Supplementary Materials

I. Proofs

A. Proof about Complete probabilities

What we need to prove is that $\mathbf{p}'$ differs from the original $\mathbf{s}$ by a constant bias.

Recall, we have the following definitions:

$$\mathbf{p}' = \texttt{CLR}(\mathbf{p}) = \log\left(\frac{\mathbf{p}}{g(\mathbf{p})}\right),\ g(\mathbf{p}) = \left(\prod_{i=1}^{|\mathcal{V}|} p_i\right)^{1/|\mathcal{V}|}. \quad (1)$$

$$\mathbf{p} = \text{softmax}(\mathbf{s}) = \frac{e^{\mathbf{s}}}{\sum_{i=1}^{|\mathcal{V}|} e^{s_i}}. \quad (2)$$

Proof. Directly expanding Eq. (1), we get:

$$\begin{aligned} p'_i &= \log\left(\frac{p_i}{g(\mathbf{p})}\right) \\ &= \log(p_i) - \log g(\mathbf{p}) \\ &= \log(p_i) - \frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\log(p_i) \end{aligned}$$

For a certain $\mathbf{p}$, the sum of all $\log(p_i)$ is a constant. And logarithmic has a well-defined inverse transformation. Therefore, $\mathbf{p}'$ differs from the original $\mathbf{s}$ by a constant bias. □

B. Proof about Top-k probabilities

We aim to demonstrate that the unbiased probabilities p_i for the $k-1$ tokens can be calculated using Eq. (3) in the top-k scenario.

$$p_i = p_{\text{ref}} \cdot p_i^b / p_{\text{ref}}^b,\ 1 \le i \le |\mathcal{V}|. \quad (3)$$

Recall that we have the following definitions:

$$\mathbf{p}^b = \text{softmax}(\mathbf{s}^b),\ s_i^b = \begin{cases} s_i + b & i \in \{1, 2, \ldots, m\}, \\ s_i & \text{otherwise.} \end{cases} \quad (4)$$

Proof. First, we have the following equation:

$$p_{\text{ref}} = \frac{e^{s_r}}{\sum_{j=1}^{|\mathcal{V}|} e^{s_j}}. \quad (5)$$

$$p_i = \frac{e^{s_i}}{\sum_{j=1}^{|\mathcal{V}|} e^{s_j}}. \quad (6)$$

Then, we add bias b to the other $k-1$ tokens and reference token, assigning their indices to $\mathcal{M}$. We have the following equation:

$$p_{\text{ref}}^b = \frac{e^{s_r + b}}{\sum_{j=1, j\notin\mathcal{M}}^{|\mathcal{V}|} e^{s_j} + \sum_{j\in\mathcal{M}} e^{s_j + b}}. \quad (7)$$

$$p_i^b = \frac{e^{s_i + b}}{\sum_{j=1, j\notin\mathcal{M}}^{|\mathcal{V}|} e^{s_j} + \sum_{j\in\mathcal{M}} e^{s_j + b}}. \quad (8)$$

We can derive the new equations by rearranging Eq. (5) and Eq. (6):

$$\frac{p_{\text{ref}}}{p_i} = \frac{e^{s_r}}{e^{s_i}} \quad (9)$$

By rearranging Eq. (7) and Eq. (8), we have:

$$\begin{aligned} \frac{p_{\text{ref}}^b}{p_i^b} &= \frac{e^{s_r + b}}{e^{s_i + b}} \\ &= \frac{e^{s_r}}{e^{s_i}} \end{aligned} \quad (10)$$

Rearranging Eq. (9) and Eq. (10), we get:

$$p_i = p_{\text{ref}} \cdot \frac{p_i^b}{p_{\text{ref}}^b}. \quad (11)$$

□

C. Proof of Top-1 probabilities

Our goal is to prove that the unbiased probability p_i for token i can be calculated using Eq. (12) in the top-1 scenario.

$$p_i = (e^{b - \log p_i^b} - e^b + 1)^{-1},\ 1 \le i \le |\mathcal{V}|. \quad (12)$$

We have the following definitions:

$$\mathbf{p}^b = \text{softmax}(\mathbf{s}^b),\ s_j^b = \begin{cases} s_j + b & j = i, \\ s_j & \text{otherwise.} \end{cases} \quad (13)$$

Proof. First, we have the following equation:

$$p_i = \frac{e^{s_i}}{\sum_{j=1}^{|\mathcal{V}|} e^{s_j}}. \quad (14)$$

Then, we add bias b to the token i to the top position, we get:

$$p_i^b = \frac{e^{s_i + b}}{\sum_{j=1, j\neq i}^{|\mathcal{V}|} e^{s_j} + e^{s_i + b}}. \quad (15)$$

Rewriting the Eq. (15), we get:

$$\begin{aligned} p_i^b &= \frac{e^{s_i + b}}{\sum_{j=1, j\neq i}^{|\mathcal{V}|} e^{s_j} + e^{s_i + b}} \\ &= \frac{e^{s_i + b}}{\sum_{j=1}^{|\mathcal{V}|} e^{s_j} - e^{s_i} + e^{s_i + b}} \end{aligned} \quad (16)$$

By substituting Eq. (14) into the right-side of Eq. (16), we obtain:

$$\begin{aligned} p_i^b &= \frac{p_i \cdot \sum_{j=1}^{|\mathcal{V}|} e^{s_j} \cdot e^b}{\sum_{j=1}^{|\mathcal{V}|} e^{s_j} \cdot (1 - p_i + p_i \cdot e^b)} \\ &= \frac{p_i \cdot e^b}{1 - p_i + p_i \cdot e^b} \end{aligned} \quad (17)$$

Rewriting the Eq. (17), we get:

$$p_i^b = \frac{e^b}{p_i^{-1} - 1 + e^b} \quad (18)$$

Rearranging the Eq. (18), we get:

$$p_i^{-1} = \frac{e^b}{p_i^b} - e^b + 1 \tag{19}$$

$\square$

II. Pseudocode

The pseudocode to determine Δr is presented below for better understanding.

Algorithm 1 Pseudocode for dimension difference calculation

Input: $\mathbf{W}$: the parameter matrix of the last linear layer in the victim model; $\mathcal{M}$: the suspect model; $\mathcal{Q}$: the query set; N: the least number of samples; e: the error term.

Output: Δr: the dimension difference.

```
Initialize Δr = 0, n = 0, S = ∅, W_sum = W.
while n ≤ N do
    Randomly sample a query q from Q
    Get the logits outputs O = {s_1, s_2, ...} by querying
    the suspected model M with q
    S ← S ∪ O
    n = size(S)
end while
for i = 1, 2, ..., n do
    Solve W_sum · x_i = s_i to obtain x̂_i
    Calculate d_i = ||s − W_sum · x̂||
    if d_i > e then
        Δr = Δr + 1
        W_sum = [W_sum, s_i]
    end if
end for
return Δr
```