1 Problem setting

We define two prompts, the erroneous prompt with only the question about the fact we want to insert and the gold prompt that also contains the fact that answers the question. (e.g ”Who is the US president as of February 2025?” and ”Donald trump is the president of USA since January 20th 2025. Who is the US president as of February 2025?”)
We can then define X^{err} and X^{gld} as the input of a single FFN layer given each prompt. And reciprocally, we define Y'^{err} and Y'^{gld} as the output of this same layer for the same prompts. We can now define our knowledge insertion goal as:

$$FFN(X^{err}) \approx Y'^{gld}$$

2 Forward path

We will focus on a single FFN layer. Such that :

- X is the input to the FFN
- X' is the normalized input weighted by W_N (i.e $X' = norm(X) \times W_N$)
- Z is the output of the first linear layer parametrized by W_U (i.e $Z = X'W_U$)
- Z' is the output of the SwiGLU activation function parametrized by W_G (i.e $Z' = SiLU(X'W_G) \times Z$)
- Y is the output of the second linear layer parametrized by W_D (i.e $Y = Z'W_D$)
- Y' is the output of the FFN after skip-connection (i.e $Y' = Y + X$)

The batch dimension is given by n_{edit}, the number of edits we want to insert. The token dimension n_{tok}, however, only considers the token that we want to edit, the others are ignored. d_{model} and d_{ff} describe, respectively, the dimension of the input/output and the hidden dimension.
The tensors can be defined as such:

- $X, X', Y \in \mathbb{R}^{(n_{\text{edit}} \times n_{\text{tok}}) \times d_{\text{model}}}$
- $Z, Z' \in \mathbb{R}^{(n_{\text{edit}} \times n_{\text{tok}}) \times d_{\text{ff}}}$
- $W_N \in \mathbb{R}^{d_{\text{model}}}$
- $W_U, W_G \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$
- $W_D \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$

The compute graph of this forward path is illustrated in Fig.1. In this case, $n_{\text{edit}} \times n_{\text{tok}} = 2$, $d_{\text{model}} = 4$ and $d_{\text{ff}} = 6$

3 Update weight

Our goal is to append weights along the d_{ff} dimension that will encode our edits. Therefore, the new weights can be described as such:

$$W_U^{new} = Concat(W_U, W_U^{edit}),\ W_G^{new} = Concat(W_G, W_G^{edit}) \text{ and } W_D^{new} = Concat(W_D, W_D^{edit})$$

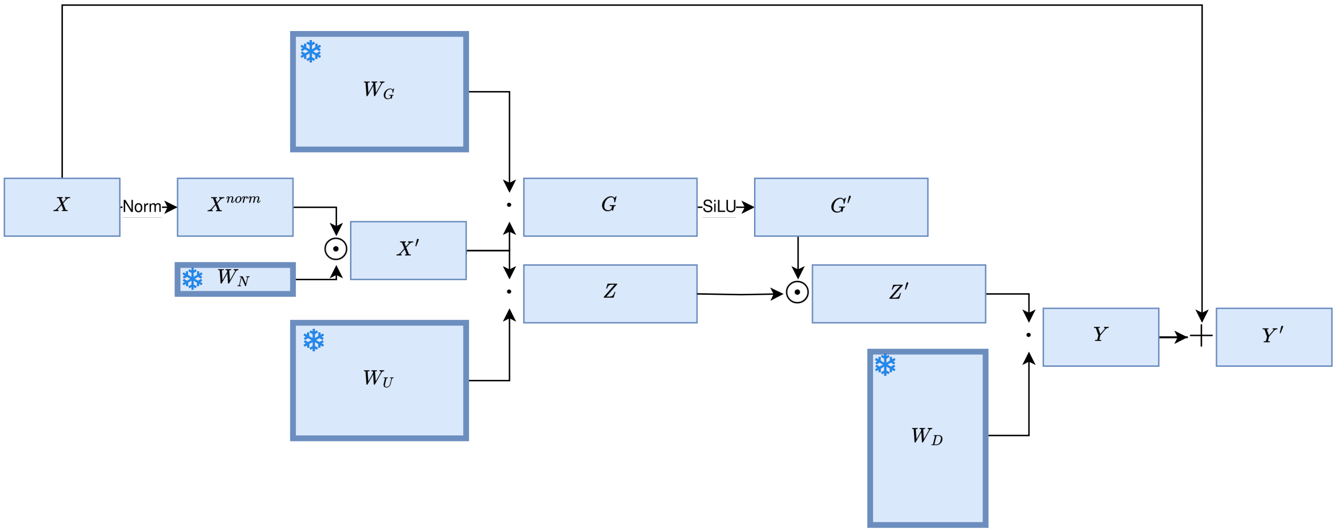

Figure 1: FFN forward path

These new edit extensions can be described similarly as the original tensors where d_{ff} is replaced by $n_{\text{edit}} \times n_{\text{tok}}$ the dimensionality of our edits:

- $Z^{edit}, Z'^{edit} \in \mathbb{R}^{(n_{\text{edit}} \times n_{\text{tok}}) \times (n_{\text{edit}} \times n_{\text{tok}})}$
- $W_U^{edit}, W_G^{edit} \in \mathbb{R}^{d_{\text{model}} \times (n_{\text{edit}} \times n_{\text{tok}})}$
- $W_D^{edit} \in \mathbb{R}^{(n_{\text{edit}} \times n_{\text{tok}}) \times d_{\text{model}}}$

4 Updated forward path

With these new definitions, we can rewrite our forward path post-edit, here illustrated in Fig.2. X is replaced by X^{err}, the input without the fact in the context of the prompt. The blue color in Fig.2 represents the matrices that are not changed by the addition of the edit weights. Therefore, we only need to rewrite the orange matrices that change post-edit. In Fig.3, we present the forward path without the frozen path that we can't modify. For simplicity, we arbitrarily decide that $W_U^{edit} = W_G^{edit}$.

- $Z^{edit} = X'^{err} W_U^{edit}$
- $$\begin{aligned} Z'^{edit} &= SiLU(X'^{err} W_G^{edit}) \times Z^{edit} \\ &= SiLU(X'^{err} W_U^{edit}) \times Z^{edit} \\ &= SiLU(Z^{edit}) \times Z^{edit} \\ &= Z^{edit^2} \times Sigmoid(Z^{edit}) \end{aligned}$$
- $Y^{new} = Z'^{new} W_D^{new}$
- $Y'^{new} = Y^{new} + X^{err}$

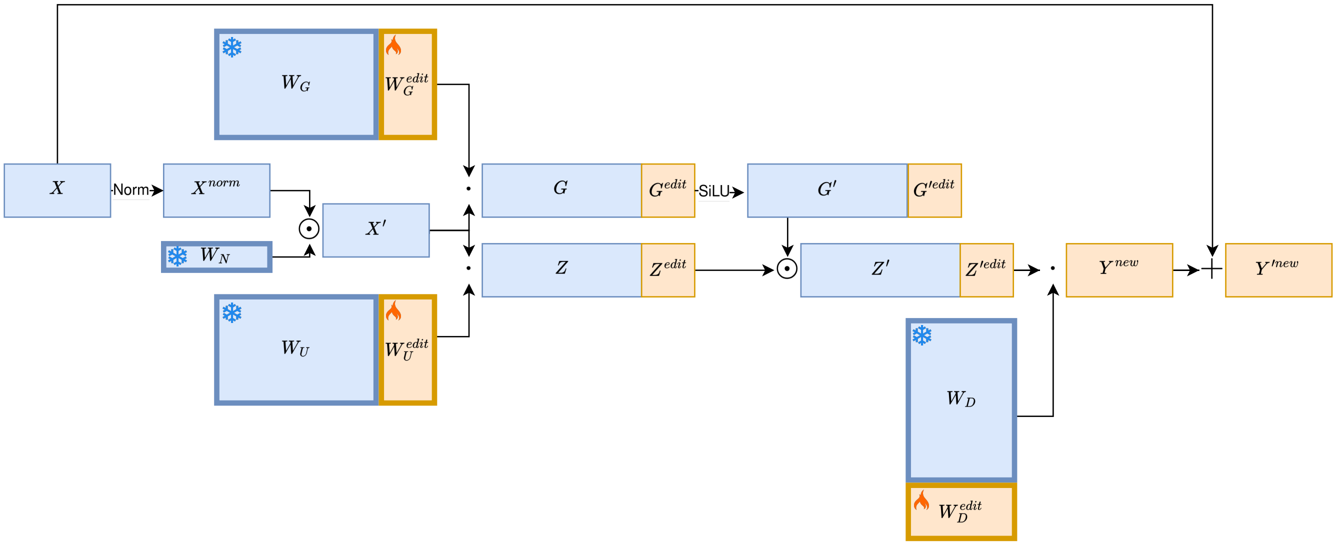

Figure 2: Updated FFN forward path

5 W_U^{edit} computation

To maximize specificity, we want Z^{edit} to trigger only when the input is related to X^{err}. As $Z^{edit} = X'^{err}W_U^{edit}$, we could simply set $W_U^{edit} = X'^{err^T}$. However, since X'^{err} is weighted by W_N, to obtain a dot product of 1, we need to divide it by the square of the 2-norm along the $n_{\text{edit}} \times n_{\text{tok}}$ dimensions, obtaining:

$$W_U^{edit} = W_G^{edit} = \frac{X'^{err^T}}{||X'^{err^T}||_2^2}$$

6 W_D^{edit} computation

To maximize edit success, we want Y'^{new} to be as close as possible to Y'^{gld}, the output with the fact in the context of the prompt. We can therefore solve the following equation:

$$\begin{aligned}
Y'^{new} &= Y^{new} + Xerr \\
Y'^{new} &= Z'^{new}W_D^{new} + Xerr \\
Y'^{gld} &= Z'^{new}W_D^{new} + Xerr \\
Y'^{gld} &= Concat(Z', Z'^{edit})Concat(W_D, W_D^{edit}) + Xerr \\
Y'^{gld} &= Z'W_D + Z'^{edit}W_D^{edit} + Xerr \\
Y'^{gld} &= Y'^{err} + Z'^{edit}W_D^{edit} \\
Y'^{gld} - Y'^{err} &= Z'^{edit}W_D^{edit} \\
W_D^{edit} &= Z'^{edit^{-1}}(Y'^{gld} - Y'^{err})
\end{aligned}$$

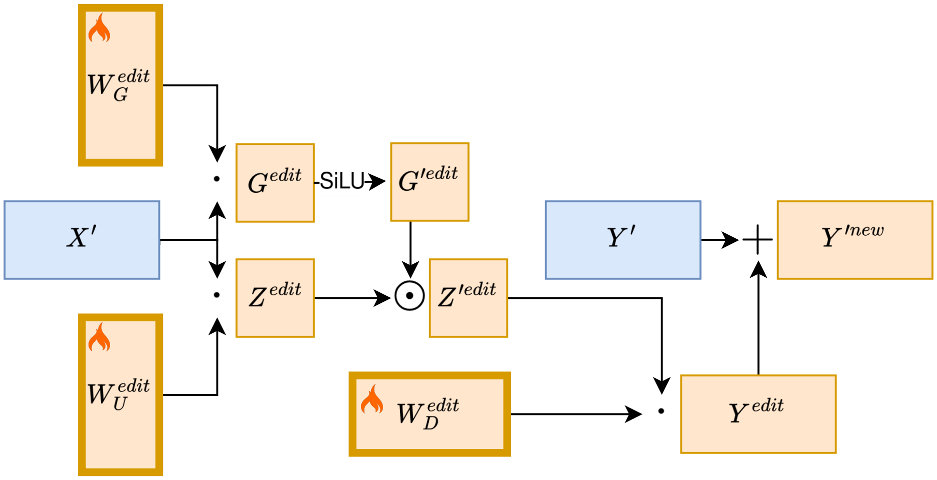

Figure 3: Updated FFN forward path without frozen path

7 Acknowledgment

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).