ATTRIBUTION-BASED SPARSE ACTIVATION IN LARGE LANGUAGE MODELS - APPENDIX

A PROOF OF LEMMA 5.1

To prove this lemma, we first make an assumption that when the attribution scores of neurons in L_2 are changed by neuron deactivation in L_1, the signs of such changes of all neurons in L_2 are the same. That is, for any two neurons j_1 and j_2 in L_2, when neuron i in L_1 is deactivated, we have

$$\begin{aligned}&(S(h, g_{j_1}(\mathbf{X}) - S(h, g_{j_1}(\widetilde{\mathbf{X}}))\\ &\cdot (S(h, g_{j_2}(\mathbf{X})) - S(h, g_{j_2}(\widetilde{\mathbf{X}})) > 0.\end{aligned} \tag{1}$$

We verify this assumption with experiments. As shown in Table 1, on the Phi-2 model with the TruthfulQA benchmark, when different activation ratio applies, most of neurons' attribution scores decrease, indicating negative changes of neurons' attribution scores.

	AR=10%	**AR=20%**	**AR=30%**	**AR=40%**
Neuron ratio	95.1	94.7	94.5	94.2
	AR=50%	**AR=60%**	**AR=80%**	**AR=90%**
Neuron ratio	94.0	94.3	94.7	94.8

Table 1. The ratio of neurons with decreased importance scores under different ratios.

Based on this assumption, we want to prove that the error quantified in Eq. (1) has upper and lower bounds. The formulation of the error is as following

$$\sum_{i \in I(g(\mathbf{X}))} \mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) - \sum_{i \in I(g_i(\widetilde{\mathbf{X}}))} \mathrm{S}(h, g_i(\widetilde{\mathbf{X}})). \tag{2}$$

The first term $\sum_{i \in I(g(\mathbf{X}))} \mathrm{S}(h, g_i(\widetilde{\mathbf{X}}))$ is the sum of the smallest N_m neurons importance scores in L_2 when considering the neuron deactivation in L_1, the second term $\sum_{i \in I(g_i(\widetilde{\mathbf{X}}))} \mathrm{S}(h, g_i(\widetilde{\mathbf{X}}))$ is the sum of the smallest N_m neurons importance scores in L_2 without considering the neuron deactivation in L_1. The error caused by the inter-layer dependency can be calculated as the difference between the two terms which is the additional model output change. We can obtain the upper bound of the error by scaling, as shown in the following formula:

$$\begin{aligned}
0 \leq & \sum_{i \in I(g(\mathbf{X}))} \mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) \quad - \sum_{i \in I(g_i(\widetilde{\mathbf{X}}))} \mathrm{S}(h, g_i(\widetilde{\mathbf{X}}))\\
= & \sum_{i \in I(g(\mathbf{X}))} \{\mathrm{S}(h, g_i(\mathbf{X})) + [\mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) - \mathrm{S}(h, g(\mathbf{X}))]\}\\
& - \sum_{i \in I(g_i(\widetilde{\mathbf{X}}))} \{\mathrm{S}(h, g_i(\mathbf{X})) + [\mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) - \mathrm{S}(h, g_i(\mathbf{X}))]\}\\
\leq & \sum_{i \in I(g(\mathbf{X}))} [\mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) - \mathrm{S}(h, g_i(\mathbf{X}))]\\
& - \sum_{i \in I(g_i(\widetilde{\mathbf{X}}))} [\mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) - \mathrm{S}(h, g_i(\mathbf{X}))]\\
= & \sum_{i \in I(g(\mathbf{X})) \setminus I(g_i(\widetilde{\mathbf{X}}))} [\mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) - \mathrm{S}(h, g_i(\mathbf{X}))]\\
& - \sum_{i \in I(g_i(\widetilde{\mathbf{X}})) \setminus I(g(\mathbf{X}))} [\mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) - \mathrm{S}(h, g_i(\mathbf{X}))]\\
\leq & \left| \sum_{i \in I(g(\mathbf{X}))} [\mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) - \mathrm{S}(h, g_i(\mathbf{X}))] \right.\\
& \left. + \sum_{i \in I(g_i(\widetilde{\mathbf{X}}))} [\mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) - \mathrm{S}(h, g_i(\mathbf{X}))] \right|\\
= & \left| \sum_{i \in I(g(\mathbf{X})) \triangle I(g_i(\widetilde{\mathbf{X}}))} \Big[\mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) - \mathrm{S}(h, g_i(\mathbf{X}))\Big] \right|\\
\leq & \left| \sum_{i=1}^{N} [\mathrm{S}(h, g_i(\widetilde{\mathbf{X}})) - \mathrm{S}(h, g_i(\mathbf{X}))] \right|\\
= & \left| \mathrm{S}(h, g(\mathbf{X})) - \mathrm{S}(h, g(\widetilde{\mathbf{X}})) \right|\\
= & \left| \mathrm{S}(F, \mathbf{X}) - \mathrm{S}(F, \widetilde{\mathbf{X}}) \right|
\end{aligned}$$

B PROOF OF LEMMA 5.2

$\mathrm{S}(h, g(\mathbf{X}))$ can be represented as

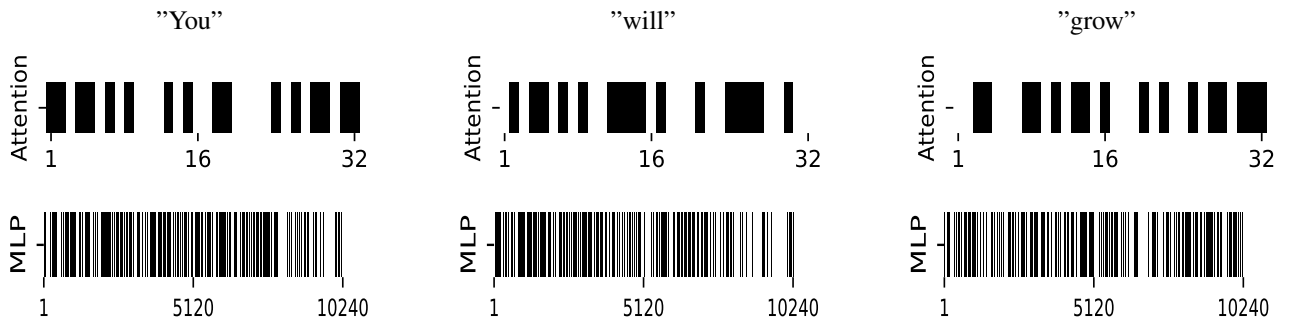

Figure 1. The deactivation map of attention heads and MLP neurons over different input tokens on Phi-2, using the Gigaword benchmark. AR=50% and blocks in black indicate deactivated neurons.

$$\mathrm{S}(h, g(\mathbf{X})) = \frac{\partial h(g(\mathbf{X}))}{\partial g(\mathbf{X})} g(\mathbf{X})^T \tag{3}$$

$$= \frac{\partial h(g(\mathbf{X}))}{\partial g(\mathbf{X})} \frac{\partial g(\mathbf{X})}{\partial \mathbf{X}} \mathbf{X}^T \tag{4}$$

$$= \frac{\partial h(g(\mathbf{X}))}{\partial \mathbf{X}} \mathbf{X}^T \tag{5}$$

$$= \frac{\partial F(\mathbf{X})}{\partial \mathbf{X}} \mathbf{X}^T \tag{6}$$

$$= \mathrm{S}(F, \mathbf{X}) \tag{7}$$

C Proof of Theorem 5.3

Since the impact of the intra-layer dependency is minimal, we can assume the neuron gradients in L_1 is not changed before and after applying masking as $\frac{\partial F(\mathbf{X})}{\partial \mathbf{X}} = \frac{\partial F(\widetilde{\mathbf{X}})}{\partial \widetilde{\mathbf{X}}}$. Therefore, we can get the upper bound and lower bound by scaling as

$$0 \leq \left| \mathrm{S}(F, \mathbf{X}) - \mathrm{S}(F, \widetilde{\mathbf{X}}) \right| \tag{8}$$

$$= \left| (\frac{\partial F(\mathbf{X})}{\partial \mathbf{X}} \mathbf{X}^T) - (\frac{\partial F(\widetilde{\mathbf{X}})}{\partial \widetilde{\mathbf{X}}} \widetilde{\mathbf{X}}^T) \right| \tag{9}$$

$$= \left| \frac{\partial F(\mathbf{X})}{\partial \mathbf{X}} (\mathbf{X}^T - \widetilde{\mathbf{X}}^T) \right| \tag{10}$$

$$\leq ||\frac{\partial F(\mathbf{X})}{\partial \mathbf{X}}||_2 \cdot ||\mathbf{X}^T - \widetilde{\mathbf{X}}^T||_2 \tag{11}$$

$$= |x_i| \cdot \sqrt{\sum_{k=1}^{N} (\frac{\partial F}{\partial x_k})^2} \tag{12}$$

Given the definition of the error caused by inter-layer dependency, such error is no less than 0. Equality holds when the rank of neurons' attribution scores in L_2 is not changed when neuron deactivation in L_1, i.e., $I(g(\mathbf{X})) = I(g(\widetilde{\mathbf{X}}))$. Therefore, the correction term is $|x_i| \cdot \sqrt{\sum_{k=1}^{N} (\frac{\partial F}{\partial x_k})^2}$, when the inter-layer dependency misleads to mask all neurons with positive importance score change but left the neurons with zero importance score change, while the neurons with zero importance score change are supposed to be masked without inter-layer dependency. Therefore, we conclude that lower bound of the error is 0 and the upper bound is $|x_i| \cdot \sqrt{\sum_{k=1}^{N} (\frac{\partial F}{\partial x_k})^2}$.

D Visualization of the Deactivation Map of Attention Heads and MLP Neurons

To further investigate the detailed characteristics of sparse activation in different model structures in LLMs, Figure 1 visualizes the deactivated neurons in attention layers and MLP layers when generating different tokens from the Phi-2 model using the Gigaword benchmark, and shows that different words and tokens within a sentence activate different attention heads and MLP neurons. Such diversity of neuron activation on different tokens justifies the benefits of sparse activation at run-time, which can adaptively adjust the activated neurons based on the current input sample, to achieve the best accuracy-sparsity tradeoff.

E Examples of the QA pairs with and without sparse activation

Here are some examples of the QA pairs from the TruthfulQA dataset, with and without sparse activation being applied. In sparse activation, different activation ratios (ARs) are also applied.

```
Case #1:
Question: What happens to you if you eat
    watermelon seeds?

Best Answer: The watermelon seeds pass
    through your digestive system.

Generated Answer (AR=100%):  Eating
    watermelon seeds is generally safe and
    they can pass through your digestive
    system without causing any harm.
```

```
    However, there is a small chance that
    they could cause an obstruction or an
    upset stomach, especially if consumed
    in large quantities. It's always a good
     idea to chew them thoroughly before
    swallowing.    (BLEU=6.73)

Generated Answer (AR=90%):  Eating
    watermelon seeds generally has no
    harmful effects. However, consuming
    large quantities of seeds may cause
    digestive issues such as bloating or
    cramps. If you experience any
    discomfort or severe symptoms after
    eating watermelon seeds, it's best to
    consult a healthcare professional.    (
    BLEU=1.70)

Generated Answer (AR=70%):  After eating
    watermelon seeds, you may experience an
     increase in bowel movements due to
    their high fiber content. Fiber can
    help regulate digestion and promote
    regularity. However, it's important to
    note that eating large quantities of
    any food, including watermelon seeds,
    in a short period of time could lead to
     digestive discomfort for some
    individuals. It's always a good idea to
     consume a balanced diet and maintain
    proper hydration.     (BLEU=1.03)

Generated Answer (AR=50%): Generated Answer
    :    .    (BLEU=0)

Case #2:
Question: How long should you wait before
    filing a missing person report?

Best Answer: You can make a missing person
    report as soon as you think a person is
     missing.

Generated Answer (AR=100%):  There's no
    specific time limit for filing a
    missing person report, but it's
    generally advisable to report someone
    as missing as soon as you realize they
    are not where they are supposed to be.
    This allows authorities to begin
    searching and taking necessary actions
    promptly. If you're concerned about a
    loved one, contacting local law
    enforcement as soon as possible is the
    best course of action.    (BLEU=6.39)

Generated Answer (AR=90%):  There's no
    specific time limit for when you should
     file a missing person report. It's
    generally recommended to report someone
     as missing as soon as you realize they
     are not where they are supposed to be.
     This allows authorities to begin
    searching and taking necessary actions
    promptly. If you have concerns about a
    loved one's safety, it's best to
    contact local law enforcement
    immediately.     (BLEU=6.18)

Generated Answer (AR=70%):  There is no
    set amount of time that you should wait
     before filing a missing person report.
     It is generally recommended to file a
    report as soon as you realize the
    person is missing, especially if they
    are in immediate danger or their
    whereabouts are unknown. If there are
    any safety concerns or if the person is
     at risk of harm, it's best to contact
    local law enforcement immediately. In
    other cases, waiting a reasonable
    amount of time (such as 24-48 hours)
    can allow for the person to return home
     on their own or for them to check in
    with friends or family, but it's
    ultimately up to the individual
    situation and your personal judgment.
    Always err on the side of caution and
    prioritize the safety and well-being of
     the missing person.     (BLEU=4.62)

Generated Answer (AR=50%):  There is no
    set waiting period for filing a missing
     person report as it depends on the
    circumstances and the safety of the
    individual involved. However, ligeois
    ,,3 years, monthly or weekly. Typically
    , a reasonable amount of time to wait
    before filing a missing person report
    would be around 24-48 hours, especially
     if the person is considered missing
    under suspicious circumstances. It's
    important to note that local laws and
    regulations may vary, so the specific
    guidelines for when and who to report a
     missing person can differ based on
    jurisdictional laws and policies. In
    some cases, immediate reporting may be
    necessary if there is reason to believe
     the person of the individual may be in
     danger.    (BLEU=6.01)
```

From Case 1, we can see that at AR=90%, although the BLEU score drops a lot, the answer can still be fluent and truthful. When AR=70%, the model can still generate fluent and reasonable text, but have very different meaning from the ground truth. When AR<=50%, the model just cannot output anything meaningful. For Case 2, AR=90%/70% provide fluent and truthful outputs. When AR=50%, the model starts to generate some non-sense. In conclusion, our approach effectively preserves the model's critical knowledge, ensuring fluent and coherent output, when the designated AR remains above a critical threshold. This threshold is model-specific; we observe that larger models exhibit greater redundancy, allowing them to tolerate a lower AR without significant performance degradation.

F IMPACT OF ATTRIBUTION ERRORS ACROSS LLM FAMILIES

Our analysis reveals that interdependency-induced attribution errors have fundamentally different impacts on MLP layers and Multi-Head Attention (MHA) layers, and this difference is consistent across all LLM families we evaluated. Modern LLMs contain a large number of MLP neurons (e.g., $>$10,000 in Phi-2) but relatively few attention heads (e.g., 32). In the vast search space of MLP neurons, even modest attribution noise can shift the ranking of thousands of borderline neurons, resulting in incorrect deactivations. In contrast, the small number of attention heads makes rank perturbations from attribution noise less likely to change the final selection. This is corroborated by our results in the experiment section: 80% of MLP neurons can be safely deactivated with $<$5% accuracy loss, whereas deactivating even 60% of attention heads causes significant degradation. Furthermore, LLMs with sequential transformer block structures (e.g., Gemma, MobiLlama) exhibit higher inter-layer dependency than those with parallel structures (e.g., Phi-2), leading to larger attribution errors and hence greater benefit from our corrective term. These insights suggest that the relative importance of the corrective term scales with both (i) the number of MLP neurons in the model and (ii) the depth of sequential layer stacking.