IntactKV: Improving Large Language Model Quantization by Keeping Pivot Tokens Intact

Ruikang Liu[1] **Haoli Bai**[2] **Haokun Lin**[3] **Yuening Li**[4] **Han Gao**[2]
Zhengzhuo Xu[1] **Lu Hou**[2] **Jun Yao**[2] **Chun Yuan**[1†*]
[1]Shenzhen International Graduate School, Tsinghua University [2]Huawei Noah's Ark Lab
[3]Institute of Automation, Chinese Academy of Sciences [4]The Chinese University of Hong Kong
{liuruikang.cs, xzzthu}@gmail.com {baihaoli, han.g, houlu3, yaojun97}@huawei.com
haokun.lin@cripac.ia.ac.cn yuening@link.cuhk.edu.hk yuanc@sz.tsinghua.edu.cn

Abstract

Large language models (LLMs) excel in natural language processing but demand intensive computation. To mitigate this, various quantization methods have been explored, yet they compromise LLM performance. This paper unveils a previously overlooked type of outliers in LLMs. Such outliers are found to allocate most of the attention scores on initial tokens of input, termed as *pivot tokens*, which are crucial to the performance of quantized LLMs. Given that, we propose *IntactKV* to generate the KV cache of pivot tokens losslessly from the full-precision model. The approach is simple and easy to combine with existing quantization solutions with no extra inference overhead. Besides, IntactKV can be calibrated as additional LLM parameters to boost the quantized LLMs further with minimal training costs. Mathematical analysis also proves that IntactKV effectively reduces the upper bound of quantization error. Empirical results show that IntactKV brings consistent improvement over various quantization methods across different LLMs and downstream tasks, leading to the new state-of-the-art for LLM quantization. The codes are available at https://github.com/ruikangliu/IntactKV.

1 Introduction

Large language models (LLMs) have achieved remarkable progress in various tasks and benchmarks in natural language processing (Brown et al., 2020; Bubeck et al., 2023; Touvron et al., 2023a; Team et al., 2023). Nonetheless, the rise of LLMs also increases computational intensity and memory requirements. This motivates various research to decrease the inference cost of LLMs, e.g., quantizaiton (Frantar et al., 2022; Shao et al., 2024; Lin et al., 2023), pruning (Frantar and Alistarh, 2023; Liu et al., 2023b; Sun et al., 2023; Zhang et al., 2024), and speculative decoding (Chen et al., 2023; Leviathan et al., 2023; Cai et al., 2024), e.t.c.

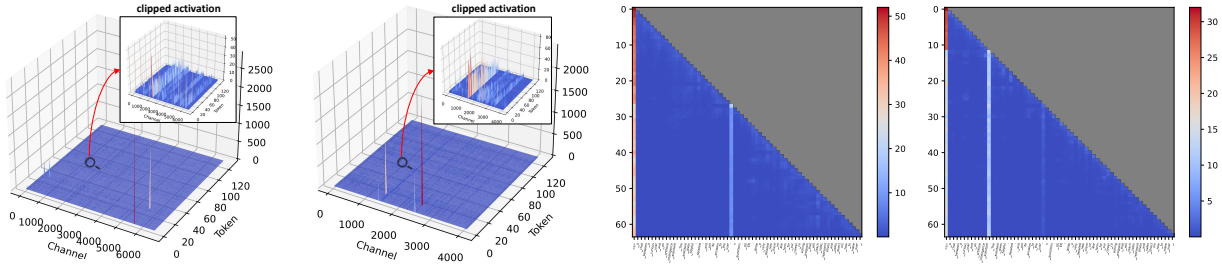
Among these methods, network quantization converts the network parameters or activations from floating-point to fixed-point formats, which is a popular technique to reduce the model size and computational resources. Nevertheless, quantization inevitably affects the performance of LLMs. The leading cause comes from the outliers in LLM activations, which are sensitive to network quantization (Dettmers et al., 2022; Xiao et al., 2023; Lin et al., 2023). As workarounds, there are efforts to either use mixed-precision formats (Dettmers et al., 2022) or re-scale network weights of the outlier channels (Lin et al., 2023). These methods are all built based on the premise that outliers persist in fixed channels across all tokens. However, we find this is not the case for all outliers in LLMs.

In this paper, we discover a new type of outlier that is overlooked by previous quantization methods. These outliers exhibit extremely high values at only the `[BOS]` and some other common tokens (e.g., "," and ".") at the beginning of the input, which is referred to as ***pivot tokens***. We find the extreme values of these outliers make the self-attention concentrate on the pivot tokens, leaving the rest of the tokens untouched. This is also known as attention sinks (Xiao et al., 2024), which is critical to the model performance (Xiao et al., 2024; Bondarenko et al., 2023). The effect of quantization on these pivot tokens should be carefully studied to improve the quantized LLMs.

Towards that end, we are motivated to propose IntactKV, a simple strategy that is orthogonal to most existing quantization solutions. The key idea behind IntactKV is to *generate the lossless KV cache of pivot tokens from the full-precision model*. By keeping the KV cache of pivot tokens intact, quantization error accumulated on the output of self-attention will be effectively alleviated in the rest of the decoding steps. The integration

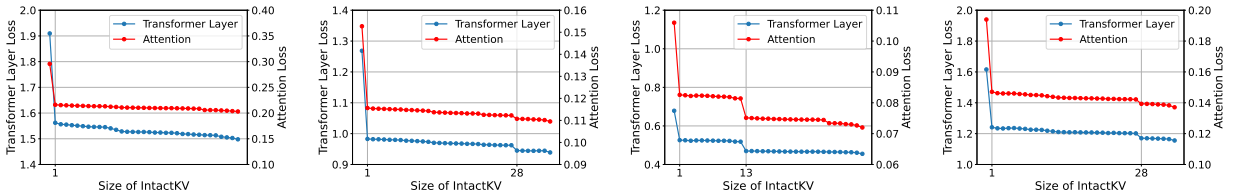
*† Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 7716–7741
August 11-16, 2024



(a) Output activations of LLaMA-30B Layer 24 (b) Output activations of LLaMA-2-7B Layer 24 (c) Attention map of LLaMA-30B Layer 24 (d) Attention map of LLaMA-2-7B Layer 24

Figure 1: Visualizations of Transformer output and attention scores of LLaMA-30B and LLaMA-2-7B. Observations: (1) There are token-specific outliers that can be orders of magnitudes larger than the rest of the tokens (enlarged in the box). Such tokens occur at the [BOS] token, the 28th token "'" in LLaMA-30B and 13th token "." in LLaMA-2-7B, which are referred to as ***pivot tokens***; (2) These outliers over pivot tokens make the attention scores concentrated on themselves, which are likely to be affected by quantization. More details can be found in Appendix C.1.



(a) LLaMA-13B (b) LLaMA-30B (c) LLaMA-2-7B (d) LLaMA-2-70B

Figure 2: The mean squared error (MSE) of the last Transformer layer and attention layers w.r.t. the varying sizes of IntactKV. Observations: (1) The MSE continues to drop as the size of IntactKV increases. (2) Including the pivot tokens' KV cache in IntactKV leads to the most significant decrease in the quantization loss, demonstrating the importance of the pivot tokens' KV cache. More experiment details can be found in Appendix D.

of IntactKV comes with no additional inference overhead. Moreover, IntactKV can also serve as *extra trainable parameters* in addition to the LLM backbone. The calibration process of IntactKV follows the convention of PTQ (Bai et al., 2022; Frantar et al., 2022), which further decreases the quantization error. To get more insights from IntactKV, we also provide mathematical analysis and the results show that IntactKV can effectively lower the upper bound of quantization error.

Empirical results show that IntactKV consistently improves the capability of different quantization methods (e.g. AWQ (Lin et al., 2023), GPTQ (Frantar et al., 2022), OmniQuant (Shao et al., 2024) and QuaRot (Ashkboos et al., 2024)) on various open-sourced LLMs (e.g., LLaMA and Vicuna) across different tasks and benchmarks such as PPL, MMLU, commonsense QA, and MT-bench, achieving new state-of-the-art results for weight-only quantization as well as weight and activation quantization, e.g., lossless INT4 weight-only quantization for Vicuna-v1.3-13B on commonsense QA tasks. Moreover, calibrating IntactKV with INT4 quantization even matches the full-precision model on aligning with human preferences, as evaluated by GPT-4 (Bubeck et al., 2023) on MT-bench.

2 Motivation

2.1 Preliminaries on LLM Quantization

Network quantization is popularly studied in the literature of efficient LLMs (Frantar et al., 2022; Lin et al., 2023; Shao et al., 2024). It allows larger throughput by reducing the model size and leads to practical inference speedup. Given the full-precision weight $\mathbf{w}$, quantization aims to convert it to the low-bit representation $\hat{\mathbf{w}}$. The general b-bit uniform quantization $\mathcal{Q}_b(\cdot)$ can be represented as

$$\hat{\mathbf{w}} = \mathcal{Q}_b(\mathbf{w}) = s \cdot \Pi_{\Omega(b)}(\mathbf{w}/s), \quad (1)$$

where s is the quantization step size, and $\Pi_{\Omega(b)}$ is the projection function onto the set of b-bit integers $\Omega(b) = \{0, 1, ..., 2^b - 1\}$. While we mainly focus on weight-only quantization, Equation 1 can be similarly used to quantize activations and KV cache of LLMs to increase the inference throughput (Xiao et al., 2023; Shao et al., 2024; Hooper et al., 2024).

Following most existing works in LLM quantization, we focus on post-training quantiza-

tion (PTQ) (Frantar et al., 2022; Lin et al., 2023), since it does not introduce extra training overhead as those in quantization-aware training (QAT) (Liu et al., 2023a; Li et al., 2024). Quantization inevitably downgrades LLMs in low-bit settings, where the outliers in quantized LLMs are found to be the cause of the deterioration (Dettmers et al., 2022). In the next, we study the details of how these outliers affect the LLM quantization.

2.2 Revisiting Outliers in LLMs

We discover a new type of outlier that is specific to particular tokens, which leads the attention sink (Xiao et al., 2024) that is critical to the performance of LLMs.

A New Variant of Outlier. Different from the outliers that persist in several fixed channels across different tokens (Dettmers et al., 2022; Xiao et al., 2023; Lin et al., 2023), we find a new variant of outlier that is specific to some initial tokens of the input sequence. By visualizing the activation of Transformer layer output in Figure 1a and Figure 1b, there exist peaks with magnitudes over 1e3. These outliers can be hundreds of times larger than the previous outliers that persist in fixed channels across all tokens, as enlarged in Figure 1a and Figure 1b. More visualizations can be found in Appendix C. It is found that such huge outliers usually occur at the `[BOS]` token and some other uninformative initial tokens (e.g., "." or ",") at particular channels, regardless of the rest of the input sequence. We thus name these tokens ***pivot tokens*** given their dominating values in the activation. Recently, a concurrent work (Sun et al., 2024) also discovers such outliers with more detailed studies.

Pivot Tokens Exhibit Attention Sinks. We hypothesize that the outliers over these pivot tokens may propagate to queries and keys in the self-attention. Consequently, the attention scores will be concentrated on these pivot tokens than the rest ones, a.k.a *attention sinks* (Xiao et al., 2024). To verify the hypothesis, we plot the attention scores in Figure 1c and Figure 1d. It can be found that the pivot tokens indeed dominate the attention scores, especially for the first token (i.e., `[BOS]`). This corresponds to the observations in attention sinks (Xiao et al., 2024), which are empirically verified to be critical to the model performance. The recent study by (Bondarenko et al., 2023) also shows that concentrating on these tokens naturally helps the attention head do nothing but simply a partial update of the residual. In the decoding stage of LLMs, all generated tokens need to interact with pivot tokens through self-attention. However, as mentioned in Section 2.1, network quantization would inevitably distort the output from the full-precision model. The concentrated scores of pivot tokens thus can be further deviated by quantization, which downgrades the model performance.

3 Method

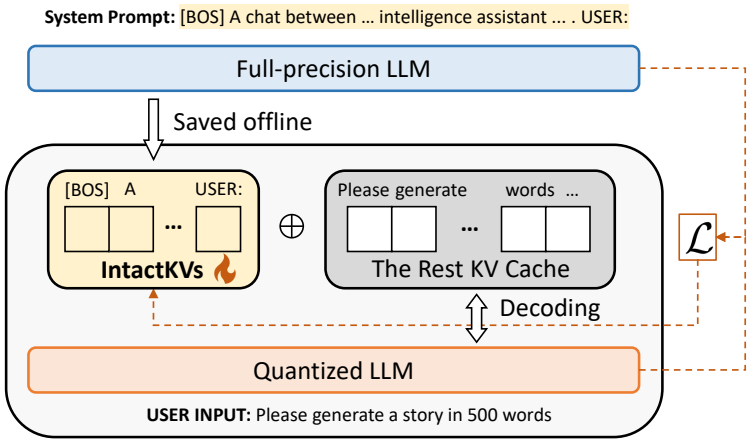
In this section, we introduce IntactKV, a simple and easy-to-implement method to improve the quantized LLMs. The key idea behind this is to keep the KV cache of the pivot tokens intact, i.e., without any distortion raised by quantization. An overview of our method can be found in Figure 3.

3.1 Preserving the KV Cache of Pivot Tokens

According to Section 2.2, the attention sinks of pivot tokens are likely to deteriorate by quantization. To alleviate this issue, we propose IntactKV, a simple yet effective strategy to keep these pivot tokens intact. Specifically, as illustrated in Figure 3a, we leverage the full-precision LLM to generate the lossless KV cache of pivot tokens, which is saved offline. The quantized LLM then loads IntactKV as the prefix to concatenate with the rest of the KV cache and continues with the regular auto-regressive decoding process. The pseudo code of the inference scheme with IntactKV is presented in Figure 3b.

In order to study the benefits of IntactKV, we conduct a preliminary test on the mean squared error (MSE) of the attention and transformer layer output. From Figure 2, it is natural that the increasing size of IntactKV gives the monotonically decreasing MSE on both the attention and transformer layers. More importantly, it is found the pivot tokens observed in Section 2.2 (e.g., `[BOS]` and other delimiter tokens) give the most significant decrease on the MSE, which demonstrates the importance of their KV cache. This aligns with the observations in Figure 1 that pivot tokens exhibit outliers with extreme values and attention sinks.

The Choice of Pivot Tokens and IntactKV. It is the key design to choose the pivot tokens and the associated IntactKV. Given the observations in Figure 2, one can naively pick pivot tokens with the most MSE reduction for IntactKV. However, this is in fact not the case. Since IntactKV acts as the prefix to the KV cache of quantized LLMs, it must



(a) The overview of IntactKV.

Algorithm: Pseudo code for Inference with IntactKV

```
def inference(fp16_model, quantized_model,
              queries, prompt):
    # get IntactKV
    intactkv = fp16_model(
        input=prompt
    ).past_key_values

    # discard the full-precision model
    del fp16_model

    # inference with IntactKV
    outputs = []
    for query in queries:
        output = quantized_model(
            input=query[len(prompt):],
            past_key_values=intactkv
        )
        outputs.append(output)

    return outputs
```

(b) Pseudo Code of Inference.

Figure 3: The overview of the proposed IntactKV applied for the supervised fine-tuned LLM. The full-precision model takes the system prompt as input and generates the IntactKV losslessly as the prefix concatenated with the rest of the KV cache of quantized LLMs. IntactKV can be further calibrated by minimizing the mean squared error $\mathcal{L}$ between the full-precision and quantized LLMs.

start from the very first token, and be consecutive in length. This ensures it to be input agnostic, and the full-precision LLMs can be safely discarded once IntactKV is generated. Next, we provide practical solutions to this problem for different LLMs.

- For pre-trained LLMs, we propose the IntactKV of size one that only contains `[BOS]` KV cache. It is a convention to prepend `[BOS]` to the input of pre-trained LLMs. Moreover, as illustrated in Section 2, `[BOS]` is the pivot token with extreme outlier and attention scores. Besides, the KV cache of `[BOS]` has a great impact on the MSE of the quantized model. Employing a lossless `[BOS]` KV cache is thus believed to decrease the quantization loss.

- For supervised fine-tuned (SFT) models, when the input follows the system prompt, we argue that extending IntactKV to the same length of the system prompt can further improve quantized LLMs. In addition to `[BOS]`, other tokens appearing at the beginning of the input sequence also have the potential to serve as pivot tokens (see Figure 1). The system prompt is usually prepended to the input, which allows it to cover more pivot tokens. As shown in Figure 2, remedying the quantization error of these pivot tokens' KV cache can be helpful to compensate for the quantization error. We find that for Vicuna models, system prompt is enough to cover all the pivot tokens, more details can be found in Appendix C.3.

Overhead of IntactKV. Finally, we highlight that IntactKV does not introduce extra latency overhead during inference. Besides, as IntactKV is pre-computed, the pre-filling stage of the quantized LLMs can be accelerated as well. The memory overhead to save IntactKV is also negligible compared with the LLM backbone. For instance, there are only 34 tokens of the system prompt for Vicuna-v1.5-7B, and thus IntactKV takes only 0.13% of the LLM model parameters.

3.2 IntactKV as Trainable Parameters

Since IntactKV is pre-computed and saved offline, it can be treated as extra trainable parameters aside from the LLM backbone to further boost the quantized LLMs. Despite there being no information loss at the pivot tokens, the quantization may still introduce errors to the KV cache during the decoding stage. As shown in Figure 3a, we calibrate IntactKV to compensate for the quantization error accumulated in the following tokens. While there are various metrics to characterize the quantization discrepancy (Frantar et al., 2022; Shao et al., 2024; Liu et al., 2023a), we adopt the mean squared error of the transformer layer output between the full-precision LLM and quantized LLM, a simple yet most widely used metric, i.e.,

$$\mathcal{L}(\Theta) = \frac{1}{2}\sum_{l=1}^{L} \|f_l(\mathbf{w}, \mathbf{x}) - f_l(\hat{\mathbf{w}}, \mathbf{x}; \Theta)\|_2^2, \quad (2)$$

where Θ denotes the set of IntactKV, f_l is the mapping function for the l-th Transformer layer,

and L is the number of Transformer layers in LLM. $\mathbf{x}$ is the input sequence, while $\mathbf{w}, \hat{\mathbf{w}}$ are full-precision and quantized weights respectively. Note that the full-precision model is only required during the calibration process, and it can be safely discarded afterward. It is empirically found that calibration of system prompt IntactKV in SFT models generally gives more improvement than the calibration of `[BOS]` IntactKV in pre-trained LLMs. This matches the intuition that a larger size of IntactKV increases the potential to compensate for quantization errors.

As we focus on the post-training quantization, the training of IntactKV is highly lightweight since the only learnable parameters introduced are IntactKV, i.e., the KV cache of pivot tokens. It takes only as few as 20 epochs on a calibration set with 128 samples. Besides, training with a quantized model further lowers the memory cost. The calibration process takes about only 10 minutes for a 7B model and less than 20 minutes for a 13B model on one computing device.

3.3 Theoretical Analysis

In this section, we provide a theoretical view of how the proposed IntactKV benefits the quantized LLM. For the clarity of presentation, our analysis is built on the self-attention module of a Transformer layer, while it can be readily extended to the FFN module and multiple layers.

Specifically, we denote $\boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{n\times d}$ as the KV cache during the decoding stage, and $\boldsymbol{q} \in \mathbb{R}^d$ is the query vector, where n and d are the sequence length and head dimension. Recall that the output of each attention head $\boldsymbol{h} \in \mathbb{R}^d$ is computed as

$$\boldsymbol{h} = \text{softmax}(\boldsymbol{q}\boldsymbol{K}^\top/\sqrt{d})\boldsymbol{V}\boldsymbol{W}^O, \qquad (3)$$

where $\boldsymbol{W}^O \in \mathbb{R}^{d\times d}$ is the weight matrix of the projection layer. By quantizing the LLMs, there will be errors accumulated on the KV cache, denoted as $\Delta\boldsymbol{K}, \Delta\boldsymbol{V} \in \mathbb{R}^{n\times d}$. Therefore, we are interested in showing how $\Delta\boldsymbol{K}$ and $\Delta\boldsymbol{V}$ are propagated to the change of attention head $\Delta\boldsymbol{h}$, and to what extent IntactKV alleviates the distortion.

Theorem 1. *Given the query vector $\boldsymbol{q} \in \mathbb{R}^d$ and the change of KV caches $\Delta\boldsymbol{K}, \Delta\boldsymbol{V} \in \mathbb{R}^{n\times d}$, the change of the attention head $\Delta\boldsymbol{h}$ is bounded by*

$$\begin{aligned}\|\Delta\boldsymbol{h}\|_2 \le\, & C_1\|\Delta\boldsymbol{K}\|_{2,\infty}\|\Delta\boldsymbol{V}\|_F + \\ & + C_2\|\Delta\boldsymbol{K}\|_{2,\infty} + C_3\|\Delta\boldsymbol{V}\|_F,\end{aligned}$$

where $C_1 = \frac{n^{3/2}}{\sqrt{d}}C_3\|\boldsymbol{q}\|_2, C_2 = C_1\|\boldsymbol{V}\|_2$ and $C_3 = \|\boldsymbol{W}^O\|_2$.

The proof to Theorem 1 can be found in Appendix 7. We preserve the terms w.r.t. $\Delta\boldsymbol{K}$ and $\Delta\boldsymbol{V}$ of interests, and leave the rest as constants. Note that $\Delta\boldsymbol{K}$ can be further separated by the pivot tokens $\Delta\boldsymbol{K}_p$ and rest tokens $\Delta\boldsymbol{K}_{\setminus p}$, and similar notations hold for $\Delta\boldsymbol{V}$. Therefore, we have $\|\Delta\boldsymbol{K}\|_{2,\infty} = \max\left(\|\Delta\boldsymbol{K}_p\|_{2,\infty}, \|\Delta\boldsymbol{K}_{\setminus p}\|_{2,\infty}\right)$, and $\|\Delta\boldsymbol{V}\|_F = \sqrt{\|\Delta\boldsymbol{V}_p\|_F^2 + \|\Delta\boldsymbol{V}_{\setminus p}\|_F^2}$. With IntactKV we have $\|\Delta\boldsymbol{K}_p\|_{2,\infty} = \|\Delta\boldsymbol{V}_p\|_F = 0$ since they are generated losslessly, which decreases the upper bound of $\|\Delta\boldsymbol{h}\|_2$. Moreover, it can further reduce the bound by incorporating more pivot tokens. This also aligns with the observation in Figure 2 that a larger size of IntactKV gives a lower MSE of the attention module.

4 Experiments

4.1 Settings

Models. We evaluate the proposed IntactKV on various sizes of open-sourced LLMs, including LLaMA (Touvron et al., 2023a) (7B-65B), LLaMA-2 (Touvron et al., 2023b) (7B-70B), Vicuna-v1.3 (Chiang et al., 2023) (7B-33B) and Vicuna-v1.5 (7B-13B). We denote models that keep intact `[BOS]` KV as IntactKV$_{[B]}$, and models that keep intact system prompt KV as IntactKV$_{[P]}$.

Quantization Methods. We mainly consider weight-only quantization methods, including round-to-nearest quantization (RTN), GPTQ (Frantar et al., 2022), the state-of-the-art OmniQuant (Shao et al., 2024) and AWQ (Lin et al., 2023). For GPTQ, we use AutoGPTQ with C4 calibration set following (Frantar et al., 2022) to reproduce all results. For AWQ and OmniQuant, we use the official code or checkpoint with Pile (Gao et al., 2020) and WikiText2 (Merity et al., 2016) calibration set respectively, following (Lin et al., 2023; Shao et al., 2024). More implementation details can be found in Appendix E. We adopt asymmetric group-wise quantization with a group size of 128 and mainly focus on INT3 and INT4 quantization since INT8 is empirically lossless on various task metrics (Dettmers et al., 2022).

Our IntactKV can be readily combined with these existing weight-only quantization methods, and the experiment results are shown in Section 4.2. Moreover, aside from weight-only quantization, the

Method	LLaMA-7B	LLaMA-13B	LLaMA-30B	LLaMA-65B	LLaMA-2-7B	LLaMA-2-13B	LLaMA-2-70B
FP16	7.36	6.82	6.15	5.83	7.28	6.75	5.73
RTN	9.15	7.89	6.85	6.33	8.97	7.60	6.27
+IntactKV$_{[B]}$	8.52	7.66	6.69	6.20	8.61	7.48	6.13
GPTQ	8.59	7.49	6.73	6.29	9.58	7.43	6.33
+IntactKV$_{[B]}$	8.30	7.42	6.62	6.23	9.27	7.36	6.28
OmniQuant	8.26	7.39	6.65	6.18	8.35	7.43	6.12
+IntactKV$_{[B]}$	8.25	7.39	6.64	6.18	8.33	7.40	6.11
AWQ	8.26	7.38	6.59	6.16	8.31	7.32	6.05
+IntactKV$_{[B]}$	**8.12**	**7.36**	**6.54**	**6.12**	**8.18**	**7.29**	**6.04**

Table 1: INT3-group128 weight-only quantization results of LLaMA and LLaMA-2 Models on C4 dataset.

Task Acc	MMLU (5 shot) average					Common Sense QA (0 shot) average				
Vicuna Family	v1.5-7B	v1.5-13B	v1.3-7B	v1.3-13B	v1.3-33B	v1.5-7B	v1.5-13B	v1.3-7B	v1.3-13B	v1.3-33B
FP16	49.84%	55.78%	47.12%	52.10%	59.30%	65.33%	68.38%	64.52%	67.22%	69.53%
RTN	44.62%	51.44%	39.33%	44.56%	53.18%	61.36%	66.12%	59.05%	63.43%	67.33%
+IntactKV$_{[B]}$	45.93%	51.89%	41.74%	46.73%	55.20%	61.94%	65.91%	61.26%	63.94%	**67.95%**
GPTQ	43.99%	52.95%	40.12%	47.83%	55.84%	58.61%	66.34%	59.56%	65.11%	66.66%
+IntactKV$_{[B]}$	44.86%	52.49%	41.55%	48.53%	56.32%	59.12%	66.53%	60.46%	**65.13%**	67.93%
OmniQuant	46.62%	52.82%	42.95%	48.23%	55.21%	62.30%	65.58%	60.89%	64.62%	67.61%
+IntactKV$_{[B]}$	46.27%	52.67%	43.85%	48.31%	55.51%	62.01%	65.67%	60.66%	64.89%	67.61%
AWQ	46.45%	52.92%	43.08%	48.56%	56.09%	62.18%	66.51%	60.75%	64.56%	67.67%
+IntactKV$_{[B]}$	**46.87%**	**53.58%**	**44.67%**	**49.05%**	**56.91%**	**62.49%**	**66.93%**	**61.93%**	65.02%	67.90%

Table 2: INT3-group128 weight-only quantization results of Vicuna models on 5-shot MMLU and 0-shot QA tasks.

proposed IntactKV can be similarly applied for KV cache quantization and extended to activation quantization, as detailed in Section 4.3 and Section 4.4. It is worth noting that the integration of IntactKV with weight-only/KV cache/activation quantization comes with no extra inference cost and works as an effective plugin to effectively boost the accuracy of quantized models.

Evaluation. For pre-trained LLMs (i.e., LLaMA and LLaMA-2), we report the perplexity (PPL) of language generation on C4 (Raffel et al., 2020) and WikiText2 (Merity et al., 2016) dataset. For SFT models (i.e., Vicuna-v1.3 and v1.5), we conduct evaluation over a wide range of downstream tasks. We test the zero and five-shot performance on the Massively Multitask Language Understanding (MMLU) (Hendrycks et al., 2020) benchmark. Meanwhile, we also evaluate seven zero-shot commonsense QA tasks: OBQA (Mihaylov et al., 2018), WinoGrande (Sakaguchi et al., 2021), ARC-Challenge, ARC-Easy (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), and LAMBADA (Paperno et al., 2016). Additionally, we evaluate quantized Vicuna on MT-bench (Zheng et al., 2023), a high-quality dataset consisting of 80 open-ended multi-turn questions, to gauge their alignment with human preferences. The responses generated by quantized models are judged by GPT-4 with a total score of 10. More evaluation details can be found in Appendix F.

Implementation Details For evaluation on PPL, MMLU, and commonsense QA tasks, we adopt IntactKV$_{[B]}$ that only includes `[BOS]` KV since the input sequence of these tasks does not use any system prompt. For evaluation of SFT models on MT-bench, we adopt IntactKV$_{[P]}$ to keep an intact system prompt KV cache. The system prompt of Vicuna can be found in Appendix B. For training the cached IntactKV, we randomly sample 128 samples from ShareGPT[1] dataset as our calibration dataset, consisting of multi-turn ChatGPT (OpenAI, 2022) conversations. The layer-wise MSE loss defined in Equation 2 is calculated on the response of ChatGPT. We use AdamW optimizer with learning rate 2×10^{-4}, training for 160 optimizer update steps with a gradient accumulation step of 16, i.e., 20 epochs. As mentioned in Section 3.2, training IntactKV$_{[B]}$ leads to comparable performance compared with vanilla IntactKV. Instead, the calibration of IntactKV$_{[P]}$ has more potential to improve quantized LLMs with longer system prompt. Thus, we primarily evaluate the IntactKV$_{[P]}$ with KV cache of system prompt as trainable parameters in the following ex-

[1] https://huggingface.co/datasets/Aeala/ShareGPT_Vicuna_unfiltered

Method	MMLU (0 shot)					MMLU (5 shot)				
	Hums	STEM	Social	Others	Avg	Hums	STEM	Social	Others	Avg
FP16	47.89%	39.96%	58.86%	57.34%	50.77%	49.78%	40.46%	60.61%	58.24%	52.10%
RTN	42.06%	32.87%	47.61%	49.51%	43.02%	42.42%	34.46%	50.34%	51.57%	44.56%
+IntactKV[B]	42.49%	35.35%	50.37%	52.44%	44.98%	44.65%	36.98%	53.04%	52.84%	46.73%
GPTQ	45.06%	35.88%	52.23%	51.26%	46.09%	45.82%	37.57%	54.83%	53.64%	47.83%
+IntactKV[B]	44.72%	35.42%	52.94%	52.07%	46.22%	45.61%	38.34%	55.83%	55.31%	48.53%
OmniQuant	43.51%	36.85%	52.16%	53.05%	46.18%	45.91%	37.44%	55.31%	54.94%	48.23%
+IntactKV[B]	44.19%	36.61%	53.33%	53.52%	46.72%	46.27%	37.54%	54.99%	54.94%	48.31%
AWQ	45.14%	36.18%	52.55%	53.79%	46.84%	46.65%	37.64%	55.54%	54.87%	48.56%
+IntactKV[B]	45.91%	36.65%	53.75%	54.60%	**47.64%**	46.57%	38.40%	56.03%	55.95%	**49.05%**

Table 3: INT3-group128 weight-only quantization results of Vicuna-v1.3-13B on MMLU benchmarks.

#bits	Method	OBQA	WinoGrande	ARC-C	ARC-E	BoolQ	HellaSwag	LAMBADA	Avg
FP16	-	45.40%	71.03%	47.70%	73.70%	82.81%	77.00%	72.91%	67.22%
w3g128	RTN	44.00%	70.96%	44.03%	67.30%	80.40%	73.33%	64.00%	63.43%
	+IntactKV[B]	44.80%	69.93%	45.05%	68.35%	79.42%	74.81%	65.22%	63.94%
	GPTQ	45.20%	69.77%	46.08%	70.33%	81.90%	74.89%	67.59%	65.11%
	+IntactKV[B]	44.00%	70.80%	44.97%	70.75%	81.35%	75.03%	69.03%	**65.13%**
	OmniQuant	45.20%	69.22%	45.22%	68.90%	80.95%	74.72%	68.15%	64.62%
	+IntactKV[B]	45.40%	70.32%	45.31%	68.86%	81.28%	74.52%	68.52%	64.89%
	AWQ	42.80%	68.98%	46.08%	68.98%	81.31%	74.97%	68.78%	64.56%
	+IntactKV[B]	43.20%	69.46%	46.16%	69.74%	81.80%	75.11%	69.67%	65.02%
w4g128	RTN	45.20%	71.43%	48.04%	73.15%	82.87%	76.56%	70.62%	66.84%
	+IntactKV[B]	44.80%	71.51%	47.44%	73.36%	82.75%	77.01%	70.99%	66.84%
	GPTQ	44.60%	70.01%	47.87%	73.32%	82.23%	76.55%	71.78%	66.62%
	+IntactKV[B]	45.00%	71.35%	46.76%	73.02%	83.33%	77.00%	71.55%	66.86%
	OmniQuant	45.60%	70.56%	46.76%	73.02%	82.81%	76.74%	70.41%	66.56%
	+IntactKV[B]	45.20%	71.43%	46.25%	72.52%	82.63%	76.90%	70.31%	66.46%
	AWQ	45.20%	70.32%	47.27%	73.91%	82.81%	76.79%	71.32%	66.80%
	+IntactKV[B]	45.60%	71.19%	47.10%	73.32%	82.72%	76.95%	71.38%	**66.89%**

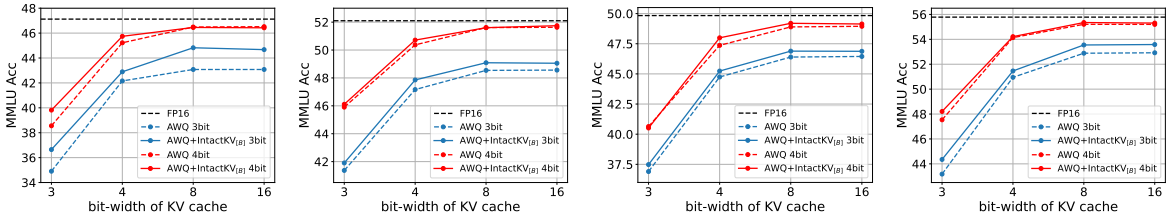
Table 4: Weight-only quantization results of Vicuna-v1.3-13B on seven 0-shot commonsense QA tasks.

periments. For weight and activation quantization, we further quantize IntactKV to lower bits to avoid extra inference overhead, which only incurs negligible accuracy loss. More details of activation quantization can be found in Section 4.4.

4.2 Main Results

Results on Language Generation Tasks. We first integrate our proposed IntactKV with RTN, GPTQ, OmniQuant, and AWQ on LLaMA and LLaMA-2 models. The effect of this integration on model accuracy is measured by the perplexity (PPL) metric, with results on the C4 dataset detailed in Table 1, and results on the WikiText2 dataset in Table 7. As indicated in these tables, IntactKV notably enhances the generative capabilities of quantized models across various LLMs and quantization methods, with AWQ+IntactKV consistently achieving new state-of-the-art (SOTA) results. These findings demonstrate the efficacy of IntactKV in improving quantized LLMs and particularly highlight the effectiveness of utilizing the lossless KV cache from full-precision models. We provide more experiment results on LLaMA-3 and other heterogeneous LLMs (e.g. OPT) in Appendix G.1. IntactKV significantly improves different quantized LLMs, especially for LLaMA-3 models with larger quantization error. These results further prove the compatibility of our IntactKV with various LLM backbones.

Results on MMLU Tasks. For SFT models, we implement IntactKV on the quantized Vicuna models and evaluate the multi-task problem-solving ability on the MMLU benchmark. Table 3 presents the detailed zero-shot and five-shot results for Vicuna-v1.3-13B. The results demonstrate that IntactKV significantly enhances the performance of quantized models across all categories of tasks and various quantization methods for Vicuna-v1.3-13B. Moreover, performance of Vicuna family under the five-shot setting is out-



(a) Vicuna-v1.3-7B (b) Vicuna-v1.3-13B (c) Vicuna-v1.5-7B (d) Vicuna-v1.5-13B

Figure 4: Results of weight and KV cache quantization with different bit-widths on 5-shot MMLU benchmark. Note that this is additional to INT3/4 weight-only quantization. Blue and red lines indicate quantizing model weights to INT3 and INT4, respectively. We apply asymmetric per-head dynamic quantization to the KV cache.

Method	Vicuna-v1.5-7B	Vicuna-v1.5-13B
FP16	5.31	5.52
RTN	4.34	5.13
+IntactKV[P]	4.72	5.27
+IntactKV[P]+Cal	**4.73**	**5.30**
OmniQuant	4.78	5.05
+IntactKV[P]	**4.94**	5.10
+IntactKV[P]+Cal	4.85	**5.24**
AWQ	4.74	5.17
+IntactKV[P]	4.68	5.34
+IntactKV[P]+Cal	**4.84**	**5.44**

Table 5: GPT-4 evaluation of INT3-group128 weight-only quantized Vicuna-v1.5 models on MT-Bench. The scores are on a scale of 10.

lined in Table 2. Remarkably, IntactKV achieves an average improvement of 1.05% over OmniQuant and 0.8% over AWQ across five model sizes, with AWQ+IntactKV exhibiting superior performance over all the other quantized models. More results on MMLU are provided in Appendix G.2.

Results on Commonsense QA Tasks. We further evaluate the quantized Vicuna models on zero-shot commonsense QA tasks. The results of Vicuna-v1.3-13B, as detailed in Table 4, indicate that IntactKV enables significant improvements over various quantization methods. For example, AWQ+IntactKV surpasses the average accuracy of AWQ by 0.46% under INT3-g128 quantization. Additionally, Table 2 presents the average accuracy for various sizes of Vicuna models. In these evaluations, our IntactKV leads to an average accuracy improvement of 0.45% across different LLMs and quantization methods, which strongly demonstrates the efficacy of our proposed IntactKV. More results on commonsense QA tasks can be found in Appendix G.3.

Results on MT-Bench. To evaluate the quantized models' generation capabilities in multi-turn conversations and their alignment with human preferences, we use GPT-4 to score the responses of quantized models on MT-Bench. We also calibrate IntactKV, denoted as IntactKV+Cal. From Table 5, IntactKV significantly boosts the quantized model and IntactKV+Cal further enhances generation quality by compensating for the quantization error. For example, the 3-bit Vicuna-v1.5-13B quantized by AWQ has been improved from 5.17 to 5.34 by using the IntactKV, which can be further boosted to 5.44 with trainable IntactKV. We provide INT4 quantization results in Table 13. Remarkably, with trainable IntactKV, AWQ+IntactKV even matches the full-precision model under INT4 quantization, while all other methods clearly lag behind the full-precision model. These results demonstrate the effectiveness of IntactKV as well as treating IntactKV as trainable parameters. Notably, the training process for the 7B model takes only 10 minutes on a single computing device, which is quite lightweight. In Appendix H, we further demonstrate the effectiveness of calibrating IntactKV by comparing it with group bias tuning, a commonly used fine-tuning strategy for quantized models. IntactKV calibration can achieve better or comparable results with group bias tuning while using significantly fewer trainable parameters. Besides, IntactKV calibration serves as a more versatile calibration strategy for quantized models, which is suitable for various quantization settings.

4.3 Extension to KV Cache Quantization

IntactKV can be readily applied to KV cache quantization to further decrease memory requirements. We employ a mixed-precision strategy that keeps IntactKV in FP16 while the rest of the KV cache is quantized to lower bits. This only induces negligible memory overhead since IntactKV only contains the KV cache of the first few tokens. Note that this does not bring any additional

Method	LLaMA-7B		LLaMA-13B		LLaMA-2-7B		LLaMA-2-13B		LLaMA-3-8B	
	C4	WikiText2	C4	WikiText2	C4	WikiText2	C4	WikiText2	C4	WikiText2
FP16	7.36	5.69	6.82	5.08	7.28	5.48	6.75	4.89	9.48	6.15
OmniQuant	17.03	12.17	15.65	11.16	21.40	14.74	16.24	12.28	-	-
+IntactKV[B]	16.24	11.32	13.87	10.04	20.01	13.70	15.91	11.00	-	-
QuaRot	8.23	6.29	7.40	5.55	8.30	6.11	7.51	5.39	13.42	8.21
+IntactKV[B]	**8.05**	**6.15**	**7.32**	**5.45**	**8.12**	**5.97**	**7.25**	**5.21**	**12.23**	**7.54**

Table 6: INT4 weight and activation quantization results of LLaMA models on C4 and WikiText2 datasets.

inference costs since in the workflow of KV cache quantization, all quantized KV cache needs to be de-quantized back to FP16 before the matrix multiplication. Keeping IntactKV in FP16 reduces the overhead of de-quantization, i,e., we only need to cheaply concatenate the FP16 IntactKV with the rest de-quantized KV cache together. From Figure 4, IntactKV notably improves AWQ across different models and KV cache bit widths under the INT3 weight quantization. For INT4 weight quantization, AWQ+IntactKV still gains an average accuracy increase of 0.27% over the original quantized model. We also notice that quantizing the KV cache to INT8 leads to almost no performance drop on the MMLU benchmark. When equipped with IntactKV, INT8 KV cache can even surpass vanilla AWQ-quantized models with FP16 KV cache, especially under INT3 weight quantization.

4.4 Extension to Activation Quantization

In Table 6, we provide experiment results of combining IntactKV with OmniQuant (Shao et al., 2024) and QuaRot (Ashkboos et al., 2024) for weight and activation quantization. The implementation details can be found in Appendix E. To avoid extra inference costs, we need to quantize the whole KV cache to lower bits and can not keep the KV cache of pivot tokens intact. However, as detailed in Appendix I, we find that IntactKV has a significantly smoother distribution compared with the rest of the KV cache. Therefore, the full-precision IntactKV can be readily quantized to lower bits with negligible accuracy loss, thus rendering IntactKV amenable to weight and activation quantization with no extra inference costs. As shown in Table 6, our IntactKV significantly surpasses the performance of original quantized models for two different quantization methods, improving the PPL by 1.07 for OmniQuant and 0.31 for QuaRot on average. When combined with QuaRot, our IntactKV archives new state-of-the-art (SOTA) results on INT4 weight and activation quantization.

5 Conclusions

In this paper, we propose IntactKV, a simple and easy-to-combine method to improve large language model quantization. The research is motivated by the previously overlooked outliers over pivot tokens, which lead to attention sinks that are critical to the performance of quantized LLMs. By generating IntactKV with the full-precision model, the quantization error accumulated over the attention scores can be effectively alleviated. IntactKV can also be calibrated as additional parameters to the LLM backbone, further improving the quantized LLMs. Experiments show that combining the proposed IntactKV gives consistent improvement on various sizes of LLMs and across multiple downstream tasks, leading to new state-of-the-art results for large language model quantization.

6 Limitations

More experiments may be needed for LLM evaluation. LLMs are being applied to a wide range of tasks, posing high demands on various model abilities. When quantizing LLMs to low bits, these abilities may be affected to varying degrees. Therefore, a comprehensive evaluation is required to gauge the capabilities of quantized LLMs. Although we experiment on several downstream tasks, such as PPL, MMLU, commonsense QA, and MT-bench, we note that this may not be enough to assess all abilities of LLMs. For example, how long context affects quantized models still remains unknown.

7 Ethics Statement

The development of LLM quantization techniques can further democratize LLMs, lowering the costs of LLM serving and enabling more people to get access to advanced AI assistants. Nonetheless, LLM itself may inherit certain social biases from training data concerning gender, race, etc. Quantization can not mitigate such biases. Therefore, caution must be taken when using quantized LLMs.

Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012), Shenzhen Key Laboratory (ZDSYS20210623092001004), and Beijing Key Lab of Networked Multimedia.

References

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. 2024. Quarot: Outlier-free 4-bit inference in rotated llms. *arXiv preprint arXiv:2404.00456*.

Haoli Bai, Lu Hou, Lifeng Shang, Xin Jiang, Irwin King, and Michael R Lyu. 2022. Towards efficient post-training quantization of pre-trained language models. *Advances in Neural Information Processing Systems*, 35:1405–1418.

Pierre Blanchard, Desmond J Higham, and Nicholas J Higham. 2021. Accurately computing the log-sum-exp and softmax functions. *IMA Journal of Numerical Analysis*, 41(4):2311–2330.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2023. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2924–2936.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*.

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pages 10323–10337. PMLR.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Optq: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. 2024. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *arXiv preprint arXiv:2401.18079*.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR.

Yixiao Li, Yifan Yu, Chen Liang, Nikos Karampatziakis, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2024. Loftq: LoRA-fine-tuning-aware quantization for large language models. In *The Twelfth International Conference on Learning Representations*.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. 2023a. Llm-qat: Data-free quantization aware training for large language models. *arXiv preprint arXiv:2305.17888*.

Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, et al. 2023b. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pages 22137–22176. PMLR.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.

OpenAI. 2022. Introducing chatgpt.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1525–1534.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. 2024. Omniquant: Omnidirectionally calibrated quantization for large language models. In *The International Conference on Learning Representations*.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Yingtao Zhang, Haoli Bai, Haokun Lin, Jialin Zhao, Lu Hou, and Carlo Vittorio Cannistraci. 2024. Plug-and-play: An efficient post-training pruning method for large language models. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Proof of Theorem 1

Proof. Denote the output of the softmax function as the score $\boldsymbol{s}$, i.e., $\boldsymbol{s} = \mathrm{softmax}(\frac{\boldsymbol{q}\boldsymbol{K}^\top}{\sqrt{d}})$, and also define the error output from the softmax function as $\Delta\boldsymbol{s}$. To show the error of the attention head, we first justify how the error propagates from the score to the attention head.

$$\begin{aligned}\|\Delta\boldsymbol{h}\|_2 &= \left\|[(\boldsymbol{s}+\Delta\boldsymbol{s})(\boldsymbol{V}+\Delta\boldsymbol{V}) - \boldsymbol{s}\boldsymbol{V}]\boldsymbol{W}^O\right\|_2 \\ &\le (\|\Delta\boldsymbol{s}\|_2\|\boldsymbol{V}+\Delta\boldsymbol{V}\|_2 + \|\boldsymbol{s}\|_2\|\Delta\boldsymbol{V}\|_2)\,\|\boldsymbol{W}^O\|_2 \\ &\le \big(\|\Delta\boldsymbol{s}\|_2(\|\boldsymbol{V}\|_2+\|\Delta\boldsymbol{V}\|_F) + \|\Delta\boldsymbol{V}\|_F\big)\|\boldsymbol{W}^O\|_2,\end{aligned}$$

where the inequalities are because

$$\|\boldsymbol{x}+\boldsymbol{y}\|_2 \le \|\boldsymbol{x}\|_2+\|\boldsymbol{y}\|_2, \quad \|\boldsymbol{s}\boldsymbol{V}\|_2 \le \|\boldsymbol{s}\|_2\|\boldsymbol{V}\|_2,$$

and $\|\boldsymbol{s}\|_2 \le \|\boldsymbol{s}\|_1 = 1, \|\boldsymbol{V}\|_2 \le \|\boldsymbol{V}\|_F$.

Next, we characterize the error of score $\|\Delta\boldsymbol{s}\|_2$. This is not easy as the error propagates through the softmax function. To proceed, we need the relative condition number of the softmax function. As indicated in (Blanchard et al., 2021),

$$\frac{\|\mathrm{softmax}(\boldsymbol{x}+\Delta\boldsymbol{x})-\mathrm{softmax}(\boldsymbol{x})\|_\infty}{\|\mathrm{softmax}(\boldsymbol{x})\|_\infty} \le \kappa(\boldsymbol{x})\frac{\|\Delta\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty},$$

where $\kappa(\boldsymbol{x}) = n\|\boldsymbol{x}\|_\infty$ ($\boldsymbol{x} \in \mathbb{R}^n$) is an upper bound of the relative condition number of the softmax function. Let $\boldsymbol{x} = \boldsymbol{q}\boldsymbol{K}^\top/\sqrt{d}$ and $\Delta\boldsymbol{x} = \boldsymbol{q}\Delta\boldsymbol{K}^\top/\sqrt{d}$, we have

$$\frac{\|\Delta\boldsymbol{s}\|_\infty}{\|\boldsymbol{s}\|_\infty} \le n\|\Delta\boldsymbol{x}\|_\infty \le \frac{n}{\sqrt{d}}\|\boldsymbol{q}\|_2\|\Delta\boldsymbol{K}\|_{2,\infty}.$$

Considering that the output of the softmax function is a probability, we have $\|\boldsymbol{s}\|_\infty \le 1$. Therefore, we obtain

$$\|\Delta\boldsymbol{s}\|_2 \le \sqrt{n}\|\Delta\boldsymbol{s}\|_\infty \le \frac{n^{2/3}}{\sqrt{d}}\|\boldsymbol{q}\|_2\|\Delta\boldsymbol{K}\|_{2,\infty}.$$

Combining the above ingredients, we derive the main results of the Theorem 1.

□

B System Prompt of Vicuna Models

[BOS] A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER:

Figure 5: System Prompt of Vicuna Models.

C Visualization of Activations and Attention Map

C.1 Implementation Details

We use ShareGPT dataset for our visualizations, where each sample starts with Vicuna system prompt of length 34. We use a randomly sampled sequence of length 128 to visualize the output activations and plot the corresponding attention map of the first 64 tokens for better visualization. The attention score is mean pooled over different heads.

C.2 Visualization of LLaMA Models

We provide more visualizations of the output activations and attention map of LLaMA models in Figure. 6–14. Similar to our observations in Section 2, we find that pivot tokens only appear at the very beginning of the input sequence, and `[BOS]` always serves as a pivot token.

C.3 Visualization of Vicuna Models

We provide more visualizations of the output activations and attention map of Vicuna models in Figure. 15–19. Although Vicuna models demonstrate stronger performance than LLaMA models of the same size, we are surprised to find that the position of pivot tokens remains unchanged for Vicuna and LLaMA models of the same size. Besides, as shown in Figure. 15–19, we find that the Vicuna system prompt is enough to cover all the pivot tokens in all Vicuna models.

C.4 Visualization of OPT and Mistral Models

To demonstrate the prevalence of pivot tokens in LLMs, we provide more visualizations on OPT and Mistral models in Figure. 20–21. The results show that the pivot tokens with extreme outliers are ubiquitous in various LLMs.

D Experiment Details of Figure 2

We plot the quantization loss of the last Transformer layer as well as the total quantization loss of all attention layers with respect to the size of IntactKV on four different models, i.e., LLaMA-13B, LLaMA-30B, LLaMA-2-7B, and LLaMA-2-70B, covering different model types and model sizes. We use lossless IntactKV generated by the full-precision model to quantify the effect of IntactKV on the quantized model. IntactKV of size s can ensure that the KV cache of the first s tokens of the input sequence are generated by the

Method	LLaMA-7B	LLaMA-13B	LLaMA-30B	LLaMA-65B	LLaMA-2-7B	LLaMA-2-13B	LLaMA-2-70B
FP16	5.69	5.08	4.09	3.52	5.48	4.89	3.33
RTN	6.98	5.88	4.84	4.22	6.65	5.52	3.99
+IntactKV$_{[B]}$	6.52	5.70	4.69	4.05	6.40	5.44	3.84
GPTQ	6.62	5.68	4.75	4.20	7.29	5.52	4.02
+IntactKV$_{[B]}$	6.51	5.62	4.63	4.12	7.00	5.46	3.97
OmniQuant	6.20	**5.46**	4.59	3.95	**6.10**	5.32	3.81
+IntactKV$_{[B]}$	**6.18**	**5.46**	4.58	3.95	**6.10**	5.31	3.80
AWQ	6.34	5.53	4.60	3.95	6.25	5.32	3.75
+IntactKV$_{[B]}$	6.23	5.49	**4.54**	**3.89**	6.14	**5.29**	**3.72**

Table 7: INT3-group128 weight-only quantization results of LLaMA and LLaMA-2 models on WikiText2 dataset.

Method	LLaMA-3-8B		LLaMA-3-70B	
	C4	WikiText2	C4	WikiText2
FP16	9.48	6.15	7.20	2.87
RTN	18.96	12.05	18.65	8.01
+IntactKV$_{[B]}$	16.89	10.77	14.11	5.43
GPTQ	51.69	26.14	5.1E4	5.1E4
+IntactKV$_{[B]}$	13.08	8.32	3.5E4	4.5E4
OmniQuant	14.46	9.09	9.04	5.29
+IntactKV$_{[B]}$	13.99	8.88	8.83	5.02
AWQ	12.69	8.15	8.55	4.66
+IntactKV$_{[B]}$	**12.42**	**7.97**	**8.35**	**4.41**

Table 8: INT3-group128 weight-only quantization results of LLaMA-3 on C4 and WikiText2 datasets.

Method	OPT-6.7B		Mistral-7B	
	C4	WikiText2	C4	WikiText2
FP16	12.75	10.83	8.39	5.30
RTN	36.18	23.91	9.65	6.20
AWQ	13.39	11.38	9.29	5.95
+IntactKV$_{[B]}$	**13.37**	**11.32**	**9.25**	**5.93**

Table 9: INT3-group128 weight-only quantization results of OPT and Mistral on C4 and WikiText2 datasets.

full-precision model and thus lossless. Quantization loss is computed with MSE loss between the output activations of the quantized model and the full-precision model. We sample 128 sequences from the ShareGPT dataset to construct the validation set, each with a common prompt prefix of length 34. MSE loss is calculated on the tokens after the common prompt prefix. We quantize the model weights to 3 bits using round-to-nearest quantization with a group size of 128.

E Quantization Method Details

We carefully reproduce the results of various quantization methods with their official code or released checkpoint.

Weight-only Quantization. For GPTQ, we use AutoGPTQ[2] with C4 calibration set following (Frantar et al., 2022) to reproduce all results. We turn the quantization option "–desc_act" on to quantize weight columns in order of decreasing activation size, which is a heuristic rule empirically found to be effective for GPTQ. For AWQ (Lin et al., 2023), we directly load the officially released quantization parameters of LLaMA models for evaluation and reproduce results on Vicuna models with their official code[3] using Pile (Gao et al., 2020) calibration set. For weight-only quantization of OmniQuant, we reproduce results with their official code[4] using WikiText2 (Merity et al., 2016) calibration set. We only activate the option "–lwc" to learn the weight clipping parameters for both LLaMA and Vicuna models, following (Shao et al., 2024). Additionally, for OmniQuant+IntactKV$_{[B]}$, we directly integrate IntactKV$_{[B]}$ into the training process of OmniQuant to adapt the weight clipping parameters to IntactKV$_{[B]}$, which is found to be effective and introduces no extra training costs.

Weight and Activation Quantization. For weight and activation quantization of OmniQuant, it is difficult to integrate IntactKV$_{[B]}$ into training with the learnable equivalent transformation, so we reuse the official checkpoint of LLaMA and LLaMA-2 models. When combining IntactKV with OmniQuant, we quantize IntactKV to lower bits to avoid additional inference overhead. We do not include OmniQuant results on LLaMA-3 models since the option "–let" is not compatible with GQA (Group Query Attention). For QuaRot (Ashkboos et al., 2024), we reproduce all the results with their official code[5] using WikiText2 (Merity et al., 2016) calibration set. We do not quantize IntactKV to lower bits since QuaRot adopts a

[2] https://github.com/AutoGPTQ/AutoGPTQ

[3] https://github.com/mit-han-lab/llm-awq

[4] https://github.com/OpenGVLab/OmniQuant

[5] https://github.com/spcl/QuaRot

Model	Method	MMLU (0 shot)					MMLU (5 shot)				
		Hums	STEM	Social	Others	Avg	Hums	STEM	Social	Others	Avg
Vicuna-v1.5-7B	FP16	45.40%	38.67%	56.16%	55.92%	48.74%	45.78%	39.50%	58.14%	57.46%	49.84%
	RTN	42.06%	34.16%	50.47%	50.59%	44.17%	40.68%	38.60%	50.31%	50.56%	44.62%
	GPTQ	39.89%	33.00%	48.10%	48.46%	42.19%	40.30%	36.28%	50.76%	50.09%	43.99%
	OmniQuant	42.72%	36.38%	51.93%	53.55%	**45.88%**	42.70%	37.97%	54.31%	53.08%	46.62%
	AWQ	42.08%	35.55%	51.61%	51.54%	44.95%	42.55%	38.93%	53.10%	52.78%	46.45%
	+IntactKV$_{[B]}$	42.42%	35.42%	51.71%	51.57%	45.06%	42.95%	38.60%	54.37%	53.15%	**46.87%**
Vicuna-v1.5-13B	FP16	50.48%	43.70%	62.72%	62.74%	54.54%	51.97%	44.96%	65.26%	62.40%	55.78%
	RTN	46.61%	41.32%	58.92%	57.53%	50.69%	47.14%	42.81%	59.38%	58.17%	51.44%
	GPTQ	48.35%	40.99%	59.25%	57.99%	51.38%	49.63%	43.04%	60.22%	60.09%	52.95%
	OmniQuant	49.73%	41.02%	59.31%	58.33%	51.94%	49.18%	44.17%	60.45%	58.91%	52.82%
	AWQ	48.82%	41.72%	61.03%	58.30%	52.16%	49.52%	43.01%	61.72%	58.73%	52.92%
	+IntactKV$_{[B]}$	49.31%	42.18%	61.20%	59.28%	**52.68%**	50.31%	43.37%	61.91%	59.93%	**53.58%**
Vicuna-v1.3-7B	FP16	44.31%	36.28%	53.23%	53.70%	46.71%	44.23%	38.34%	53.82%	53.15%	47.12%
	RTN	38.09%	31.58%	42.35%	44.32%	39.06%	36.81%	32.77%	43.87%	44.79%	39.33%
	GPTQ	39.09%	32.57%	44.59%	46.73%	40.66%	36.94%	33.90%	45.08%	45.81%	40.12%
	OmniQuant	41.40%	34.06%	48.07%	48.06%	42.82%	40.98%	35.19%	48.23%	48.03%	42.95%
	AWQ	40.49%	32.44%	47.06%	49.57%	42.29%	39.64%	36.22%	48.72%	49.11%	43.08%
	+IntactKV$_{[B]}$	41.76%	32.94%	47.74%	49.72%	**43.01%**	41.93%	36.58%	50.37%	50.77%	**44.67%**
Vicuna-v1.3-13B	FP16	47.89%	39.96%	58.86%	57.34%	50.77%	49.78%	40.46%	60.61%	58.24%	52.10%
	RTN	42.06%	32.87%	47.61%	49.51%	43.02%	42.42%	34.46%	50.34%	51.57%	44.56%
	GPTQ	45.06%	35.88%	52.23%	51.26%	46.09%	45.82%	37.57%	54.83%	53.64%	47.83%
	OmniQuant	43.51%	36.85%	52.16%	53.05%	46.18%	45.91%	37.44%	55.31%	54.94%	48.23%
	AWQ	45.14%	36.18%	52.55%	53.79%	46.84%	46.65%	37.64%	55.54%	54.87%	48.56%
	+IntactKV$_{[B]}$	45.91%	36.65%	53.75%	54.60%	**47.64%**	46.57%	38.40%	56.03%	55.95%	**49.05%**
Vicuna-v1.3-33B	FP16	53.73%	44.14%	67.63%	63.54%	56.98%	57.66%	46.32%	69.32%	64.25%	59.30%
	RTN	49.88%	40.13%	61.33%	58.42%	52.26%	51.26%	42.54%	61.75%	57.71%	53.18%
	GPTQ	51.22%	40.03%	61.85%	59.47%	53.05%	54.05%	44.04%	64.35%	61.35%	55.84%
	OmniQuant	51.22%	42.18%	64.06%	60.39%	54.21%	53.94%	44.10%	63.21%	59.81%	55.21%
	AWQ	51.69%	42.74%	63.41%	61.38%	54.57%	54.56%	44.10%	65.36%	60.67%	56.09%
	+IntactKV$_{[B]}$	52.09%	42.68%	63.70%	62.03%	**54.91%**	55.79%	44.90%	65.62%	61.47%	**56.91%**

Table 10: INT3-group128 weight-only quantization results of Vicuna models on MMLU benchmarks.

mixed-precision self-attention quantization strategy and can not utilize the integer multiplications for self-attention operations. Therefore, maintaining IntactKV in FP16 will not bring any extra inference costs for QuaRot.

F Evaluation Details

PPL. We evaluate PPL following the new evaluation setting in GPTQ official code[6], except that we substitute the first token of each text segment with [BOS] token to evaluate the performance of IntactKV.

MMLU. We evaluate MMLU following the original MMLU implementation[7] for 0-shot and 5-shot tasks. We note that when using Vicuna, it is considered more appropriate to fit the input sequences into the Vicuna system prompt. However, the original MMLU implementation does not use the Vicuna system prompt for Vicuna models. In our experiments on Vicuna models, we find that naively fitting the original MMLU prompt into the Vicuna system prompt will harm the final accuracy. Since prompt engineering is out of scope for this paper, we choose to follow the original evaluation setting that does not use the Vicuna system prompt for MMLU evaluation on Vicuna models.

Common Sense Reasoning Tasks. For the seven zero-shot common sense reasoning tasks, we adopt the open-sourced lm-evaluation-harness[8] library for evaluation. Similar to PPL evaluation, to assess the performance of IntactKV, we prepend [BOS] token to the beginning of each input sequence. For the evaluation of Vicuna models, we also follow the evaluation protocol in lm-evaluation-harness and do not use a system prompt.

MT-bench. MT-bench employs a GPT-4 model to score the generated content. In our experiments, we find that the scores given by GPT-4 can vary for the same generated content even when the gen-

[6] https://github.com/ist-daslab/gptq

[7] https://github.com/hendrycks/test/pull/13

[8] https://github.com/EleutherAI/lm-evaluation-harness

Model	Method	MMLU (0 shot)					MMLU (5 shot)				
		Hums	STEM	Social	Others	Avg	Hums	STEM	Social	Others	Avg
Vicuna-v1.5-7B	FP16	45.40%	38.67%	56.16%	55.92%	48.74%	45.78%	39.50%	58.14%	57.46%	49.84%
	RTN	44.65%	38.47%	53.95%	54.41%	47.61%	44.87%	39.13%	56.45%	55.34%	48.59%
	GPTQ	44.87%	37.08%	54.44%	53.86%	47.37%	45.44%	38.83%	57.33%	56.14%	49.10%
	OmniQuant	44.97%	38.80%	55.57%	56.32%	**48.59%**	45.53%	39.40%	57.20%	57.50%	**49.53%**
	AWQ	45.08%	37.41%	55.64%	55.31%	48.11%	45.44%	38.97%	56.94%	55.74%	48.95%
	+IntactKV$_{[B]}$	45.25%	37.51%	55.93%	55.58%	48.31%	45.33%	39.60%	57.36%	55.74%	49.14%
Vicuna-v1.5-13B	FP16	50.48%	43.70%	62.72%	62.74%	54.54%	51.97%	44.96%	65.26%	62.40%	55.78%
	RTN	50.01%	43.41%	62.33%	62.00%	54.06%	51.31%	43.14%	63.54%	61.63%	54.61%
	GPTQ	50.20%	42.31%	61.62%	61.41%	53.60%	50.10%	43.97%	62.72%	61.01%	54.07%
	OmniQuant	49.99%	43.97%	62.40%	62.03%	**54.19%**	51.67%	43.90%	63.05%	61.81%	54.84%
	AWQ	50.10%	42.94%	61.68%	61.66%	53.77%	52.31%	44.43%	63.18%	61.84%	55.20%
	+IntactKV$_{[B]}$	50.14%	42.84%	61.78%	61.91%	53.84%	52.31%	44.37%	63.67%	61.91%	**55.31%**
Vicuna-v1.3-7B	FP16	44.31%	36.28%	53.23%	53.70%	46.71%	44.23%	38.34%	53.82%	53.15%	47.12%
	RTN	42.78%	36.55%	51.74%	51.48%	45.41%	42.23%	37.08%	52.10%	51.94%	45.53%
	GPTQ	43.40%	34.46%	52.06%	53.45%	45.70%	43.78%	36.41%	53.49%	52.41%	46.32%
	OmniQuant	43.12%	34.59%	52.45%	52.31%	45.46%	43.04%	37.67%	52.75%	53.08%	46.33%
	AWQ	43.53%	36.22%	53.01%	52.53%	46.11%	43.36%	37.74%	53.46%	52.68%	**46.52%**
	+IntactKV$_{[B]}$	43.57%	36.51%	52.29%	53.27%	**46.20%**	43.51%	37.44%	53.17%	52.62%	46.43%
Vicuna-v1.3-13B	FP16	47.89%	39.96%	58.86%	57.34%	50.77%	49.78%	40.46%	60.61%	58.24%	52.10%
	RTN	47.16%	39.00%	56.52%	56.63%	49.64%	49.25%	39.63%	57.85%	57.74%	51.03%
	GPTQ	46.95%	39.30%	57.39%	56.23%	49.74%	49.05%	39.46%	59.02%	57.65%	51.16%
	OmniQuant	47.52%	39.40%	57.98%	57.37%	**50.34%**	49.03%	40.09%	59.34%	58.11%	51.47%
	AWQ	48.03%	39.43%	56.94%	56.76%	50.15%	49.44%	40.49%	59.57%	57.65%	51.63%
	+IntactKV$_{[B]}$	47.91%	39.60%	57.69%	56.79%	50.31%	49.54%	40.23%	60.12%	57.71%	**51.74%**
Vicuna-v1.3-33B	FP16	53.73%	44.14%	67.63%	63.54%	56.98%	57.66%	46.32%	69.32%	64.25%	59.30%
	RTN	53.18%	44.27%	66.88%	62.95%	56.52%	56.73%	45.73%	68.09%	62.49%	58.18%
	GPTQ	52.92%	44.90%	67.05%	63.66%	56.77%	57.13%	45.96%	67.63%	63.11%	58.41%
	OmniQuant	53.22%	44.43%	67.73%	63.26%	56.73%	56.83%	45.46%	68.67%	62.31%	58.25%
	AWQ	53.22%	44.40%	67.63%	63.54%	56.87%	56.85%	45.69%	68.80%	63.66%	58.65%
	+IntactKV$_{[B]}$	53.37%	44.40%	67.50%	63.63%	**56.91%**	57.07%	45.96%	68.51%	63.63%	**58.70%**

Table 11: INT4-group128 weight-only quantization results of Vicuna models on MMLU benchmarks.

eration temperature of GPT-4 is set to 0. Besides, content generation for the writing and roleplay categories has a relatively high generation temperature of 0.7, which also results in variations in the final score. To faithfully assess the performance of the quantized model and decrease the variations in the final score, we run the content generation process of each model 3 times with random seeds 42, 43, and 44. We report the mean score of three trials as the final score in Table 5 and Table 13. Also, we note that GPT-4-Turbo has been shown to be smarter than GPT-4[9], and in our experiments, we find that GPT-4-Turbo can give more stable scores than GPT-4 while having a much lower price. Therefore, we evaluate the generation results on MT-bench with the latest gpt-4-0125-preview API (i.e., GPT-4-Turbo) provided by OpenAI to further reduce variations in the final score.

[9] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

G More Experiment Results

G.1 PPL Results

We provide PPL results of LLaMA and LLaMA-2 models on WikiText2 in Table 7, and PPL results of LLaMA-3 models in Table 8. These results affirm IntactKV's effectiveness in restoring the capabilities of quantized models. Moreover, in Table 9, we conduct experiments on more heterogeneous backbones like OPT and Mistral, which further proves the compatibility of our IntactKV with various LLM backbones.

G.2 MMLU Results

We provide INT3-group128 weight-only quantization results on MMLU in Table 10, and INT4-group128 weight-only quantization results on MMLU in Table 11. For INT3-group128 quantization, AWQ+IntactKV consistently improves AWQ in every experiment setting and outperforms OmniQuant for nine out of ten settings. For INT4-group128 quantization, AWQ+IntactKV leads to relatively less improvement over AWQ compared

Model	#bits	Method	OBQA	WinoGrande	ARC-C	ARC-E	BoolQ	HellaSwag	LAMBADA	Avg
Vicuna-V1.5-7B	FP16	-	45.00%	69.53%	45.73%	71.25%	80.92%	73.78%	71.12%	65.33%
	w3g128	RTN	40.60%	66.22%	43.77%	67.89%	77.86%	71.46%	61.75%	61.36%
		GPTQ	39.40%	64.72%	40.87%	65.07%	74.77%	66.32%	59.09%	58.61%
		OmniQuant	43.00%	66.46%	43.69%	67.72%	78.59%	70.53%	66.12%	62.30%
		AWQ	41.60%	67.56%	42.66%	67.85%	78.96%	71.32%	65.28%	62.18%
		+IntactKV$_{[B]}$	42.20%	67.64%	41.98%	68.52%	79.02%	71.24%	66.82%	**62.49%**
	w4g128	RTN	43.40%	68.98%	44.80%	71.09%	82.05%	73.32%	69.28%	64.70%
		GPTQ	43.60%	69.77%	44.62%	70.20%	74.01%	72.61%	68.27%	63.30%
		OmniQuant	43.40%	69.06%	44.37%	71.17%	81.83%	72.90%	70.13%	64.69%
		AWQ	43.80%	68.59%	45.73%	71.09%	82.02%	73.51%	69.42%	64.88%
		+IntactKV$_{[B]}$	44.00%	68.90%	45.90%	71.63%	82.29%	73.52%	69.61%	**65.12%**
Vicuna-v1.5-13B	FP16	-	45.40%	71.51%	50.68%	74.87%	85.29%	77.50%	73.43%	68.38%
	w3g128	RTN	43.60%	71.27%	48.55%	72.81%	82.91%	74.55%	69.18%	66.12%
		GPTQ	43.00%	70.09%	48.98%	72.98%	84.43%	74.80%	70.11%	66.34%
		OmniQuant	43.60%	69.85%	47.78%	71.17%	82.45%	74.16%	70.04%	65.58%
		AWQ	45.40%	69.38%	48.38%	71.89%	84.46%	75.24%	70.85%	66.51%
		+IntactKV$_{[B]}$	45.40%	70.32%	48.38%	72.14%	85.20%	75.23%	71.86%	**66.93%**
	w4g128	RTN	44.80%	71.51%	49.15%	73.78%	85.20%	76.70%	72.62%	67.68%
		GPTQ	45.80%	70.96%	50.51%	73.99%	85.47%	76.70%	73.43%	68.12%
		OmniQuant	44.40%	70.80%	50.09%	73.86%	85.29%	76.79%	72.39%	67.66%
		AWQ	45.60%	72.85%	49.49%	74.07%	85.72%	77.37%	72.37%	68.21%
		+IntactKV$_{[B]}$	45.40%	73.09%	49.57%	74.45%	85.66%	77.32%	72.75%	**68.32%**
Vicuna-V1.3-7B	FP16	-	43.80%	69.46%	44.54%	71.89%	78.07%	73.93%	69.98%	64.52%
	w3g128	RTN	41.80%	63.38%	38.91%	63.47%	76.57%	68.92%	60.29%	59.05%
		GPTQ	40.00%	65.90%	41.55%	66.16%	70.73%	69.66%	62.95%	59.56%
		OmniQuant	42.00%	66.06%	39.68%	66.67%	75.69%	70.45%	65.65%	60.89%
		AWQ	42.40%	66.69%	39.51%	65.40%	77.06%	70.53%	63.69%	60.75%
		+IntactKV$_{[B]}$	43.60%	68.43%	39.16%	67.30%	77.28%	71.20%	66.54%	**61.93%**
	w4g128	RTN	42.20%	67.80%	43.00%	70.66%	75.50%	73.16%	68.37%	62.96%
		GPTQ	45.20%	68.82%	42.41%	70.45%	67.58%	72.50%	67.40%	62.05%
		OmniQuant	43.40%	67.96%	44.28%	71.46%	76.42%	73.22%	68.81%	63.65%
		AWQ	43.60%	68.03%	43.26%	71.68%	75.87%	73.44%	68.45%	63.48%
		+IntactKV$_{[B]}$	43.80%	68.59%	42.92%	71.84%	76.79%	73.49%	69.57%	**63.86%**
Vicuna-V1.3-13B	FP16	-	45.40%	71.03%	47.70%	73.70%	82.81%	77.00%	72.91%	67.22%
	w3g128	RTN	44.00%	70.96%	44.03%	67.30%	80.40%	73.33%	64.00%	63.43%
		GPTQ	45.20%	69.77%	46.08%	70.33%	81.90%	74.89%	67.59%	**65.11%**
		OmniQuant	45.20%	69.22%	45.22%	68.90%	80.95%	74.72%	68.15%	64.62%
		AWQ	42.80%	68.98%	46.08%	68.98%	81.31%	74.97%	68.78%	64.56%
		+IntactKV$_{[B]}$	43.20%	69.46%	46.16%	69.74%	81.80%	75.11%	69.67%	65.02%
	w4g128	RTN	45.20%	71.43%	48.04%	73.15%	82.87%	76.56%	70.62%	66.84%
		GPTQ	44.60%	70.01%	47.87%	73.32%	82.23%	76.55%	71.78%	66.62%
		OmniQuant	45.60%	70.56%	46.76%	73.02%	82.81%	76.74%	70.41%	66.56%
		AWQ	45.20%	70.32%	47.27%	73.91%	82.81%	76.79%	71.32%	66.80%
		+IntactKV$_{[B]}$	45.60%	71.19%	47.10%	73.32%	82.72%	76.95%	71.38%	**66.89%**
Vicuna-V1.3-33B	FP16	-	47.80%	74.35%	51.79%	74.71%	83.91%	80.38%	73.74%	69.53%
	w3g128	RTN	46.60%	72.53%	49.06%	72.18%	83.12%	78.06%	69.73%	67.33%
		GPTQ	44.80%	71.74%	47.01%	70.12%	83.64%	77.79%	71.51%	66.66%
		OmniQuant	45.40%	73.64%	48.63%	72.35%	83.55%	77.98%	71.73%	67.61%
		AWQ	45.60%	73.32%	50.68%	71.63%	82.39%	78.55%	71.49%	67.67%
		+IntactKV$_{[B]}$	44.80%	73.56%	51.11%	72.60%	82.78%	78.55%	71.90%	**67.90%**
	w4g128	RTN	47.20%	73.88%	51.62%	74.12%	83.58%	79.86%	73.24%	69.07%
		GPTQ	47.00%	73.48%	50.85%	73.06%	83.67%	80.31%	72.50%	68.70%
		OmniQuant	48.80%	74.19%	50.68%	73.91%	83.79%	79.83%	73.28%	**69.21%**
		AWQ	47.00%	73.16%	50.85%	73.82%	84.19%	79.77%	73.32%	68.87%
		+IntactKV$_{[B]}$	45.60%	73.24%	50.94%	74.12%	84.28%	79.70%	73.14%	68.72%

Table 12: Weight-only quantization results of Vicuna models on seven 0-shot commonsense QA tasks.

with INT3-group128 quantization, but still outperforms AWQ in nine out of ten experiment settings, and performs on par with OmniQuant.

G.3 Commonsense QA Results

We conduct experiments on seven zero-shot commonsense QA tasks for the Vicuna family with both INT3-group128 and INT4-group128 weight-only quantization. The results are shown in Table 12. For INT3-group128 quantization, AWQ+IntactKV significantly surpasses all baselines in four out of five experiment settings. For INT4-group128 quantization, AWQ+IntactKV improves AWQ and outperforms OmniQuant in

Method	Vicuna-v1.5-7B	Vicuna-v1.5-13B
FP16	5.31	5.52
RTN	5.18	5.47
OmniQuant	5.09	5.48
AWQ	5.22	5.28
+IntactKV$_{[P]}$	5.32	5.35
+IntactKV$_{[P]}$+Cal	**5.36**	**5.50**

Table 13: GPT-4 evaluation of INT4-group128 weight-only quantized Vicuna-v1.5 models on MT-Bench. The scores are on a scale of 10.

four out of five experiment settings, demonstrating the superiority of IntactKV.

G.4 MT-Bench Results

We provide INT4-group128 quantization results on MT-bench in Table 13. As can be seen, IntactKV leads to an average increase of 0.09 in the final score. Remarkably, with trainable IntactKV, AWQ even matches the full-precision model under INT4 quantization, while all other methods clearly lag behind the full-precision model.

H Effectiveness of Calibrating IntactKV

We conduct more experiments on MT-Bench to further demonstrate the effectiveness of calibrating IntactKV. We adopt a commonly used fine-tuning method for quantized models that tunes the quantization bias term (used in non-symmetric quantization) in every quantization group as a baseline method, termed "group bias tuning". Both "group bias tuning" and "calibrating IntactKV" are further tuned based on "AWQ+IntactKV$_{[P]}$". We use the same calibration set containing 128 samples and train 20 epochs for a fair comparison. As shown in Table 14, although calibrating IntactKV uses fewer trainable parameters, it still achieves better or comparable results compared with group bias tuning, demonstrating the effectiveness of calibrating IntactKV. Also, we note that "calibrating IntactKV" can be adopted for any quantization setting, while "group bias tuning" is only suitable for non-symmetric and group-wise quantization, making our proposed method a more versatile calibration strategy for quantized models.

I Adapting IntactKV for Activation Quantization

It is non-trivial to integrate IntactKV into activation quantization. For activation quantization, the whole KV cache needs to be quantized to low bits to exploit integer multiplications in self-attention, which contradicts our idea of keeping pivot tokens' KV cache intact. However, as shown in Table 15, the distribution of the pivot tokens' KV cache is much smoother than that of the non-pivot tokens' KV cache, which implies that IntactKV is amenable to quantization. Therefore, we adopt a straightforward solution to adapt IntactKV for activation quantization that directly quantizes IntactKV to lower bits with RTN. As shown in Table 16, quantizing IntactKV incurs minimal accuracy loss. For example, quantizing IntactKV to 4 bits only results in an average PPL increase of 0.05 on WikiText2 compared with full-precision IntactKV, which is negligible.

Method	INT3-group128	INT4-group128
AWQ	5.17	5.28
+IntactKV$_{[P]}$	5.34	5.35
+IntactKV$_{[P]}$+gbias	5.31	5.47
+IntactKV$_{[P]}$+Cal	**5.44**	**5.50**

Table 14: Evaluation of different calibration methods on MT-bench. "gbias" denotes group bias tuning and "Cal" denotes calibrating IntactKV.

J Links to Officially Released LLMs

We provide download links to some officially released LLMs used in our experiments in Table 17.

Method	Pivot K Cache		Pivot V Cache		Non-pivot K Cache		Non-pivot V Cache	
	AbsMax	Std	AbsMax	Std	AbsMax	Std	AbsMax	Std
LLaMA-7B	3.15	0.38	0.63	0.04	13.91	1.58	2.34	0.46
LLaMA-13B	3.02	0.35	0.73	0.05	13.69	1.56	2.62	0.49
LLaMA-2-7B	2.76	0.30	0.79	0.05	14.28	1.65	2.23	0.42
LLaMA-2-13B	2.73	0.27	0.75	0.05	14.60	1.62	2.57	0.44
LLaMA-3-8B	3.30	0.37	0.57	0.03	15.86	2.19	1.54	0.27

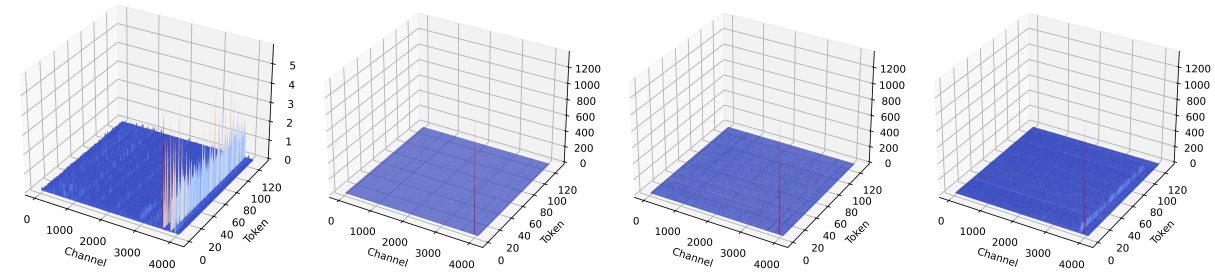
Table 15: The statistical results of pivot tokens' and non-pivot tokens' KV cache. The maximum absolute value and standard deviation are calculated on a sequence of length 1024 and averaged over all layers.

Method	LLaMA-7B		LLaMA-13B		LLaMA-2-7B		LLaMA-2-13B	
	C4	WikiText2	C4	WikiText2	C4	WikiText2	C4	WikiText2
OmniQuant	17.03	12.17	15.65	11.16	21.40	14.74	16.24	12.28
+IntactKV$_{[B]}$ (FP16)	16.26	11.30	13.89	10.00	19.97	13.61	15.77	10.94
+IntactKV$_{[B]}$	16.24	11.32	13.87	10.04	20.01	13.70	15.91	11.00

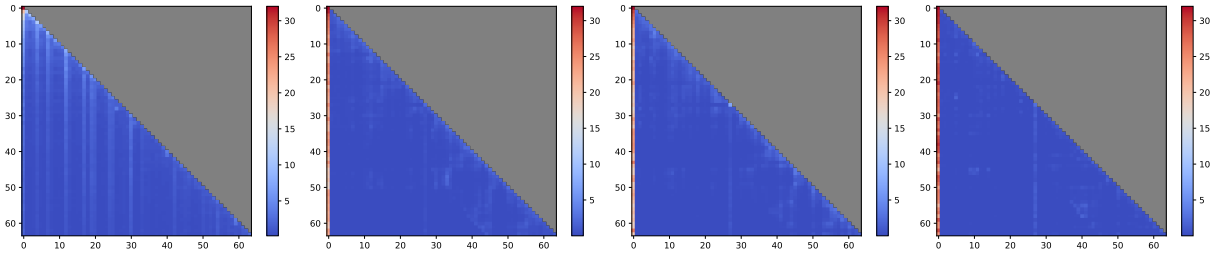
Table 16: The effect of quantizing IntactKV to lower bits. We show the INT4 weight and activation quantization results of LLaMA models on C4 and WikiText2 datasets. IntactKV$_{[B]}$ (FP16) indicates keeping IntactKV in 16 bits, which incurs extra inference costs. IntactKV$_{[B]}$ indicates quantizing IntactKV to lower bits (i.e., 4 bits).

Model	Download URL
LLaMA-2-7B	https://huggingface.co/meta-llama/Llama-2-7b
LLaMA-2-13B	https://huggingface.co/meta-llama/Llama-2-13b
LLaMA-2-70B	https://huggingface.co/meta-llama/Llama-2-70b
LLaMA-3-8B	https://huggingface.co/meta-llama/Meta-Llama-3-8B
LLaMA-3-70B	https://huggingface.co/meta-llama/Meta-Llama-3-70B
Vicuna-v1.3-7B	https://huggingface.co/lmsys/vicuna-7b-v1.3
Vicuna-v1.3-13B	https://huggingface.co/lmsys/vicuna-13b-v1.3
Vicuna-v1.3-33B	https://huggingface.co/lmsys/vicuna-33b-v1.3
Vicuna-v1.5-7B	https://huggingface.co/lmsys/vicuna-7b-v1.5
Vicuna-v1.5-13B	https://huggingface.co/lmsys/vicuna-13b-v1.5

Table 17: Download links to officially released LLMs.

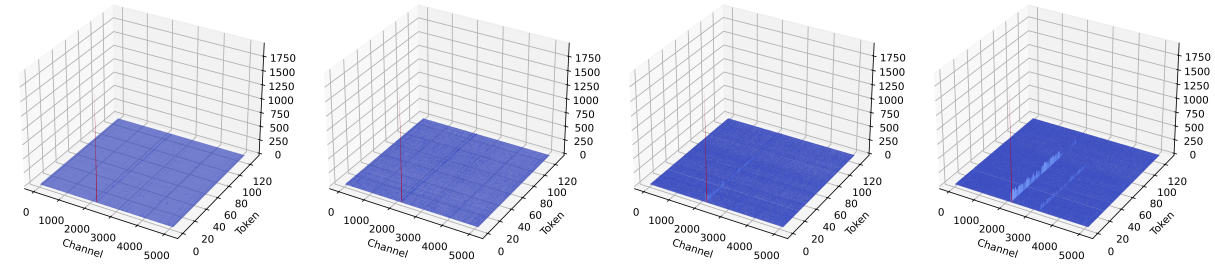


(a) Output activations of LLaMA-7B Layer 0 (b) Output activations of LLaMA-7B Layer 8 (c) Output activations of LLaMA-7B Layer 16 (d) Output activations of LLaMA-7B Layer 24

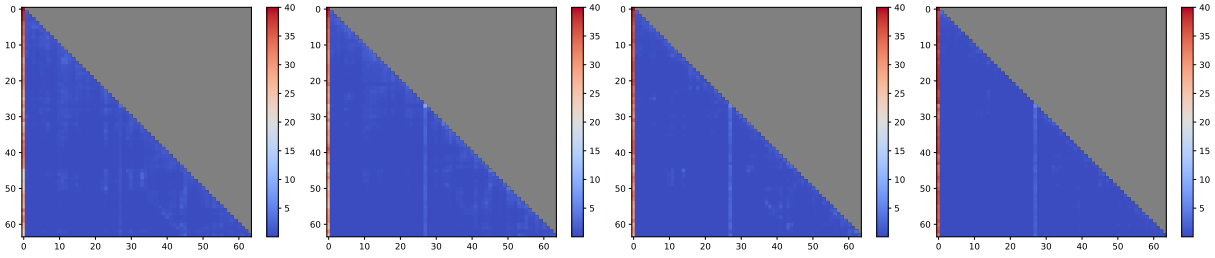


(e) Attention map of LLaMA-7B Layer 0 (f) Attention map of LLaMA-7B Layer 8 (g) Attention map of LLaMA-7B Layer 16 (h) Attention map of LLaMA-7B Layer 24

Figure 6: Magnitude of the output activations and attention map in LLaMA-7B.

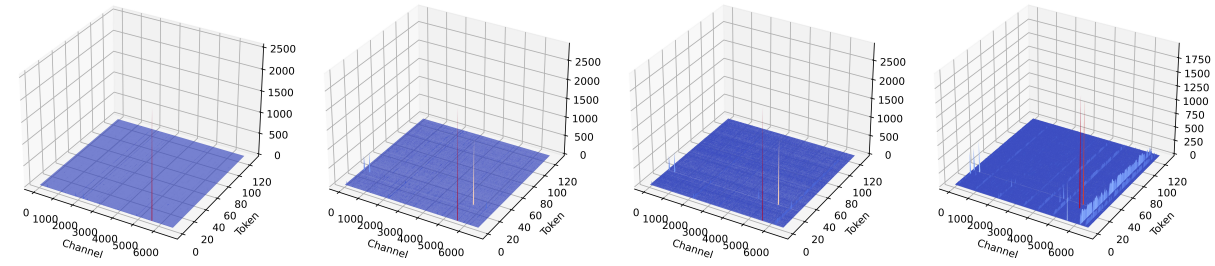


(a) Output activations of LLaMA-13B Layer 8 (b) Output activations of LLaMA-13B Layer 16 (c) Output activations of LLaMA-13B Layer 24 (d) Output activations of LLaMA-13B Layer 32

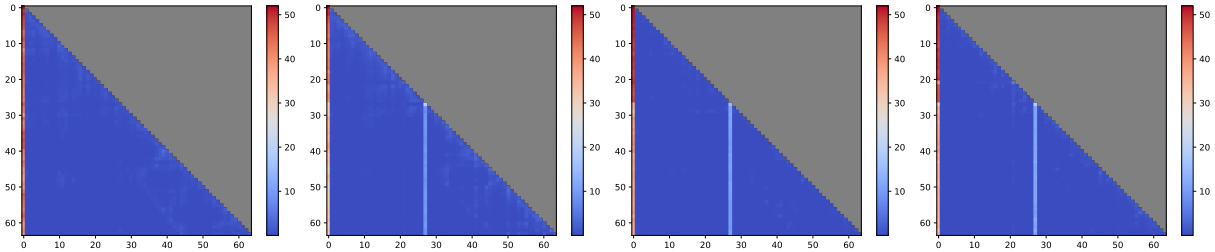


(e) Attention map of LLaMA-13B Layer 8 (f) Attention map of LLaMA-13B Layer 16 (g) Attention map of LLaMA-13B Layer 24 (h) Attention map of LLaMA-13B Layer 32

Figure 7: Magnitude of the output activations and attention map in LLaMA-13B.

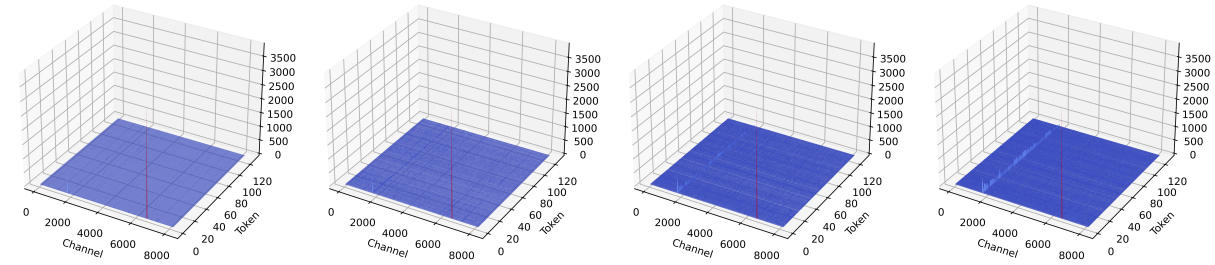


(a) Output activations of LLaMA-30B Layer 8 (b) Output activations of LLaMA-30B Layer 24 (c) Output activations of LLaMA-30B Layer 40 (d) Output activations of LLaMA-30B Layer 56

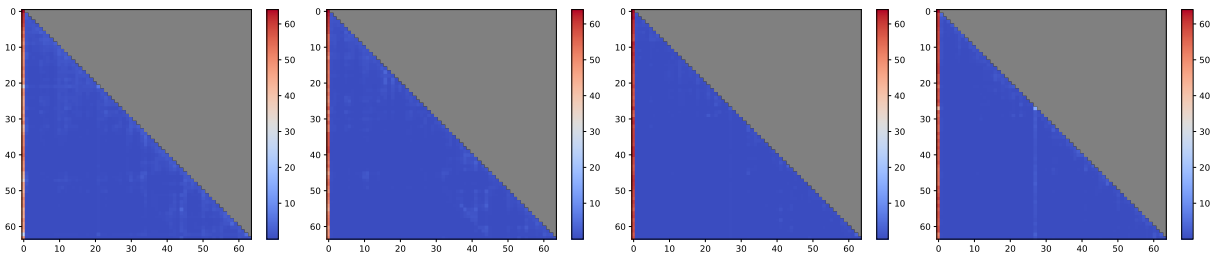


(e) Attention map of LLaMA-30B Layer 8 (f) Attention map of LLaMA-30B Layer 24 (g) Attention map of LLaMA-30B Layer 40 (h) Attention map of LLaMA-30B Layer 56

Figure 8: Magnitude of the output activations and attention map in LLaMA-30B.

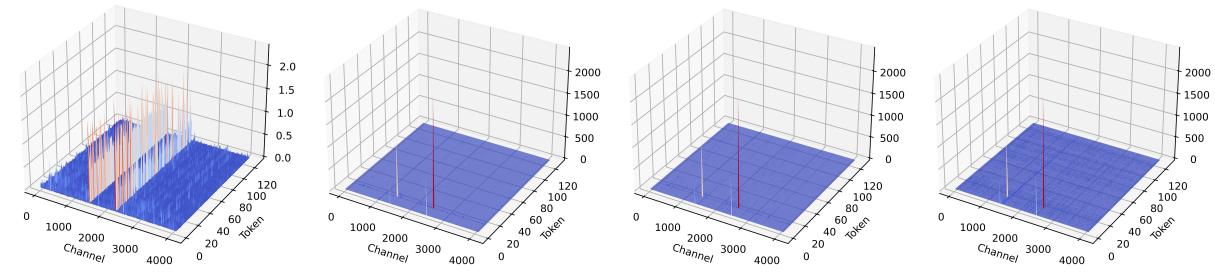


(a) Output activations of LLaMA-65B Layer 16 (b) Output activations of LLaMA-65B Layer 32 (c) Output activations of LLaMA-65B Layer 48 (d) Output activations of LLaMA-65B Layer 64

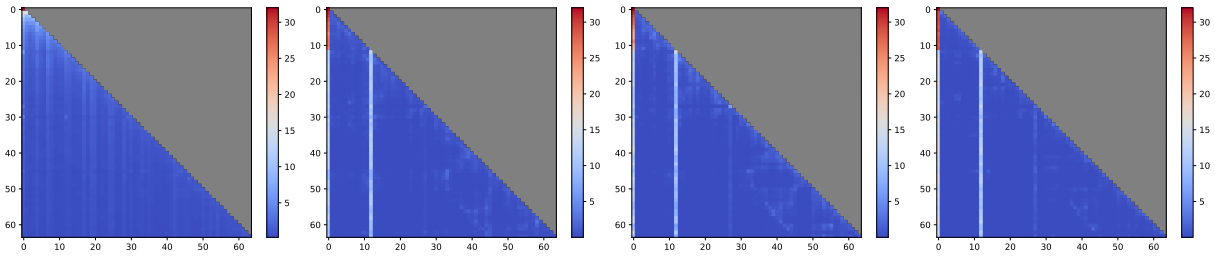


(e) Attention map of LLaMA-65B Layer 16 (f) Attention map of LLaMA-65B Layer 32 (g) Attention map of LLaMA-65B Layer 48 (h) Attention map of LLaMA-65B Layer 64

Figure 9: Magnitude of the output activations and attention map in LLaMA-65B.

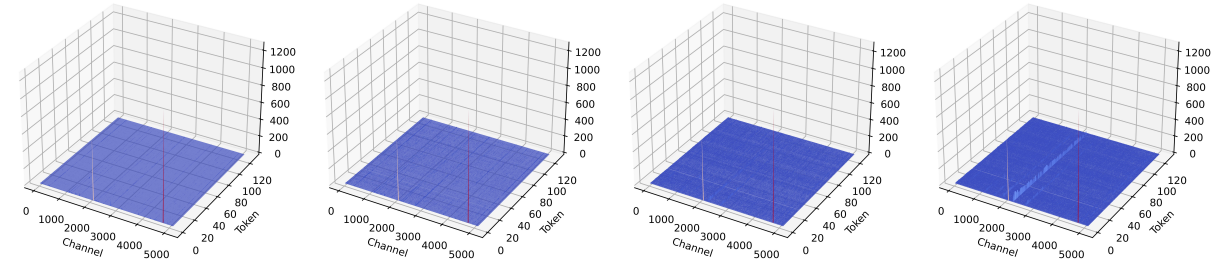


(a) Output activations of LLaMA-2-7B Layer 0 (b) Output activations of LLaMA-2-7B Layer 8 (c) Output activations of LLaMA-2-7B Layer 16 (d) Output activations of LLaMA-2-7B Layer 24

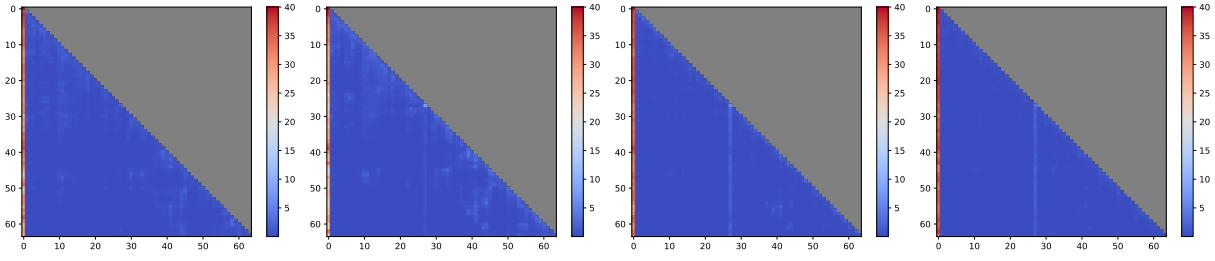


(e) Attention map of LLaMA-2-7B Layer 0 (f) Attention map of LLaMA-2-7B Layer 8 (g) Attention map of LLaMA-2-7B Layer 16 (h) Attention map of LLaMA-2-7B Layer 24

Figure 10: Magnitude of the output activations and attention map in LLaMA-2-7B.

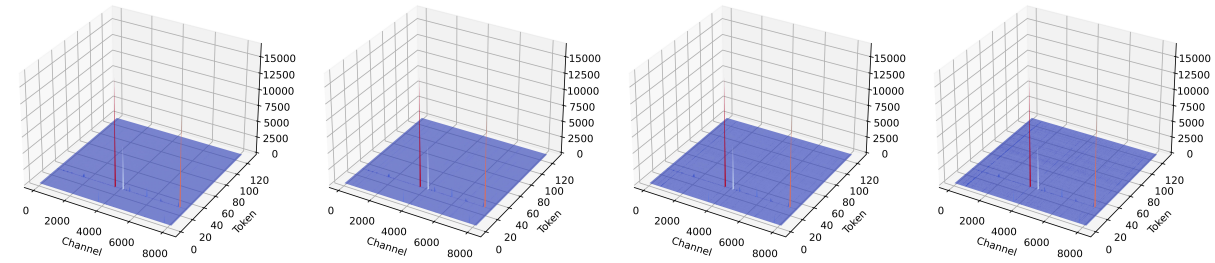


(a) Output activations of LLaMA-2-13B Layer 8 (b) Output activations of LLaMA-2-13B Layer 16 (c) Output activations of LLaMA-2-13B Layer 24 (d) Output activations of LLaMA-2-13B Layer 32



(e) Attention map of LLaMA-2-13B Layer 8 (f) Attention map of LLaMA-2-13B Layer 16 (g) Attention map of LLaMA-2-13B Layer 24 (h) Attention map of LLaMA-2-13B Layer 32

Figure 11: Magnitude of the output activations and attention map in LLaMA-2-13B.

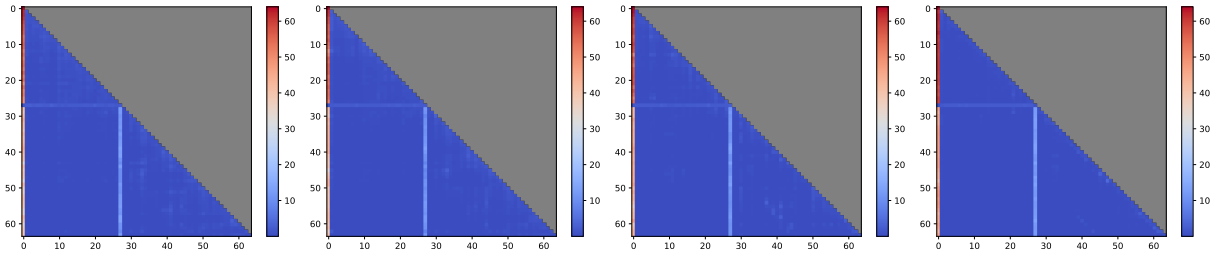


(a) Output activations of LLaMA-2-70B Layer 16

(b) Output activations of LLaMA-2-70B Layer 32

(c) Output activations of LLaMA-2-70B Layer 48

(d) Output activations of LLaMA-2-70B Layer 64



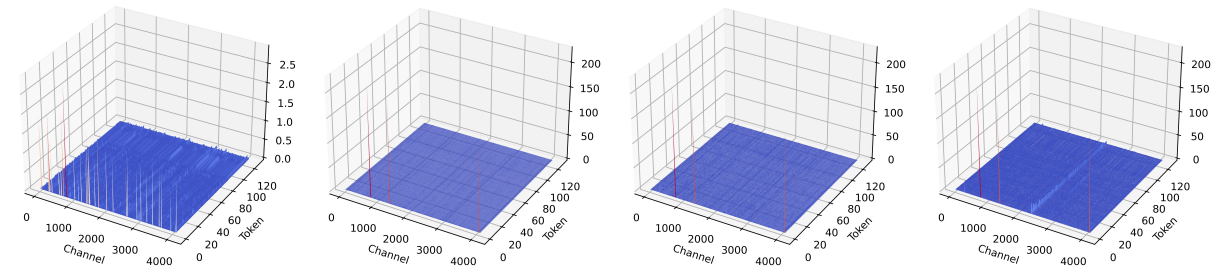
(e) Attention map of LLaMA-2-70B Layer 16

(f) Attention map of LLaMA-2-70B Layer 32

(g) Attention map of LLaMA-2-70B Layer 48

(h) Attention map of LLaMA-2-70B Layer 64

Figure 12: Magnitude of the output activations and attention map in LLaMA-2-70B.

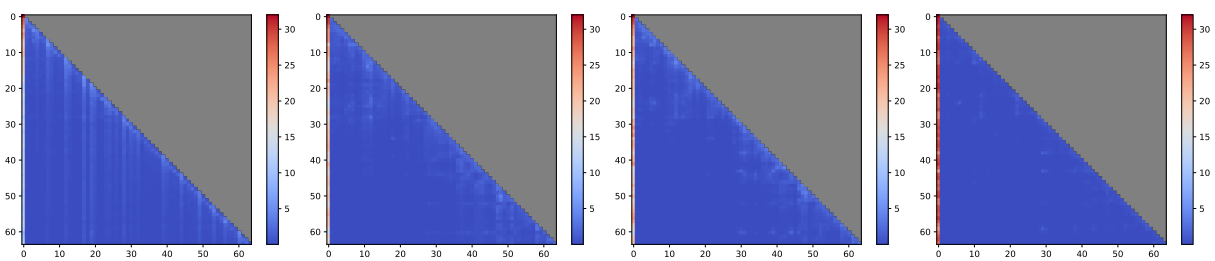


(a) Output activations of LLaMA-3-8B Layer 0

(b) Output activations of LLaMA-3-8B Layer 8

(c) Output activations of LLaMA-3-8B Layer 16

(d) Output activations of LLaMA-3-8B Layer 24



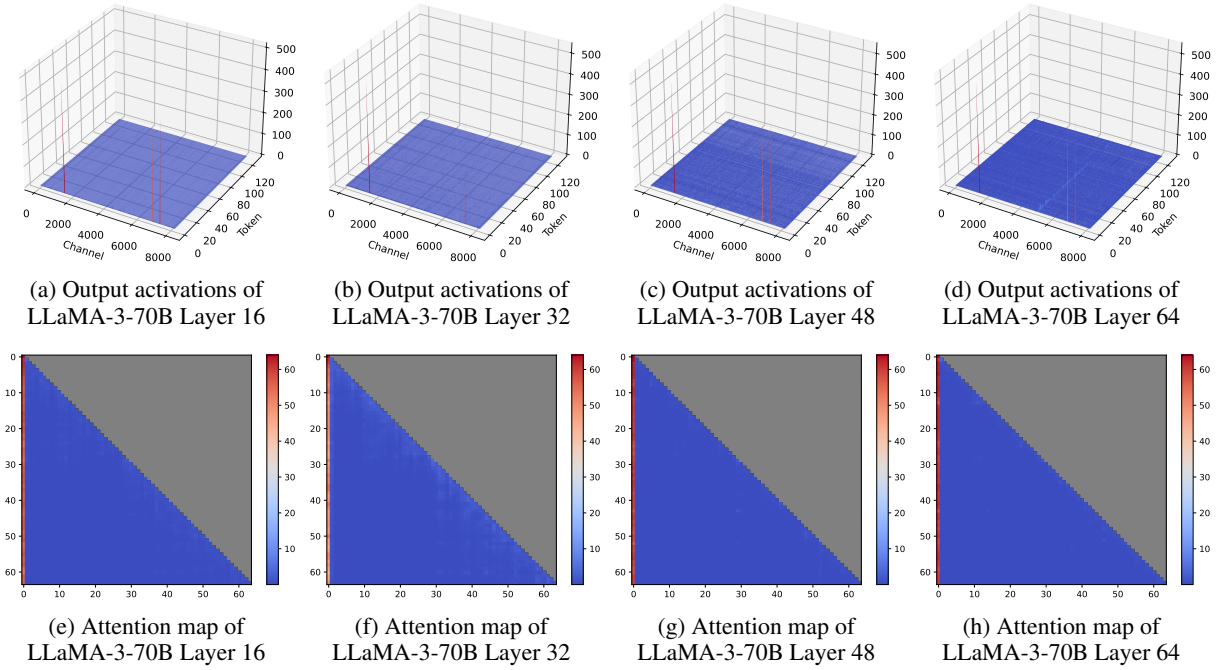
(e) Attention map of LLaMA-3-8B Layer 0

(f) Attention map of LLaMA-3-8B Layer 8

(g) Attention map of LLaMA-3-8B Layer 16

(h) Attention map of LLaMA-3-8B Layer 24

Figure 13: Magnitude of the output activations and attention map in LLaMA-3-8B.



(a) Output activations of LLaMA-3-70B Layer 16

(b) Output activations of LLaMA-3-70B Layer 32

(c) Output activations of LLaMA-3-70B Layer 48

(d) Output activations of LLaMA-3-70B Layer 64

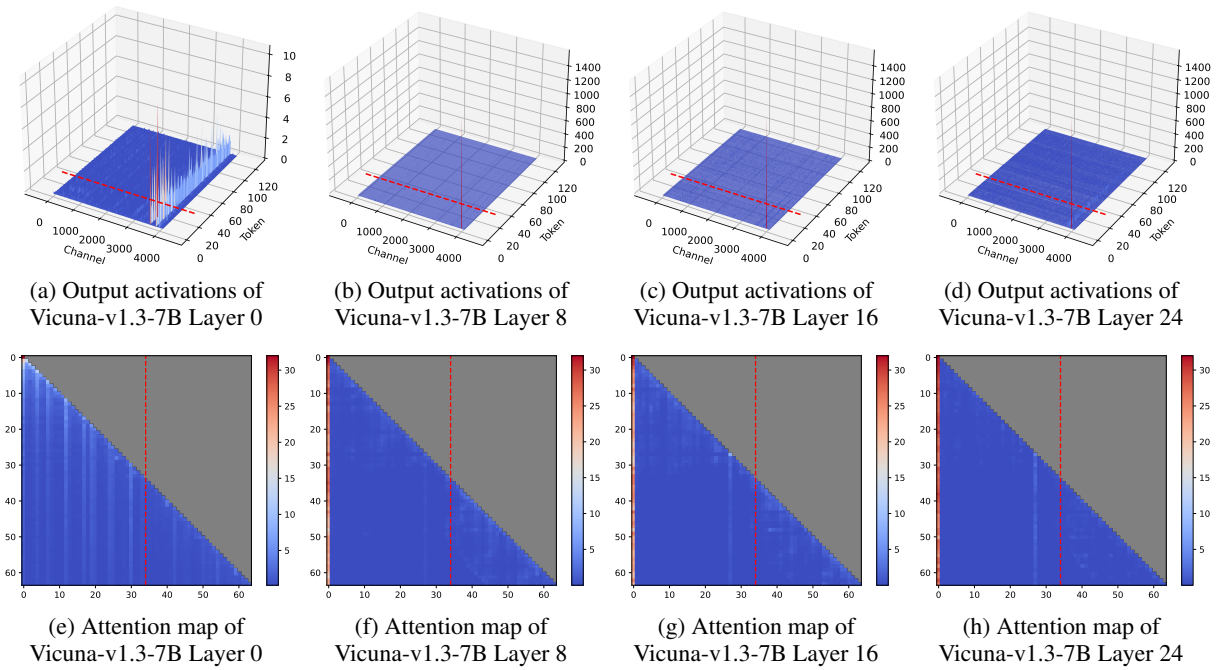
(e) Attention map of LLaMA-3-70B Layer 16

(f) Attention map of LLaMA-3-70B Layer 32

(g) Attention map of LLaMA-3-70B Layer 48

(h) Attention map of LLaMA-3-70B Layer 64

Figure 14: Magnitude of the output activations and attention map in LLaMA-3-70B.



(a) Output activations of Vicuna-v1.3-7B Layer 0

(b) Output activations of Vicuna-v1.3-7B Layer 8

(c) Output activations of Vicuna-v1.3-7B Layer 16

(d) Output activations of Vicuna-v1.3-7B Layer 24

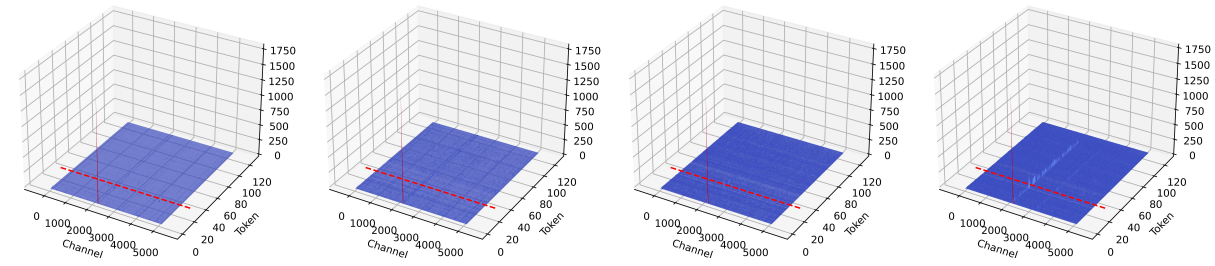
(e) Attention map of Vicuna-v1.3-7B Layer 0

(f) Attention map of Vicuna-v1.3-7B Layer 8

(g) Attention map of Vicuna-v1.3-7B Layer 16

(h) Attention map of Vicuna-v1.3-7B Layer 24

Figure 15: Magnitude of the output activations and attention map in Vicuna-v1.3-7B. The tokens before the red dashed line correspond to the Vicuna system prompt.

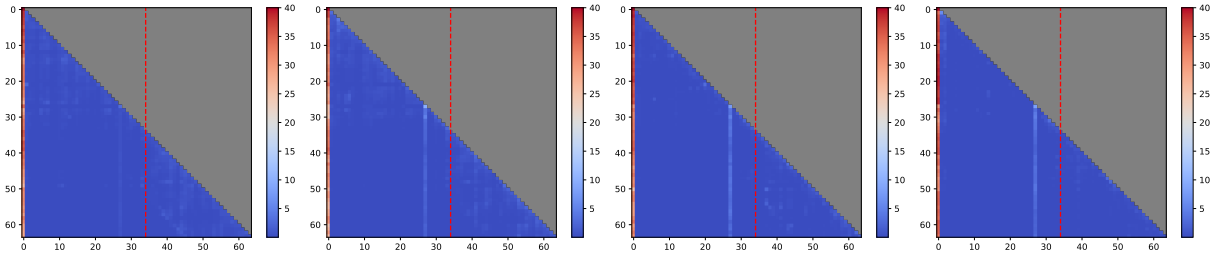


(a) Output activations of Vicuna-v1.3-13B Layer 8

(b) Output activations of Vicuna-v1.3-13B Layer 16

(c) Output activations of Vicuna-v1.3-13B Layer 24

(d) Output activations of Vicuna-v1.3-13B Layer 32



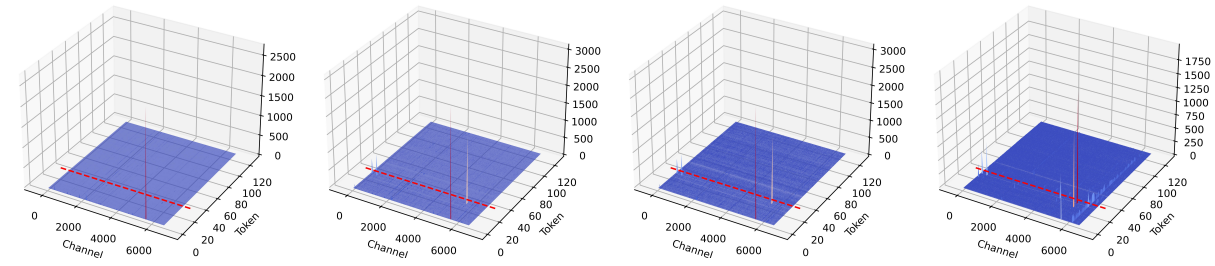
(e) Attention map of Vicuna-v1.3-13B Layer 8

(f) Attention map of Vicuna-v1.3-13B Layer 16

(g) Attention map of Vicuna-v1.3-13B Layer 24

(h) Attention map of Vicuna-v1.3-13B Layer 32

Figure 16: Magnitude of the output activations and attention map in Vicuna-v1.3-13B. The tokens before the red dashed line correspond to the Vicuna system prompt.

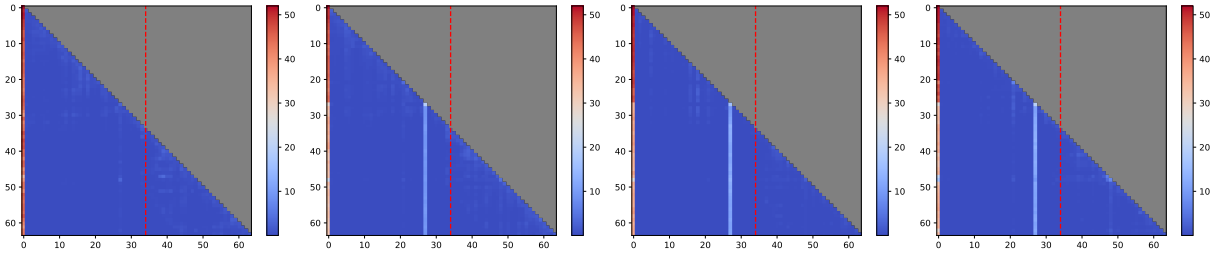


(a) Output activations of Vicuna-v1.3-33B Layer 8

(b) Output activations of Vicuna-v1.3-33B Layer 24

(c) Output activations of Vicuna-v1.3-33B Layer 40

(d) Output activations of Vicuna-v1.3-33B Layer 56



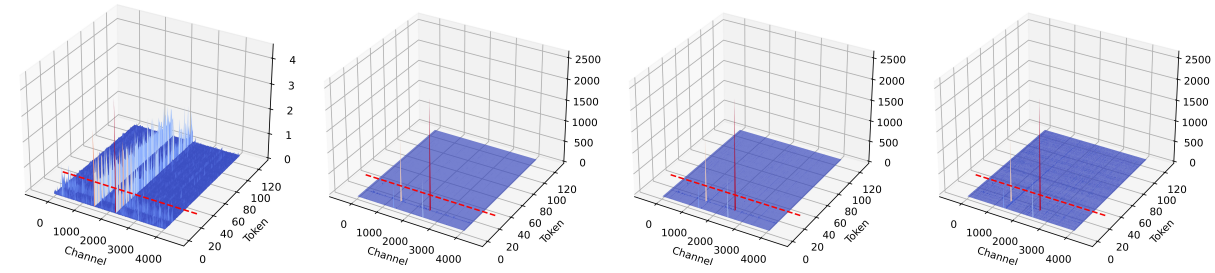
(e) Attention map of Vicuna-v1.3-33B Layer 8

(f) Attention map of Vicuna-v1.3-33B Layer 24

(g) Attention map of Vicuna-v1.3-33B Layer 40

(h) Attention map of Vicuna-v1.3-33B Layer 56

Figure 17: Magnitude of the output activations and attention map in Vicuna-v1.3-33B. The tokens before the red dashed line correspond to the Vicuna system prompt.

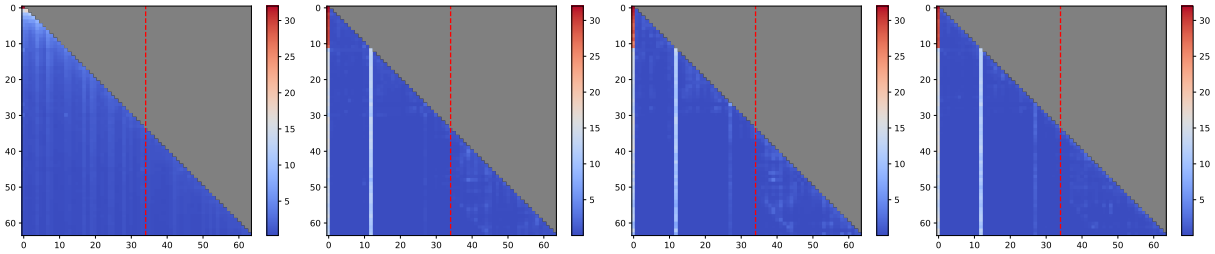


(a) Output activations of Vicuna-v1.5-7B Layer 0

(b) Output activations of Vicuna-v1.5-7B Layer 8

(c) Output activations of Vicuna-v1.5-7B Layer 16

(d) Output activations of Vicuna-v1.5-7B Layer 24



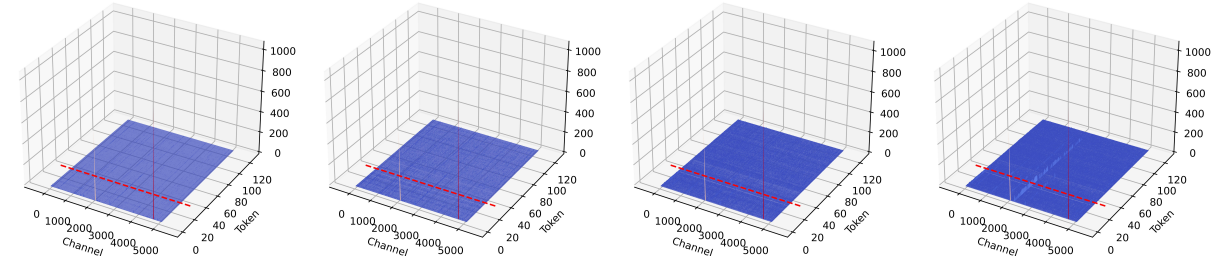
(e) Attention map of Vicuna-v1.5-7B Layer 0

(f) Attention map of Vicuna-v1.5-7B Layer 8

(g) Attention map of Vicuna-v1.5-7B Layer 16

(h) Attention map of Vicuna-v1.5-7B Layer 24

Figure 18: Magnitude of the output activations and attention map in Vicuna-v1.5-7B. The tokens before the red dashed line correspond to the Vicuna system prompt.

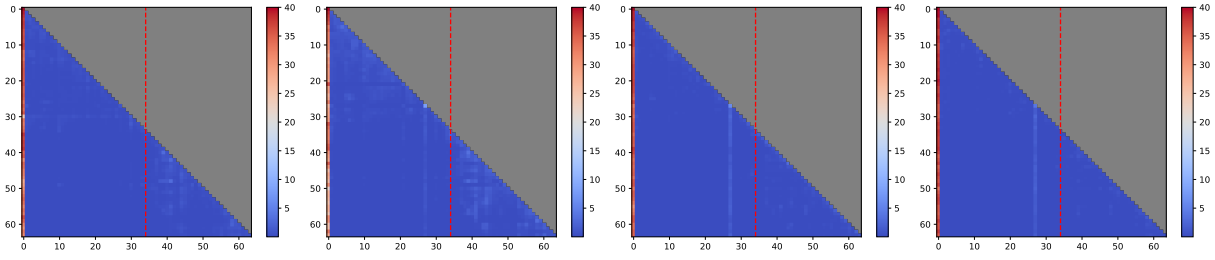


(a) Output activations of Vicuna-v1.5-13B Layer 8

(b) Output activations of Vicuna-v1.5-13B Layer 16

(c) Output activations of Vicuna-v1.5-13B Layer 24

(d) Output activations of Vicuna-v1.5-13B Layer 32



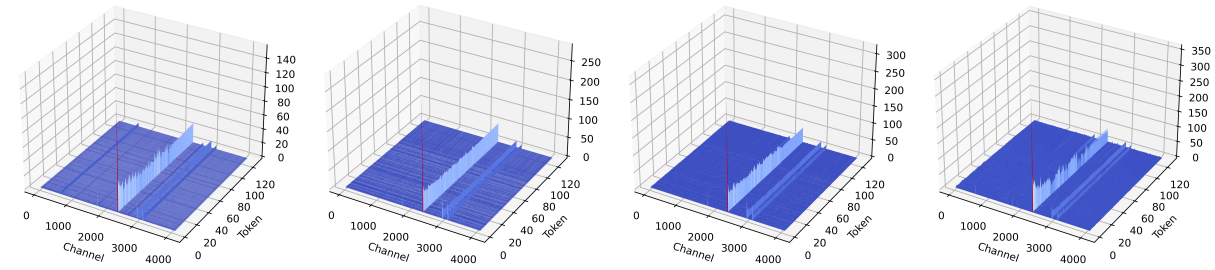
(e) Attention map of Vicuna-v1.5-13B Layer 8

(f) Attention map of Vicuna-v1.5-13B Layer 16

(g) Attention map of Vicuna-v1.5-13B Layer 24

(h) Attention map of Vicuna-v1.5-13B Layer 32

Figure 19: Magnitude of the output activations and attention map in Vicuna-v1.5-13B. The tokens before the red dashed line correspond to the Vicuna system prompt.

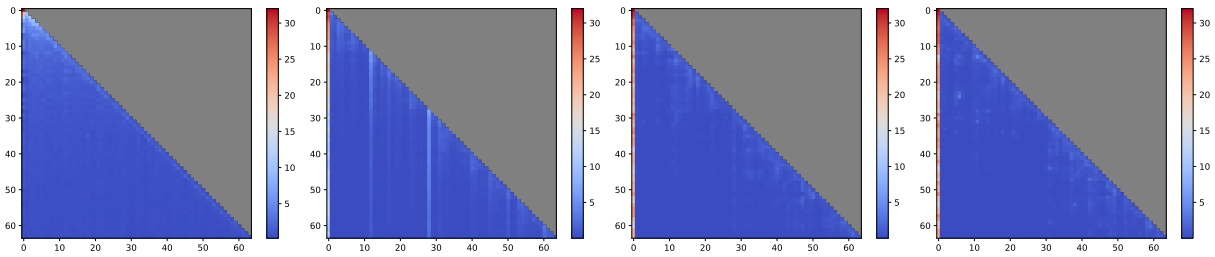


(a) Output activations of OPT-6.7B Layer 0

(b) Output activations of OPT-6.7B Layer 8

(c) Output activations of OPT-6.7B Layer 16

(d) Output activations of OPT-6.7B Layer 24



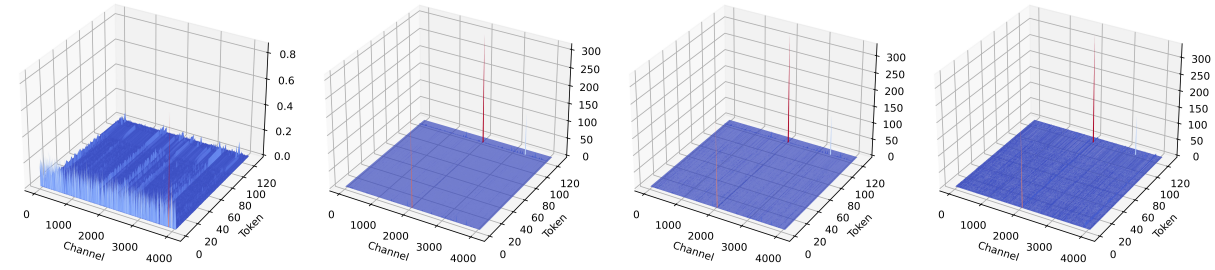
(e) Attention map of OPT-6.7B Layer 0

(f) Attention map of OPT-6.7B Layer 8

(g) Attention map of OPT-6.7B Layer 16

(h) Attention map of OPT-6.7B Layer 24

Figure 20: Magnitude of the output activations and attention map in OPT-6.7B.

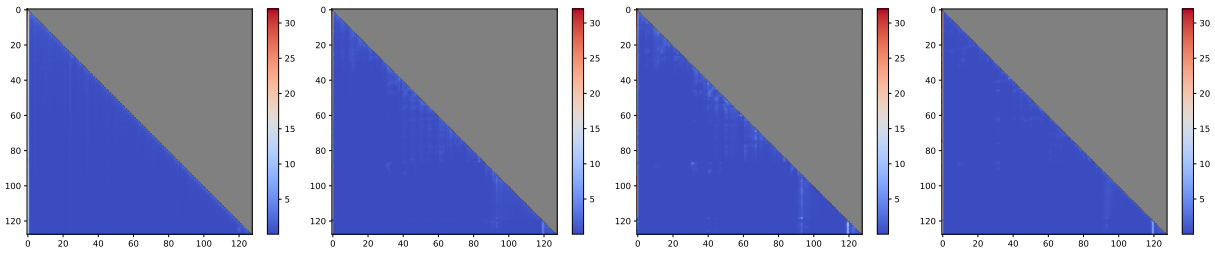


(a) Output activations of Mistral-7B Layer 0

(b) Output activations of Mistral-7B Layer 8

(c) Output activations of Mistral-7B Layer 16

(d) Output activations of Mistral-7B Layer 24



(e) Attention map of Mistral-7B Layer 0

(f) Attention map of Mistral-7B Layer 8

(g) Attention map of Mistral-7B Layer 16

(h) Attention map of Mistral-7B Layer 24

Figure 21: Magnitude of the output activations and attention map in Mistral-7B.