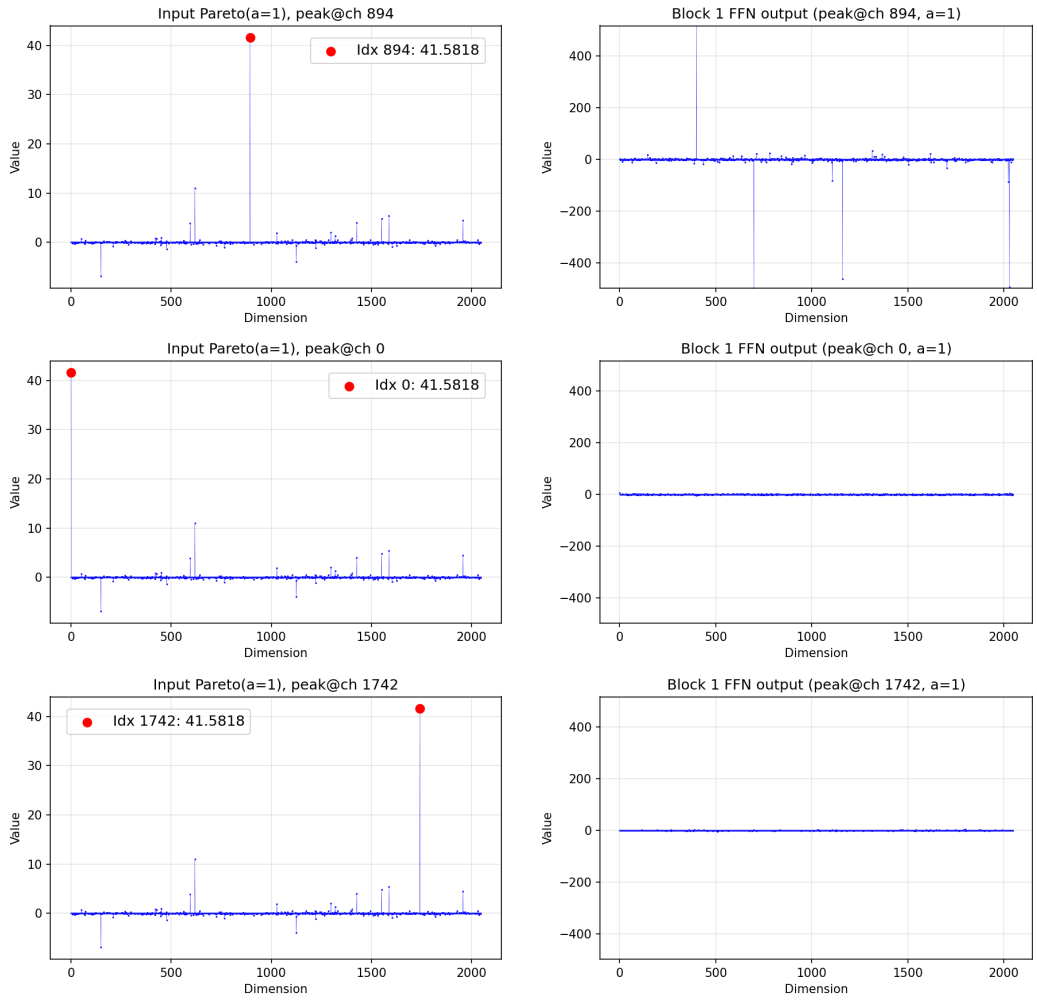


Figure 2: From top to bottom, the figure shows the output of the spike FFN when the input vector values are randomly sampled from a Pareto distribution ($a = 1$), with the input value of the 894th (MAGC), 0th, or 1742nd channel swapped with the maximum input value, respectively.

3.2 MAGC Triggers Massive Activations

Next, we show that this single-channel spike triggers massive activations through the computation of the spike FFN. To empirically validate this claim, we conduct the following experiment. Specifically, we randomly generate the input embedding to the spike FFN from a Pareto distribution with random signs: each channel value of the input embedding is i.i.d. drawn as $X = \text{random}(\pm)Z$ with

$Z \sim \text{Pareto}(a), Z \geq 0$, where each element takes $+Z$ and $-Z$ with equal probability 0.5, and a is a fixed shape parameter. Next, we swap the input value of the 894th channel with the maximum input value. We then input this randomly generated embedding into the spike FFN. Meanwhile, we conduct a controlled experiment. Specifically, we used the same randomly sampled input embedding but swapped the input value of a randomly selected channel other than the 894th, such as the 0th or 1742nd, with the maximum value. We report the results for $a = 1$ in this section and present the results for $a = 3$ and $a = 5$ in Appendix B.3.

Figure 2 depicts the randomly generated inputs and their corresponding outputs after the spike FFN. It can be observed that when swapping the value of the 894th channel with the maximum value, the outputs of the spike FFN demonstrate massive activations. Since the randomly generated inputs carry no semantic meaning, this phenomenon mainly indicates that the FFN learns to recognize MAGC. However, swapping the value of any other channel, such the 0th or 1742nd channel, with the maximum value does not lead to the output of massive activations. We therefore conclude that MAGC controls the emergence of massive activations.

Moreover, as shown in Figure A3 and A4 in Appendix B.3, decreasing a increases the maximum value generated by the Pareto distribution. More importantly, assigning a larger maximum value to the 894th channel of the spike FFN input results in a correspondingly larger activation magnitude in its output. Specifically, as a decreases from 5 to 1, the largest output activation magnitude increases from approximately 40 to 400. These results demonstrate that the spike FFN is highly sensitive to the value of the 894th input channel, with a larger value in this channel inducing more "massive" activations.

Key Takeaway 1: There exists a massive activation gating channel (MAGC) in the hidden representation of a spike token, whose position is fixed for a specific LLM. When the hidden representation is used as the input to the spike FFN, the FFN output exhibits massive activations when the value of the MAGC in the hidden representation is sufficiently large (or small, depending on the LLM).

4 From MAGC to Massive Activations

In this section, we examine how a large value of MAGC induces massive activations in LLMs. Since a large value of MAGC corresponds to a large magnitude in a certain channel of the input to the FFN, we begin by analyzing the asymptotic property of the FFN with respect to any input channel. Recall the FFN computation in Equation 2: $\boldsymbol{y}(\boldsymbol{h}) = \boldsymbol{W}_{\text{down}} \left(\text{SiLU}(\boldsymbol{W}_{\text{gate}}\boldsymbol{h}) \odot \boldsymbol{W}_{\text{up}}\boldsymbol{h}\right)$. We show that, when a fixed channel has a large magnitude (in either direction), the mechanism of the FFN can be simplified as mixing columns of $\boldsymbol{W}_{\text{down}}$. Note that the mixing mechanism depends on the direction of the channel value ($x \to +\infty$ or $x \to -\infty$), which explains why MAGC also has a fixed gating direction.

4.1 Asymptotic Property of an FFN w.r.t. an Input Channel

Figure 1a demonstrates the phenomenon that the input to the FFN contains a large magnitude in a certain channel for spike tokens. We begin by analyzing the asymptotic property of any FFN with respect to any input channel. Consider a gated FFN in Equation 2, for input $\boldsymbol{h} \in \mathbb{R}^d$,

$$\text{FFN}(\boldsymbol{h}) = \boldsymbol{y}(\boldsymbol{h}) = \boldsymbol{W}_{\text{down}} \left(\text{SiLU}(\boldsymbol{W}_{\text{gate}}\boldsymbol{h}) \odot \boldsymbol{W}_{\text{up}}\boldsymbol{h}\right). \tag{3}$$

where $\boldsymbol{W}_{\text{up}} \in \mathbb{R}^{m\times d}$, $\boldsymbol{W}_{\text{gate}} \in \mathbb{R}^{m\times d}$, and $\boldsymbol{W}_{\text{down}} \in \mathbb{R}^{d\times m}$ are parameter matrices, $\text{SiLU}(x) := x \cdot \sigma(x)$ with the logistic sigmoid function $\sigma(x) := \frac{1}{1+\exp(-x)}$, and $\odot$ is the element-wise product.

We focus on the input $\boldsymbol{h}$ controlled by a fixed single channel i. For any value $x \in \mathbb{R}$ in the fixed channel i, the input can be written as,

$$\boldsymbol{h}(x) = x \cdot \boldsymbol{e}_i + \bar{\boldsymbol{h}}. \tag{4}$$

Here, $\bar{\boldsymbol{h}} \in \mathbb{R}^d$ is the representation outside the focused channel i with $\bar{\boldsymbol{h}}(i) = 0$, and $\boldsymbol{e}_i$ is the unit vector at position i. Next, we analyze the asymptotic property of $\boldsymbol{y}(\boldsymbol{h}(x))$ for $x \to +\infty$ without

loss of generality. Let $\boldsymbol{\omega}_j = \boldsymbol{W}_{\text{down}}^{(:,j)} / \left\| \boldsymbol{W}_{\text{down}}^{(:,j)} \right\|_2$ be the normalized j-th column of $\boldsymbol{W}_{\text{down}}$. The following theorem shows the closed form of $\boldsymbol{y}(\boldsymbol{h}(x))$ when $x \to +\infty$.

Theorem 4.1. *Given input $\boldsymbol{h}(x) = x \cdot \boldsymbol{e}_i + \bar{\boldsymbol{h}}$, when $x \to +\infty$, the output of an FFN has a quadratic form $\boldsymbol{y}(\boldsymbol{h}(x)) = x^2 \cdot \boldsymbol{y}^* + o(x^2)$ with constant $\boldsymbol{y}^* \in \mathbb{R}^d$ equal to*

$$\boldsymbol{y}^* = \sum_{j=1}^{m} \alpha_j \cdot \boldsymbol{\omega}_j, \tag{5}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ is the combination of coefficients in the weight matrices:

$$\alpha_j = \begin{cases} W_{\text{gate}}^{(j,i)} W_{\text{up}}^{(j,i)} \left\| \boldsymbol{W}_{\text{down}}^{(:,j)} \right\|_2 & W_{\text{gate}}^{(j,i)} > 0 \\ 0 & W_{\text{gate}}^{(j,i)} \le 0 \end{cases}. \tag{6}$$

Proof. See Appendix C.1.

Specifically, for any input channel i in $\boldsymbol{h}$, $\text{SiLU}(\boldsymbol{W}_{\text{gate}}\boldsymbol{h})$ acts as a gating function, choosing column j only when $W_{\text{gate}}^{(j,i)} x > 0$, and α_j controls the sensitivity to each $\boldsymbol{\omega}_j$. From Theorem 4.1, when x is large enough, the output $\boldsymbol{y}(\boldsymbol{h}(x)) = x^2 \cdot \boldsymbol{y}^* + o(x^2)$ exhibits a pattern of massive activations only if $\boldsymbol{y}^*$ has the same pattern. Similarly, an analogous result holds as $x \to -\infty$.

4.2 From MAGC to Massive Activations

We introduce several metrics below to quantify the degree of massive activations. Since massive activations describe the phenomenon in which certain activations exhibit large magnitudes, we utilize the energy distribution to characterize their magnitude distribution.

Definition 4.2 (Energy Distribution)**.** Given vector $\boldsymbol{v} = (v_1, \ldots, v_d) \in \mathbb{R}^d$, let its energy distribution be $\boldsymbol{p}(\boldsymbol{v}) = (p_1(\boldsymbol{v}), \ldots, p_d(\boldsymbol{v}))$, where

$$p_i(\boldsymbol{v}) = \frac{v_i^2}{\|\boldsymbol{v}\|_2^2}, \quad \forall i = 1, \ldots, d. \tag{7}$$

The energy distribution is a probability distribution. Moreover, if $\boldsymbol{y}$ exhibits massive activations, $\boldsymbol{p}(\boldsymbol{y})$ should be highly concentrated. We define the degree of concentration as the negative entropy of $\boldsymbol{p}(\boldsymbol{y})$ plus a constant. Formally,

Definition 4.3 (Concentration Degree)**.** Given vector $\boldsymbol{v} \in \mathbb{R}^d$, let its concentration degree be

$$R(\boldsymbol{v}) = \log d + \sum_{i=1}^{d} p_i(\boldsymbol{v}) \log p_i(\boldsymbol{v}) \tag{8}$$

where $\boldsymbol{p}(\boldsymbol{v}) = (p_1(\boldsymbol{v}), \ldots, p_d(\boldsymbol{v}))$ is the energy distribution defined in Definition 4.2.

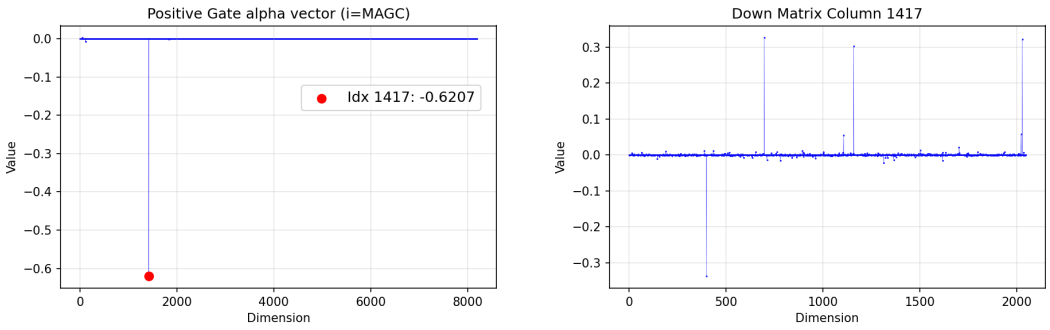
The concentration degree $R(\boldsymbol{v}) \in [0, \log d]$ is exactly the negative entropy plus a constant $\log d$. A larger $R(\boldsymbol{v})$ means $\boldsymbol{v}$ is more concentrated. $R(\boldsymbol{v}) = 0$ means $\boldsymbol{v}$ is uniform, while $R(\boldsymbol{v}) = \log d$ means $\boldsymbol{v}$ has only one non-zero entry.

We now fix MAGC as the focused channel i^* and analyze the relationship between the concentration degree of the output $\boldsymbol{y}^*$ and those of $\boldsymbol{\omega}_j$ and $\boldsymbol{\alpha}$. We first find that the following assumption is empirically satisfied.

Assumption 4.4. When MAGC is the focused channel, $\boldsymbol{\alpha}$ has a high concentration degree, i.e., $R(\boldsymbol{\alpha}) \ge \log m - \epsilon$ for a small ϵ. Let $j^* = \arg\max_j p_j(\boldsymbol{\alpha})$ be the dominant column index; then

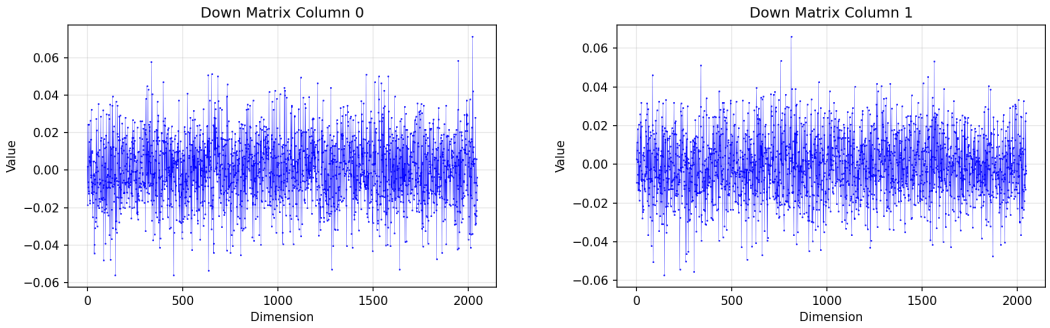
$$R(\boldsymbol{\alpha}) \ge \log m - \epsilon \Rightarrow p_{j^*}(\boldsymbol{\alpha}) \ge e^{-\epsilon}. \tag{9}$$

Empirical verification. Rather than assuming this property, we verify it directly by scanning all $d = 2{,}048$ input channels of the spike FFN in LLaMA-3.2-1B: for each channel i we form the coefficient vector $\boldsymbol{\alpha}$ from Theorem 4.1 and measure its concentration $R(\boldsymbol{\alpha})$, dominant energy share $p_{j^*}(\boldsymbol{\alpha})$, and norm $\|\boldsymbol{\alpha}\|_2$. MAGC ($i^* = 894$) attains the largest concentration among all channels,



(a) $\boldsymbol{\alpha}$ vector when $i^* = 894$ for LLaMA-3.2-1B.

(b) $\boldsymbol{\omega}_{1417}$ for LLaMA-3.2-1B.



(c) $\boldsymbol{\omega}_0$ for LLaMA-3.2-1B.

(d) $\boldsymbol{\omega}_1$ for LLaMA-3.2-1B.

Figure 3: Visualization results for LLaMA-3.2-1B.

with $R(\boldsymbol{\alpha}) = 9.009$ against $\log m = 9.011$ (i.e. $\epsilon \approx 1.9 \times 10^{-3}$) and $p_{j^*}(\boldsymbol{\alpha}) = 0.9998$, whereas the mean over all channels is only $R(\boldsymbol{\alpha}) = 2.908$. This confirms that MAGC is a significant outlier in $\boldsymbol{\alpha}$-concentration.

A high concentration degree $R(\boldsymbol{\alpha})$ indicates that a few $p_j(\boldsymbol{\alpha})$ are extremely large. We empirically demonstrate $\boldsymbol{\alpha}$ when $i^* = 894$ for LLaMA-3.2-1B in Figure 3a and highlight the largest magnitude in the vector. As shown, $\boldsymbol{\alpha}$ is highly concentrated with $j^* = 1417$. That is, the output $\boldsymbol{y}$ of the FFN is largely controlled by the j^*-th column of $\boldsymbol{W}_{\text{down}}$. Theoretically, based on the high concentration of $\boldsymbol{\alpha}$, we next quantify the contribution of the remaining columns.

Definition 4.5. Given $j^* = \arg\max_j p_j(\boldsymbol{\alpha})$, define the residual amplification factor

$$\kappa = \frac{\left\| \sum_{j \neq j^*} \alpha_j \boldsymbol{\omega}_j \right\|_2}{\sqrt{\sum_{j \neq j^*} \alpha_j^2}}.$$

The residual amplification factor κ quantifies the contribution of other channels after removing the dominant column j^*. In the following proposition, κ controls how strongly the non-dominant columns can perturb the energy distribution of the output.

Proposition 4.6. *Let*

$$\tau = \kappa \sqrt{\frac{1 - p_{j^*}(\boldsymbol{\alpha})}{p_{j^*}(\boldsymbol{\alpha})}}. \tag{10}$$

We have

$$|R(\boldsymbol{y}^*) - R(\boldsymbol{\omega}_{j^*})| \leq \Delta\left(|\sin \angle(\boldsymbol{y}^*, \boldsymbol{\omega}_{j^*})|\right) \leq \Delta(\tau), \tag{11}$$

where for $t \in \left(0, 1 - \frac{1}{d}\right)$, the function Δ is defined as $\Delta(t) = -t \log t - (1-t)\log(1-t) + t \log(d-1)$.

Proof. See Appendix C.2.

Proposition 4.6 indicates that when $p_{j^*}(\boldsymbol{\alpha})$ is large (i.e., close to 1), we have $\tau \approx 0$, and there is almost no gap between the concentration degree of $\boldsymbol{y}^*$ and that of $\boldsymbol{\omega}_{j^*}$.

In the case of LLaMA-3.2-1B, the output $\boldsymbol{y}$ of the FFN is largely controlled by the j^*-th (1417th) column of $\boldsymbol{W}_{\text{down}}$. We next depict $\boldsymbol{\omega}_{j^*}$ in Figure 3b. Comparing Figure 3b to 1b, we find that these two vectors are highly co-linear, empirically with $|\sin\angle(\boldsymbol{y}^*, \boldsymbol{\omega}_{j^*})| = 0.0055$. For other columns such as columns 0 and 1 in Figures 3c and 3d, the parameters have lower maximum magnitudes and do not contain the pattern of massive activations. Note that both $\boldsymbol{y}^*$ and $\boldsymbol{\omega}_{j^*}$ are computed only from the pretrained weight matrices. We show that when the value of channel i^* is large enough, the actual output $\boldsymbol{y}$ is dominated by $\boldsymbol{\omega}_{j^*}$, which leads to $\boldsymbol{y} \approx x^2\boldsymbol{y}^* \approx x^2\alpha_{j^*}\boldsymbol{\omega}_{j^*}$. If α_{j^*} has a large magnitude and $\boldsymbol{\omega}_{j^*}$ is highly concentrated (i.e., contains the pattern of massive activations), the output $\boldsymbol{y}$ should exhibit the same pattern. Below, we summarize the mechanism by which the MAGC in the input to a spike FFN leads to its output that exhibits massive activations.

Key Takeaway 2: When the MAGC of an input embedding to the spike FFN takes a sufficiently large (or small) value, the output of the FFN asymptotically approaches $x^2 \sum_{j=1}^{m} \alpha_j \boldsymbol{\omega}_j$, where $\boldsymbol{\omega}_j$ is the normalized j-th column of $\boldsymbol{W}_{\text{down}}$ of the FFN and x is the value of the MAGC. Moreover, the coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ concentrate on one (or a few) dimensions. As a result, the FFN output is reduced to one normalized column of $\boldsymbol{W}_{\text{down}}$ or a combination of a few normalized columns of $\boldsymbol{W}_{\text{down}}$, multiplied by their respective coefficient and x^2. Since these columns exhibit the shape of massive activations, the output therefore exhibits massive activations.

5 Extension

5.1 Extension to Other LLMs

Model Name	Spike FFN Block	MAGC(i^*)	Direction	Concentrated Columns(J)
LLaMA-3.2-1B	1	894	+	1417
LLaMA-2-7B	1	310	+	10411, 7890
Qwen2.5-1.5B	1	1395	-	5601, 3295, 3750
QwQ-32B	4	4626	-	11696, 1368
Ministral-3-3B	2	2	-	0, 2
Deepseek-R1-8B	1	717	+	2427, 198

Table 1: MAGC in different models. The spike FFN block, MAGC i^*, and concentrated column ids are counted from 0. The concentrated columns are identified by first selecting the three columns ω_j of the down-projection matrix with the largest absolute α_j values and then filtering out the selected columns that do not satisfy $|\cos(\boldsymbol{\omega}_j, \boldsymbol{y}_0)| > 0.9$.

In this section, we examine the generality of our findings across LLMs of different model families and sizes. Specifically, we study LLaMA-2-7B, Qwen2.5-1.5B, QwQ-32B, Ministral-3-3B-Instruct-2512, and DeepSeek-R1-Distill-Llama-8B. Together with LLaMA-3.2-1B, these models cover four model families and range from 1B to 32B parameters. Based on our empirical findings, we propose an empirical method to identify the MAGC as detailed in Appendix D. For each model, we repeat the procedure in the preceding sections: we localize the spike FFN, identify MAGC i^* and its gating direction, and find the dominant column(s) J of $\boldsymbol{W}_{\text{down}}$. Table 1 summarizes the results, and the corresponding visualizations are provided in Appendix D.

We highlight the following findings. First, the spike FFN consistently occurs near the beginning of the network. Notably, even for QwQ-32B with 64 transformer blocks, the spike FFN is located at block 4. Second, the position i^* of MAGC is fixed for a specific model but differs across models, as does the gating direction. Third, the dominant column is not always unique. For LLaMA-3.2-1B, the energy distribution $p_j(\boldsymbol{\alpha})$ is dominated by a single column of $\boldsymbol{W}_{\text{down}}$. For the remaining models, however, two or three columns share the dominant energy. As shown in Appendix D, the columns in J exhibit highly similar shapes and concentrate their mass on the same small set of output channels, so their

weighted combination $\boldsymbol{y}^*$ remains concentrated and reproduces the pattern of massive activations. In contrast, a randomly selected column distributes its mass much more broadly across channels.

5.2 Relation to Attention Sink

Attention sink is commonly attributed to massive activations (Sun et al., 2024; Yan et al., 2024). Our previous results show that massive activations can be controlled by MAGC. We now extend this finding to adjust attention sink using MAGC.

Experimental Setup. We feed ten natural sentences to LLaMA-3.2-1B and intervene only on MAGC of the first token: at the input to the spike FFN (block 1), we overwrite $h_0(i^*)$ with a prescribed value c and leave all other channels and all other tokens unchanged. We then measure the attention sink of the first token in the following block (block 2) using the sink score (Gu et al., 2025),

$$s_0 = \frac{1}{H(T+1)} \sum_{h=0}^{H-1} \sum_{t=0}^{T} A^h_{t,0}, \qquad (12)$$

where $A^h_{t,0}$ denotes the post-softmax attention that the t-th query assigns to the first token in the h-th head, $T+1$ is the sequence length, and H is the number of attention heads. Since the first token is visible to all queries under the causal mask, s_0 is the average fraction of attention that the first token receives, and s_0 close to 1 indicates a dominant attention sink.

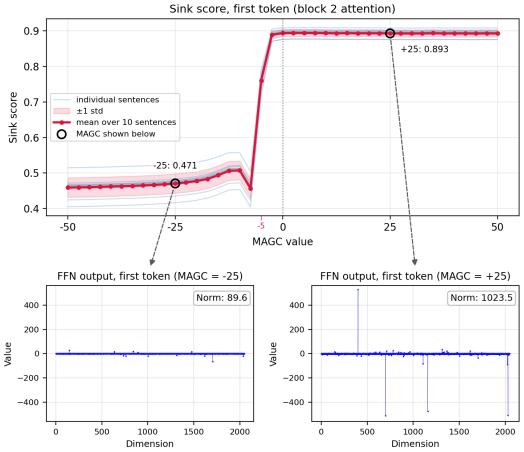


Figure 4: Sink score of the first token in LLaMA-3.2-1B, measured in the attention of block 2, as a function of the value assigned to MAGC ($i^* = 894$) of the first token at the input of the spike FFN (block 1). Thin curves correspond to the ten individual sentences, the thick curve to their mean, and the shaded area to one standard deviation. The unmodified average value of MAGC is 8.81.

Results. Figure 4 reports the sink score s_0 as a function of MAGC value c. Without intervention, the original MAGC value is 8.81, at which the first token attracts $s_0 = 0.89$ of the attention mass. As c varies, the sink score shows a sharp step-like transition. For $c \geq -2.5$, s_0 stays near 0.89, together with massive activations, as shown at $c = +25$. For sufficiently negative c, the sink collapses. The sink score drops to about 0.46 for $c \leq -7.5$, while the massive activations disappear, as shown at $c = -25$. The transition is consistent across sentences, with standard deviation below 0.03 throughout the sweep. Therefore, adjusting the MAGC value in the embedding of an input spike token to the spike FFN can adjust the sink mass on the token.

6 Conclusion and Future Work

In this study, we show how tokens develop massive activations in pretrained transformers and identify the massive activation gating channel (MAGC): a channel in the input to an early FFN whose value gates the emergence of massive activations. We show both empirically and theoretically that once MAGC is large or small enough, the spike FFN asymptotically collapses into a quadratic mixing of a few highly concentrated columns of the down-projection matrix, which reproduces the massive-activation pattern. This finding holds across six LLMs from four families ranging from 1B to 32B parameters.

Our analysis has several limitations that suggest directions for future work. First, it focuses on decoder-only LLMs with gated (such as SwiGLU) FFNs; extending the framework to mixture-of-experts architectures is a natural next step. Second, we localize a single spike FFN and a single gating channel per model, whereas a few models exhibit more diffuse behavior across several channels or columns; a multi-channel account may be needed for such cases. Finally, our study is purely mechanistic at inference time and does not explain why a particular channel becomes MAGC during pretraining, which we leave for future investigation.

AI USE STATEMENT

In this work, we used generative AI tools for paper proofreading and code writing. We have not used generative AI tools for literature review, idea formalization, and drafting the manuscript. We have reviewed all AI-assisted work. For example, LLM-generated code was verified by the authors. We take responsibility for the originality of this work.

REPRODUCIBILITY STATEMENT

We will release all code needed to reproduce the results in this work. It includes scripts for identifying the MAGC, running the intervention experiments, and generating all figures. All experiments use publicly available pretrained models.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, 2023.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael M Bronstein. Why do llms attend to the first token? In *Second Conference on Language Modeling*, 2025.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive kv cache compression for llms. *arXiv preprint arXiv:2310.01801*, 2023.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 2025.

Xu Liu, Guikun Chen, and Wenguan Wang. Sinktrack: Attention sink based context anchoring for large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeun Kim, and Jaeho Lee. Prefixing attention sinks can mitigate activation outliers for large language model quantization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2242–2252, 2024.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks. *arXiv preprint arXiv:2603.05498*, 2026.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Ruiqing Yan, Xingbo Du, Haoyu Deng, Linghan Zheng, Qiuzhuang Sun, Jifang Hu, Yuhang Shao, Penghao Jiang, Jinrong Jiang, and Lian Zhao. Unveiling and controlling anomalous attention distribution in transformers. *arXiv preprint arXiv:2407.01601*, 2024.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*, 2024.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. Attention sinks: A'catch, tag, release'mechanism for embeddings. *Advances in Neural Information Processing Systems*, 38:83140–83181, 2026.

A Generalization of Near Identity Assumption for SiLU

Following Sun et al. (2026), we compute the cosine similarity between the gate pre-activation $\boldsymbol{W}_{\text{gate}}\boldsymbol{h}_t$ and its activation $\text{SiLU}(\boldsymbol{W}_{\text{gate}}\boldsymbol{h}_t)$, and the ratio of their ℓ_2 norms for spike tokens. Using the randomly generated sequences described in Section 3, we report in Table A1 both metrics averaged over the first tokens of these sequences. For SiLU to behave approximately as an identity function, both metrics should be close to 1. Among the models examined, this condition is approximately satisfied only by LLaMA-2-7B.

Model	Metric	Token 0
LLaMA-2-7B	Cosine sim. Norm ratio	0.9719 0.8103
LLaMA-3.2-1B	Cosine sim. Norm ratio	0.8623 0.6232
Qwen2.5-1.5B	Cosine sim. Norm ratio	0.2276 0.1564
QwQ-32B	Cosine sim. Norm ratio	0.2990 0.1186
Ministral-3-3B	Cosine sim. Norm ratio	0.7895 0.7897
Deepseek-R1-8B	Cosine sim. Norm ratio	0.9215 0.7720

Table A1: Input-output characteristics of SiLU in the spike FFN across models.

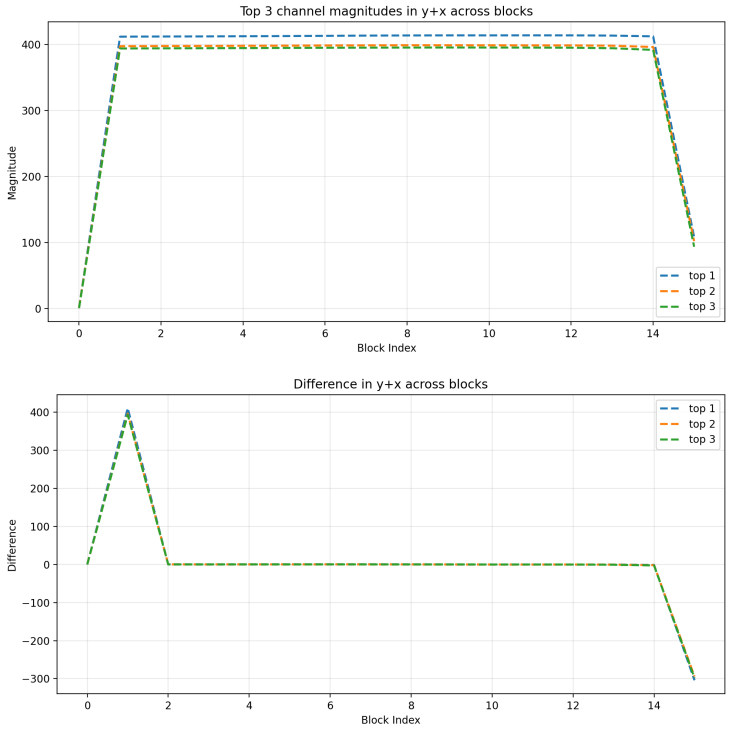


Figure A1: Top-3 channel magnitudes of transformer block outputs $x + u + y$ in Llama-3.2-1B and their differences.

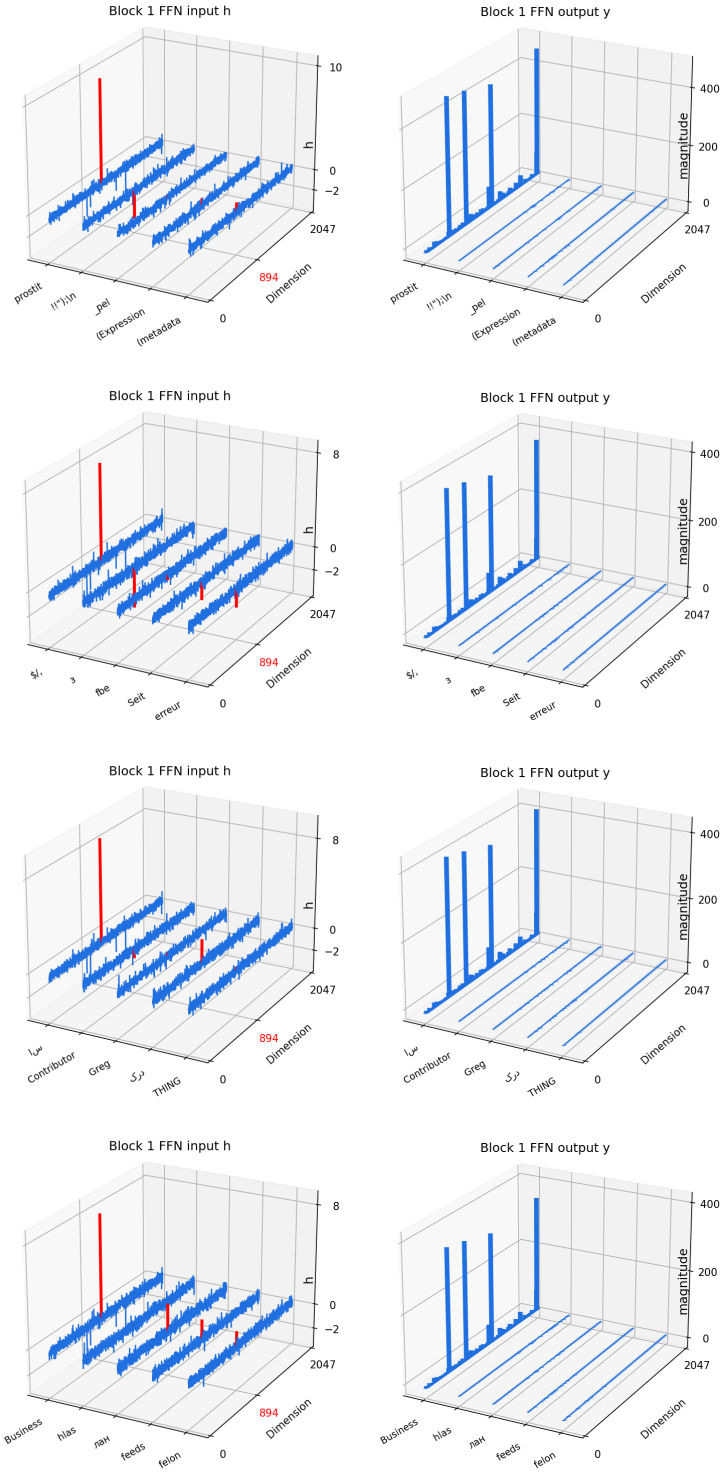


Figure A2: Inputs and Outputs of the Spike FFN.

B Additional Experiments for LLaMA-3.2-1B

B.1 Massive Activations

Figure A1 shows the block-wise differences of the top-3 channel magnitudes of the transformer block outputs $\boldsymbol{x} + \boldsymbol{u} + \boldsymbol{y}$ in LLaMA-3.2-1B. As depicted, the differences exhibit a single sharp jump at the second transformer block, where the spike FFN triggers massive activations. In the subsequent middle blocks, the differences remain close to zero, indicating that no further massive activations are triggered and that the residual stream alone preserves the existing ones until the late blocks.

B.2 Input and Output of the Spike FFN for other Sample Sequences

Following Section 3, we depict the activations of the other 4 randomly generated sequences in Figure A2.

B.3 Pareto Inputs for the Spike FFN

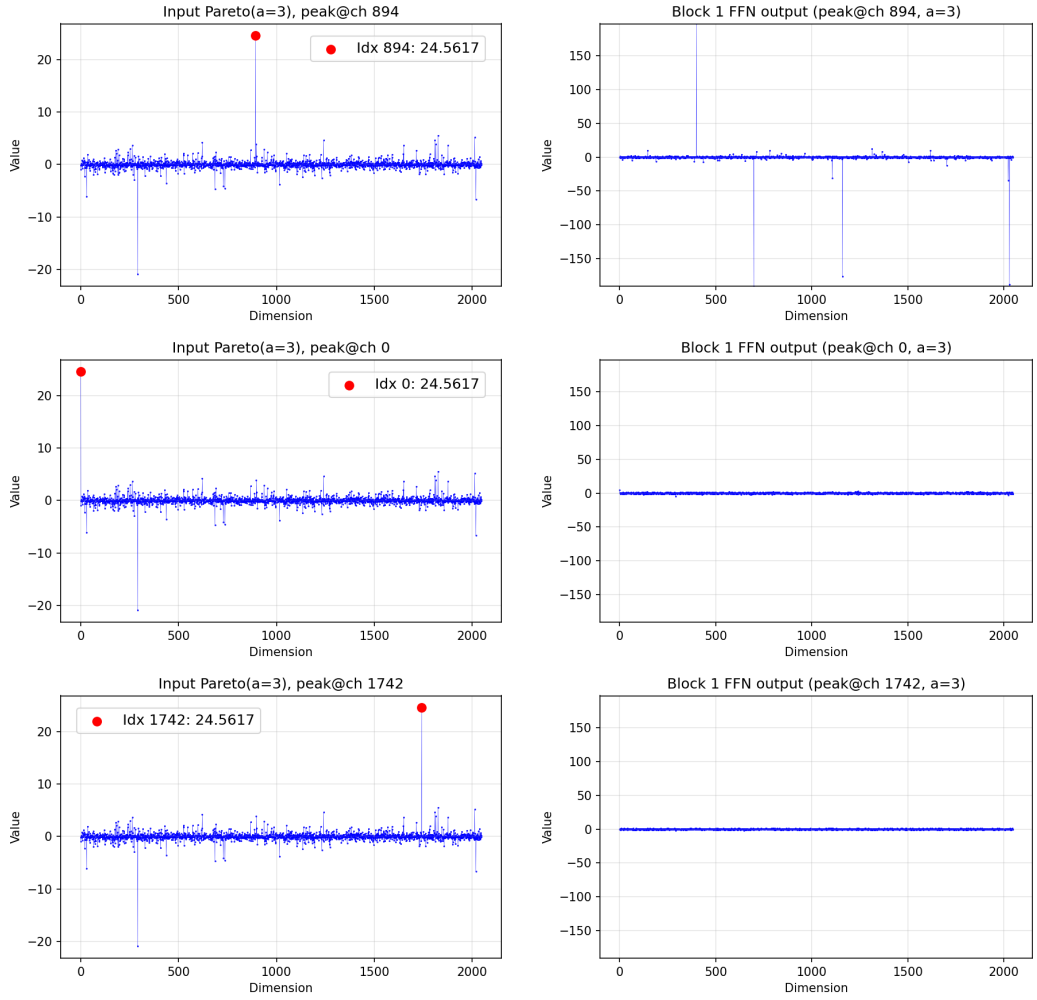


Figure A3: From top to bottom, the figure shows the output of the spike FFN when the input vector values are randomly sampled from a Pareto distribution ($a = 3$), with the input value of the 894th (MAGC), 0th, or 1742nd channel swapped with the maximum input value, respectively.

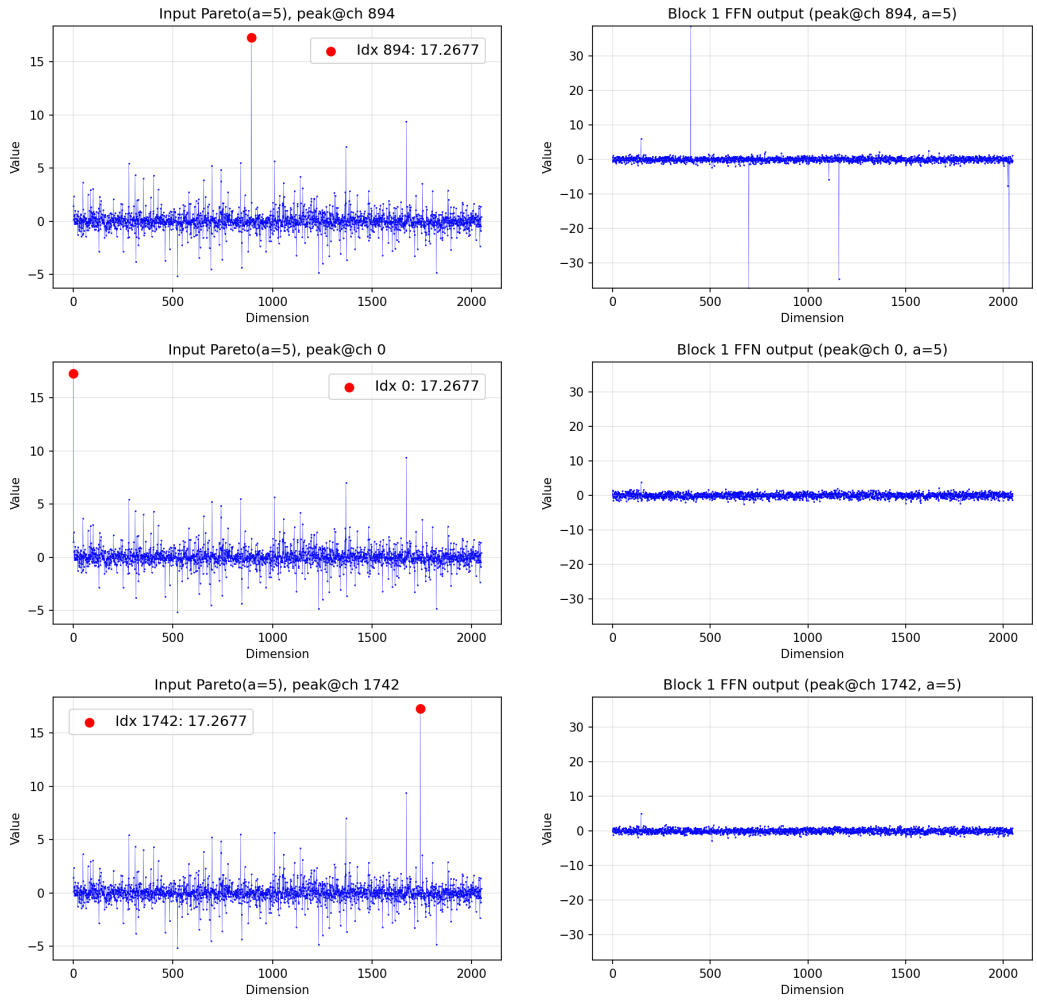


Figure A4: From top to bottom, the figure shows the output of the spike FFN when the input vector values are randomly sampled from a Pareto distribution ($a = 5$), with the input value of the 894th (MAGC), 0th, or 1742nd channel swapped with the maximum input value, respectively.

C Proofs

C.1 Proof of Theorem 4.1

Theorem C.1. *Given input $\boldsymbol{h}(x) = x \cdot \boldsymbol{e}_i + \bar{\boldsymbol{h}}$. When $x \to +\infty$, the output of an FFN has a quadratic form $\boldsymbol{y}(\boldsymbol{h}(x)) = x^2 \cdot \boldsymbol{y}^* + o(x^2)$ with constant $\boldsymbol{y}^* \in \mathbb{R}^d$ equals to*

$$\boldsymbol{y}^* = \sum_{j=1}^{m} \alpha_j \cdot \boldsymbol{\omega}_j, \tag{A1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ is the combination of coefficients in the weight matrices:

$$\alpha_j = \begin{cases} W_{\text{gate}}^{(j,i)} W_{\text{up}}^{(j,i)} \left\| \boldsymbol{W}_{\text{down}}^{(:,j)} \right\|_2 & W_{\text{gate}}^{(j,i)} > 0 \\ 0 & W_{\text{gate}}^{(j,i)} \le 0 \end{cases}. \tag{A2}$$

Proof. Based on Equation 4, the linear representation of $\boldsymbol{W}_{\text{gate}}\boldsymbol{h}(x)$ and $\boldsymbol{W}_{\text{up}}\boldsymbol{h}(x)$ can be written as

$$\boldsymbol{W}_{\text{gate}}\boldsymbol{h}(x) = x\boldsymbol{W}_{\text{gate}}^{(:,i)} + \boldsymbol{b}_{\text{gate}} \tag{A3}$$

$$\boldsymbol{W}_{\text{up}}\boldsymbol{h}(x) = x\boldsymbol{W}_{\text{up}}^{(:,i)} + \boldsymbol{b}_{\text{up}} \tag{A4}$$

where $\boldsymbol{b}_{\text{gate}} = \sum_{i=1}^{d} \bar{h}_i \boldsymbol{W}_{\text{gate}}^{(:,i)}$ and $\boldsymbol{b}_{\text{up}} = \sum_{i=1}^{d} \bar{h}_i \boldsymbol{W}_{\text{up}}^{(:,i)}$ are both constant under x.

We have the asymptotic property of $\text{SiLU}(x)$:

$$\lim_{x \to +\infty} \frac{\text{SiLU}(\boldsymbol{W}_{\text{gate}} h(x))}{x} = \begin{cases} W_{\text{gate}}^{(j,i)} & W_{\text{gate}}^{(j,i)} \geq 0 \\ 0 & W_{\text{gate}}^{(j,i)} < 0 \end{cases} \tag{A5}$$

and the asymptotic property of $\boldsymbol{W}_{\text{up}} \boldsymbol{h}(x)$:

$$\lim_{x \to +\infty} \frac{\boldsymbol{W}_{\text{up}} \boldsymbol{h}(x)}{x} = \boldsymbol{W}_{\text{up}}^{(:,i)} \tag{A6}$$

Finally,

$$\lim_{x \to +\infty} \frac{\boldsymbol{y}(\boldsymbol{h}(x))}{x^2} = \boldsymbol{W}_{\text{down}} \lim_{x \to +\infty} \frac{\text{SiLU}(\boldsymbol{W}_{\text{gate}} h(x))}{x} \odot \lim_{x \to +\infty} \frac{\boldsymbol{W}_{\text{up}} \boldsymbol{h}(x)}{x} = \sum_{j=1}^{m} \alpha_j \cdot \boldsymbol{\omega}_j, \tag{A7}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ is

$$\alpha_j = \begin{cases} W_{\text{gate}}^{(j,i)} W_{\text{up}}^{(j,i)} \left\| \boldsymbol{W}_{\text{down}}^{(:,j)} \right\|_2 & W_{\text{gate}}^{(j,i)} > 0 \\ 0 & W_{\text{gate}}^{(j,i)} \leq 0 \end{cases}. \tag{A8}$$

□

C.2 Proof of Proposition 4.6

Recall that

$$\boldsymbol{y}^* = \sum_{j=1}^{m} \alpha_j \boldsymbol{\omega}_j, \qquad \|\boldsymbol{\omega}_j\|_2 = 1.$$

The energy distribution of $\boldsymbol{\alpha}$ is

$$p_j(\boldsymbol{\alpha}) = \frac{\alpha_j^2}{\|\boldsymbol{\alpha}\|_2^2}, \qquad j = 1, \ldots, m.$$

Let $j^* = \arg\max_j p_j(\boldsymbol{\alpha})$ be the dominant coefficient index. Then $\boldsymbol{y}^*$ can be decomposed into the dominant column and the residual part:

$$\boldsymbol{y}^* = \alpha_{j^*} \boldsymbol{\omega}_{j^*} + \sum_{j \neq j^*} \alpha_j \boldsymbol{\omega}_j.$$

Define the residual vector

$$\boldsymbol{r} = \sum_{j \neq j^*} \alpha_j \boldsymbol{\omega}_j.$$

The residual amplification factor is defined as in Definition 4.5,

$$\kappa = \frac{\|\boldsymbol{r}\|_2}{\sqrt{\sum_{j \neq j^*} \alpha_j^2}}.$$

The angular deviation between the true asymptotic output and the dominant column is

$$|\sin \angle(\boldsymbol{y}^*, \boldsymbol{\omega}_{j^*})| = \sqrt{1 - \left(\frac{\boldsymbol{y}^* \cdot \boldsymbol{\omega}_{j^*}}{\|\boldsymbol{y}^*\|_2 \|\boldsymbol{\omega}_{j^*}\|_2} \right)^2}.$$

Finally, for $\tau \in (0, 1 - \frac{1}{d})$, define

$$\Delta(\tau) = -\tau \log \tau - (1 - \tau) \log(1 - \tau) + \tau \log(d - 1).$$

Lemma C.2. *Assume* $|\sin \angle(\boldsymbol{y}^*, \boldsymbol{\omega}_{j^*})| \leq 1 - \frac{1}{d}$ *. Then,*

$$|R(\boldsymbol{y}^*) - R(\boldsymbol{\omega}_{j^*})| \leq \Delta\left(|\sin \angle(\boldsymbol{y}^*, \boldsymbol{\omega}_{j^*})|\right).$$

Proof. Based on the Fannes-Audenaert inequality:

$$|R(\boldsymbol{y}^*) - R(\boldsymbol{\omega}_{j^*})| = |H(\boldsymbol{\omega}_{j^*}) - H(\boldsymbol{\omega}_{j^*})| \leq \Delta(t),$$

where $t := \frac{1}{2}\|p(\boldsymbol{y}^*) - p(\boldsymbol{\omega}_{j^*})\|_1$ and H is the entropy function. We have

$$\begin{aligned}
\frac{1}{2}\|p(\boldsymbol{y}^*) - p(\boldsymbol{\omega}_{j^*})\|_1 &= \frac{1}{2}\sum_{i=1}^{d}\left|\left(\frac{y_i^*}{\|\boldsymbol{y}^*\|_2}\right)^2 - \left(\frac{\omega_{j^*,i}}{\|\boldsymbol{\omega}_{j^*}\|_2}\right)^2\right| \\
&= \frac{1}{2}\sum_{i=1}^{d}\left|\left(\frac{y_i^*}{\|\boldsymbol{y}^*\|_2} - \frac{\omega_{j^*,i}}{\|\boldsymbol{\omega}_{j^*}\|_2}\right)\left(\frac{y_i^*}{\|\boldsymbol{y}^*\|_2} + \frac{\omega_{j^*,i}}{\|\boldsymbol{\omega}_{j^*}\|_2}\right)\right| \\
&\leq \frac{1}{2}\left[\sum_{i=1}^{d}\left(\frac{y_i^*}{\|\boldsymbol{y}^*\|_2} - \frac{\omega_{j^*,i}}{\|\boldsymbol{\omega}_{j^*}\|_2}\right)^2\right]^{1/2}\left[\sum_{i=1}^{d}\left(\frac{y_i^*}{\|\boldsymbol{y}^*\|_2} + \frac{\omega_{j^*,i}}{\|\boldsymbol{\omega}_{j^*}\|_2}\right)^2\right]^{1/2} \\
&= \frac{1}{2}\left[2 - 2\frac{\boldsymbol{y}^*\cdot\boldsymbol{\omega}_{j^*}}{\|\boldsymbol{y}^*\|_2\|\boldsymbol{\omega}_{j^*}\|_2}\right]^{1/2}\left[2 + 2\frac{\boldsymbol{y}^*\cdot\boldsymbol{\omega}_{j^*}}{\|\boldsymbol{y}^*\|_2\|\boldsymbol{\omega}_{j^*}\|_2}\right]^{1/2} \\
&= \left[1 - \left(\frac{\boldsymbol{y}^*\cdot\boldsymbol{\omega}_{j^*}}{\|\boldsymbol{y}^*\|_2\|\boldsymbol{\omega}_{j^*}\|_2}\right)^2\right]^{1/2} \\
&= |\sin\angle(\boldsymbol{y}^*, \boldsymbol{\omega}_{j^*})|.
\end{aligned}$$

Since $\Delta(t)$ is increasing on $0 \leq t \leq 1 - \frac{1}{d}$,

$$|R(\boldsymbol{y}^*) - R(\boldsymbol{\omega}_{j^*})| \leq \Delta\left(\min\left\{|\sin\angle(\boldsymbol{y}^*, \boldsymbol{\omega}_{j^*})|, 1 - \frac{1}{d}\right\}\right) = \Delta\left(|\sin\angle(\boldsymbol{y}^*, \boldsymbol{\omega}_{j^*})|\right).$$

□

Proposition C.3. *Let*

$$\tau = \kappa\sqrt{\frac{1 - p_{j^*}(\boldsymbol{\alpha})}{p_{j^*}(\boldsymbol{\alpha})}}. \tag{A9}$$

We have

$$|R(\boldsymbol{y}^*) - R(\boldsymbol{\omega}_{j^*})| \leq \Delta(\tau), \tag{A10}$$

where for $\tau \in \left(0, 1 - \frac{1}{d}\right)$, $\Delta(\tau) = -\tau\log\tau - (1-\tau)\log(1-\tau) + \tau\log(d-1)$.

Proof. Recall $\boldsymbol{r} = \sum_{j\neq j^*}\alpha_j\boldsymbol{\omega}_j$, then $\boldsymbol{y}^* = \alpha_{j^*}\boldsymbol{\omega}_{j^*} + \boldsymbol{r}$. Since $\|\boldsymbol{\omega}_{j^*}\|_2 = 1$, decompose the residual vector as

$$\boldsymbol{r} = r_\parallel\boldsymbol{\omega}_{j^*} + \boldsymbol{r}_\perp, \qquad \boldsymbol{r}_\perp \perp \boldsymbol{\omega}_{j^*}.$$

We have

$$\boldsymbol{y}^* = (\alpha_{j^*} + r_\parallel)\boldsymbol{\omega}_{j^*} + \boldsymbol{r}_\perp.$$

The angular deviation from $\boldsymbol{\omega}_{j^*}$ is caused only by the perpendicular residual component. Hence,

$$\begin{aligned}
|\sin\angle(\boldsymbol{y}^*, \boldsymbol{\omega}_{j^*})| &= \frac{\|\boldsymbol{r}_\perp\|_2}{\left[(\alpha_{j^*} + r_\parallel)^2 + \|\boldsymbol{r}_\perp\|_2^2\right]^{1/2}} \\
&\leq \frac{\sqrt{r_\parallel^2 + \|\boldsymbol{r}_\perp\|_2^2}}{|\alpha_{j^*}|} = \frac{\|\boldsymbol{r}\|_2}{|\alpha_{j^*}|}.
\end{aligned}$$

The above inequality follows because, after squaring and rearranging, it is equivalent to

$$0 \leq \left(r_\parallel(\alpha_{j^*} + r_\parallel) + \|\boldsymbol{r}_\perp\|_2^2\right)^2.$$

Moreover,

$$\frac{\|\boldsymbol{r}\|_2}{|\alpha_{j^*}|} = \kappa\frac{\sqrt{\sum_{j\neq j^*}\alpha_j^2}}{|\alpha_{j^*}|} = \kappa\sqrt{\frac{1 - p_{j^*}(\boldsymbol{\alpha})}{p_{j^*}(\boldsymbol{\alpha})}}.$$

Therefore,

$$|\sin\angle(\boldsymbol{y}^*, \boldsymbol{\omega}_{j^*})| \leq \kappa\sqrt{\frac{1-p_{j^*}(\boldsymbol{\alpha})}{p_{j^*}(\boldsymbol{\alpha})}}.$$

Based on Lemma C.2, it is sufficient to prove Proposition 4.6. □

D Additional Results

Identification of the Spike FFN and MAGC. We identify the spike FFN and MAGC by comparing the first (spike) token and the second (non-spike) token of the randomly generated sequences. Let $\boldsymbol{y}_0^l$ denote the average FFN output of the first token in the l-th transformer block. We regard an FFN as the spike FFN if its output exhibits an outlier channel (Sun et al., 2024), i.e.,

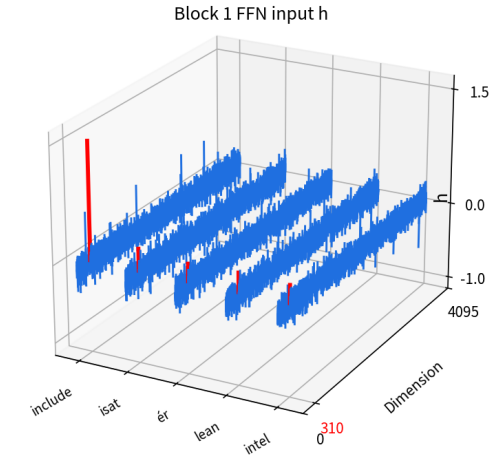
$$\max_j \left|y_0^l(j)\right| > 100 \quad \text{and} \quad \max_j \left|y_0^l(j)\right| > 10 \cdot \text{median}_j \left|y_0^l(j)\right|. \tag{A11}$$

Given the spike FFN, we then identify MAGC from its inputs. Let $\boldsymbol{h}_0$ and $\boldsymbol{h}_1$ be the average inputs to the spike FFN for the first and second tokens, respectively. The gating channel should be the channel where the two inputs deviate the most. We therefore locate MAGC as

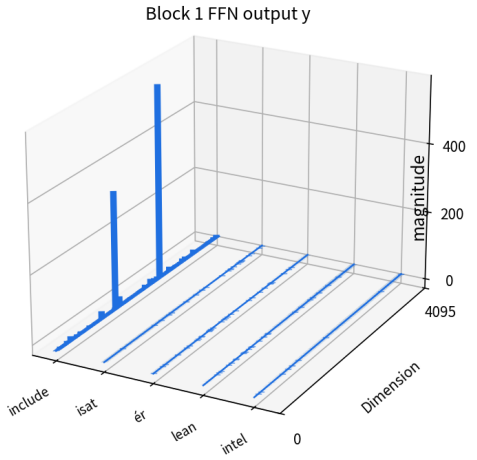
$$i^* = \arg\max_j |h_0(j) - h_1(j)|, \tag{A12}$$

and determine the gating direction by the sign of $h_0(i^*) - h_1(i^*)$: a positive (negative) sign indicates that massive activations emerge when the value of MAGC becomes sufficiently large (small).

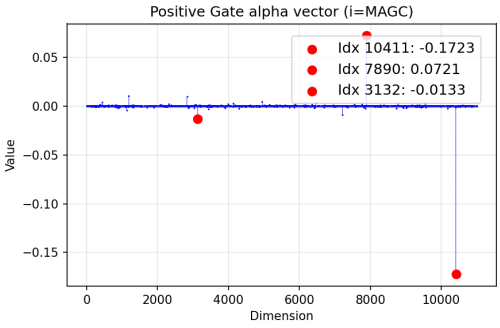
This appendix reports the detailed results for the models summarized in Table 1: LLaMA-2-7B (Figure A5), Qwen2.5-1.5B (Figure A6), QwQ-32B (Figure A7), Ministral-3-3B-Instruct-2512 (Figure A8), and DeepSeek-R1-Distill-Llama-8B (Figure A9). For each model, we present (i) the input $\boldsymbol{h}$ and outputs $\boldsymbol{y}$ of the spike FFN for a random input sequence, (ii) the coefficient vector $\boldsymbol{\alpha}$ defined in Theorem 4.1, where for models with a negative gating direction $\boldsymbol{\alpha}$ is computed under the limit $x \to -\infty$ and the condition on $W_{\text{gate}}^{(j,i)}$ flips sign accordingly, and (iii) the dominant columns (the columns most aligned with the output $\boldsymbol{y}_0$, as reported in Table 1) together with a randomly selected column for comparison.



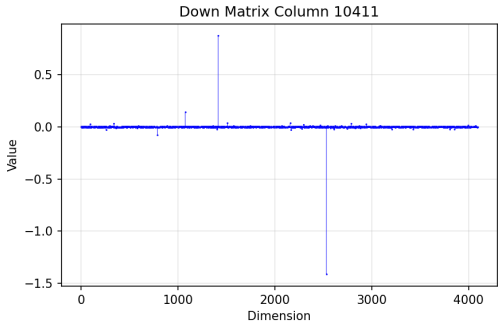
(a) Input $\boldsymbol{h}$ to FFN.



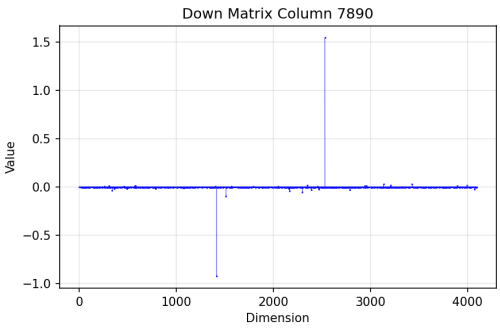
(b) Output $\boldsymbol{y}$ of FFN.



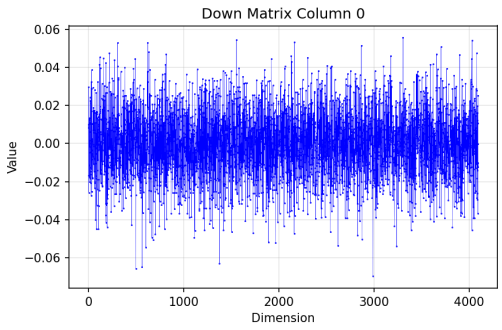
(c) α vector when $i^* = 310$.



(d) $\boldsymbol{\omega}_{10411}$.

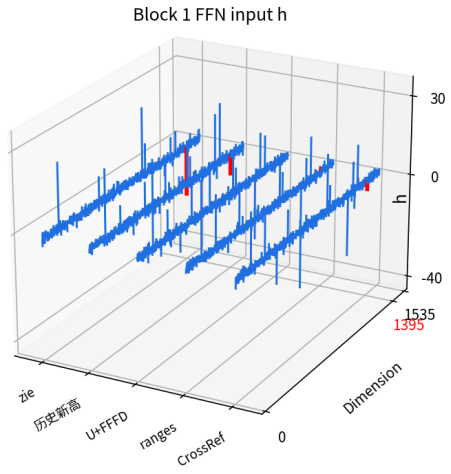


(e) $\boldsymbol{\omega}_{7890}$.

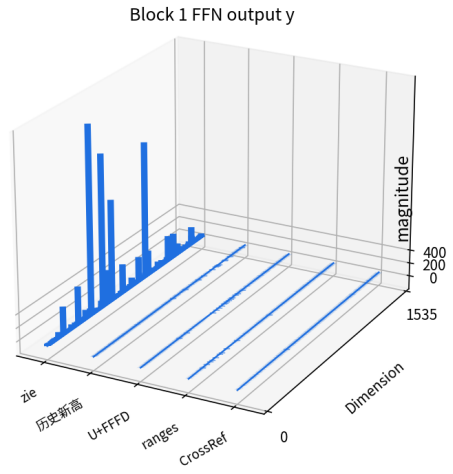


(f) $\boldsymbol{\omega}_{0}$.

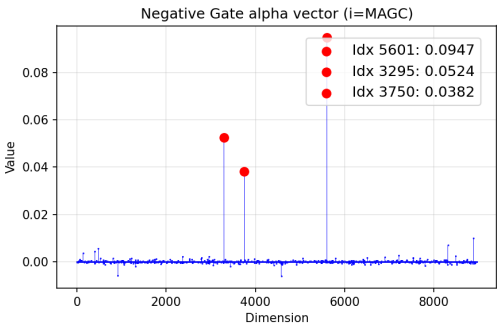
Figure A5: Results for LLaMA-2-7b-hf.



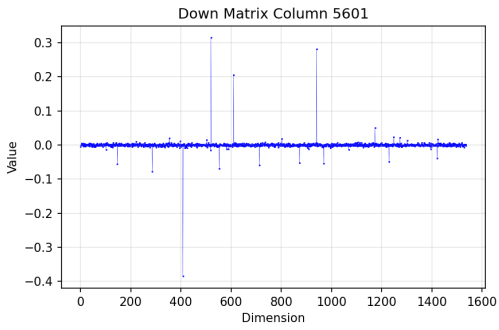
(a) Input $\boldsymbol{h}$ to FFN.



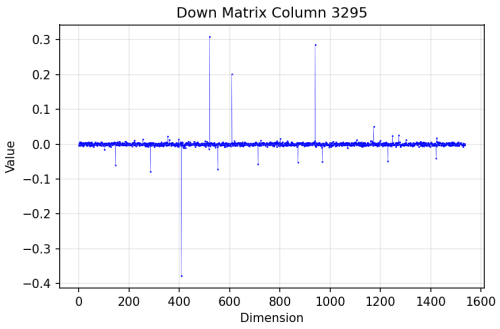
(b) Output $\boldsymbol{y}$ of FFN.



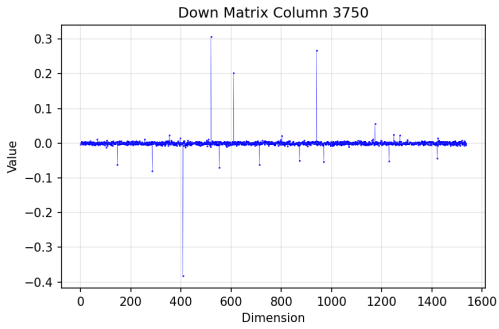
(c) α vector when $i^* = 1395$.



(d) $\boldsymbol{\omega}_{5601}$.

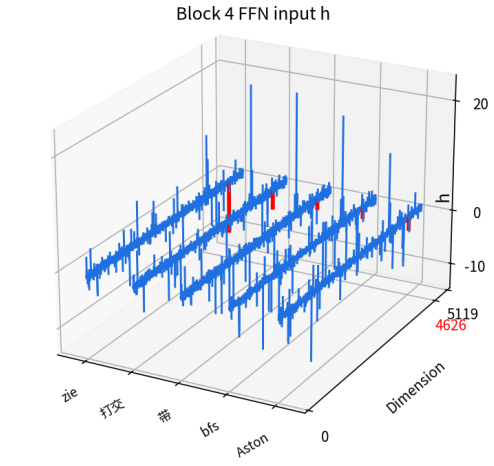


(e) $\boldsymbol{\omega}_{3295}$.

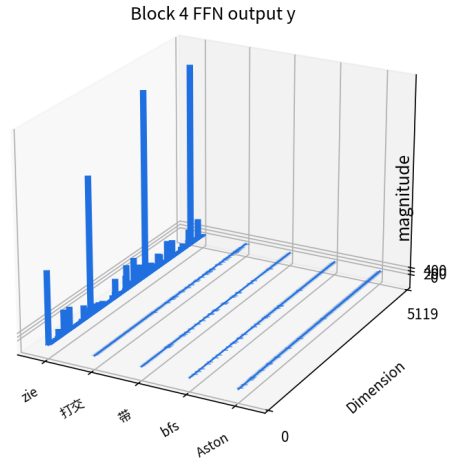


(f) $\boldsymbol{\omega}_{3750}$.

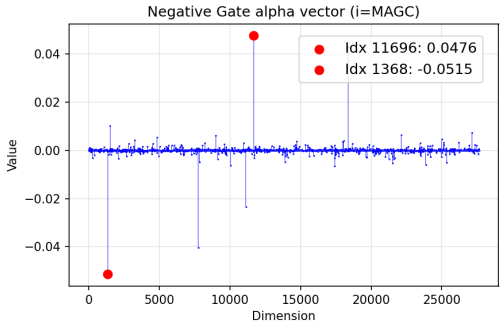
Figure A6: Results for Qwen2.5-1.5b.



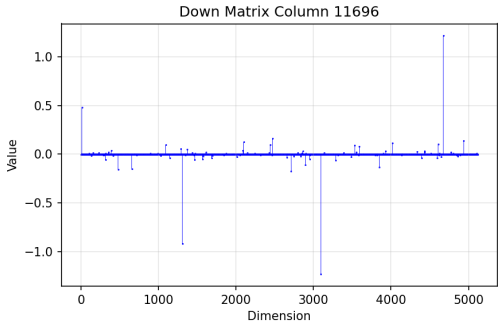
(a) Input $\boldsymbol{h}$ to FFN.



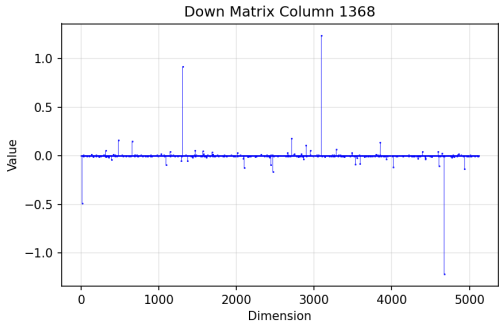
(b) Output $\boldsymbol{y}$ of FFN.



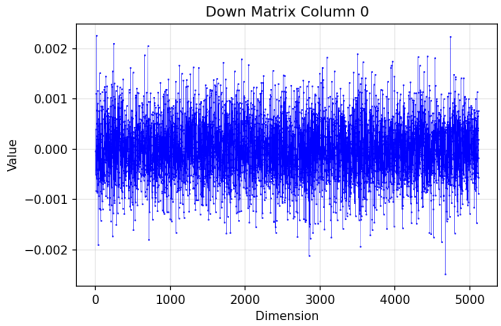
(c) α vector when $i^* = 4626$.



(d) $\boldsymbol{\omega}_{11696}$.

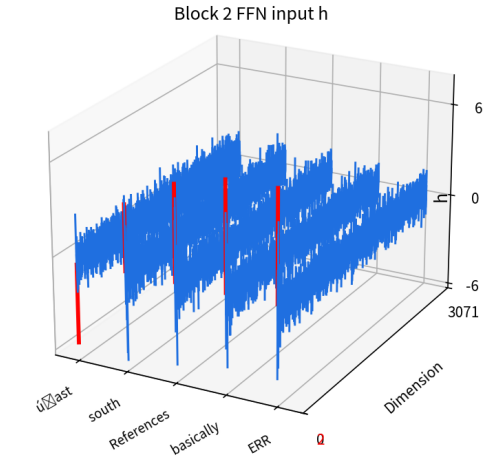


(e) $\boldsymbol{\omega}_{1368}$.

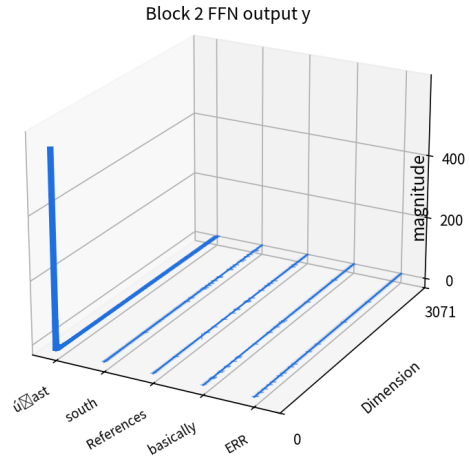


(f) $\boldsymbol{\omega}_0$.

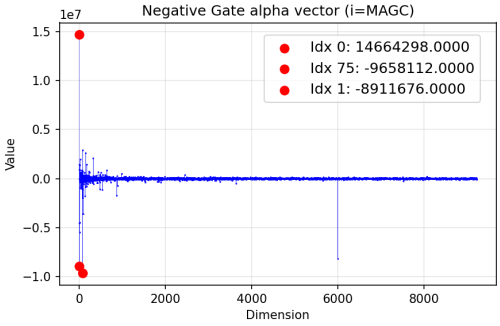
Figure A7: Results for QwQ-32B. For QwQ-32B, several of the channels with the largest input difference $|h_0(j) - h_1(j)|$ coincide with massive-activation channels propagated through the residual stream; we therefore identify the MAGC ($i^* = 4626$, negative gating direction) by traversing the ten channels with the largest such difference following Equation A12 and selecting the first one that is not itself a massive-activation channel, rather than simply taking the single largest.



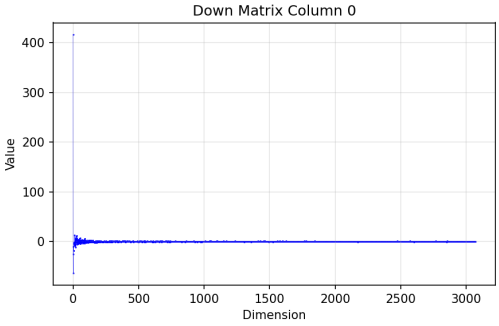
(a) Input $\boldsymbol{h}$ to FFN.



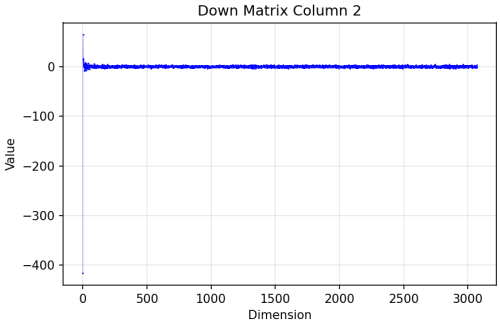
(b) Output $\boldsymbol{y}$ of FFN.



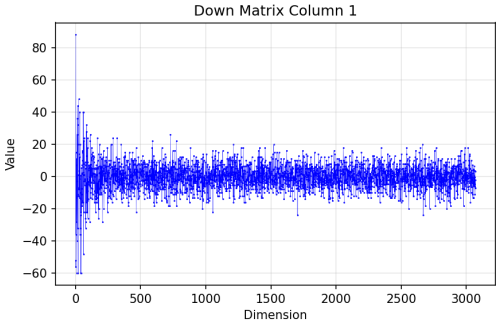
(c) α vector when $i^* = 2$.



(d) $\boldsymbol{\omega}_0$.

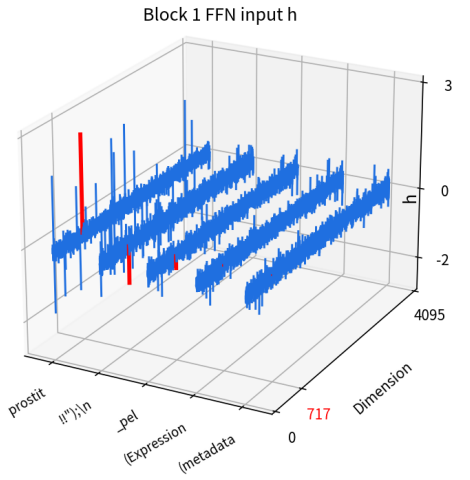


(e) $\boldsymbol{\omega}_2$.

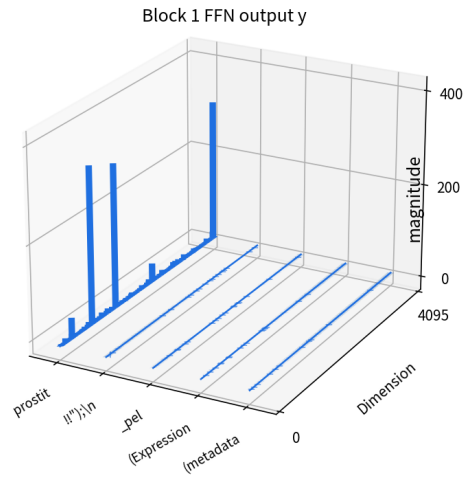


(f) $\boldsymbol{\omega}_1$.

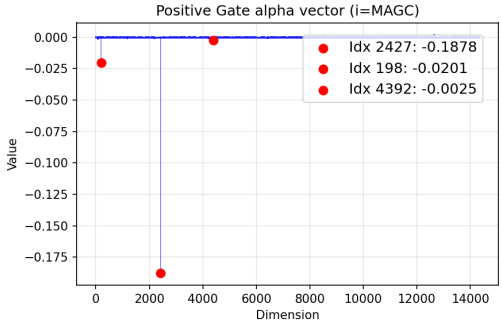
Figure A8: Results for Ministral-3-3B-Instruct-2512. Note that columns 0 and 2 belong to j^*. Instead of choosing 0 as the random column for comparison, we select column 1.



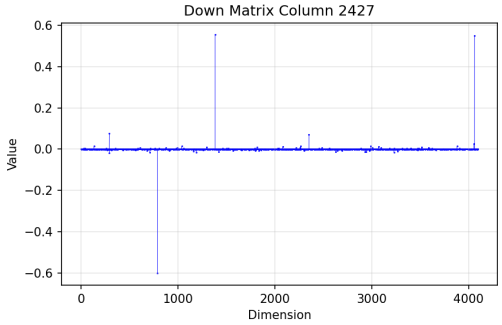
(a) Input $\boldsymbol{h}$ to FFN.



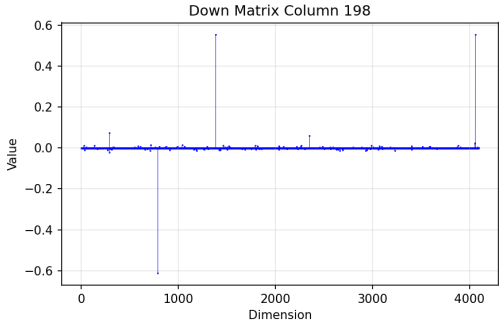
(b) Output $\boldsymbol{y}$ of FFN.



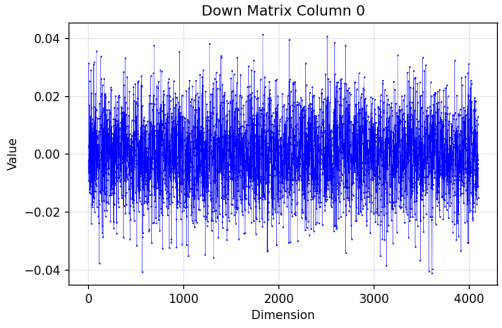
(c) α vector when $i^* = 717$.



(d) $\boldsymbol{\omega}_{2427}$.



(e) $\boldsymbol{\omega}_{198}$.



(f) $\boldsymbol{\omega}_{0}$.

Figure A9: Results for DeepSeek-R1-Distill-Llama-8B.