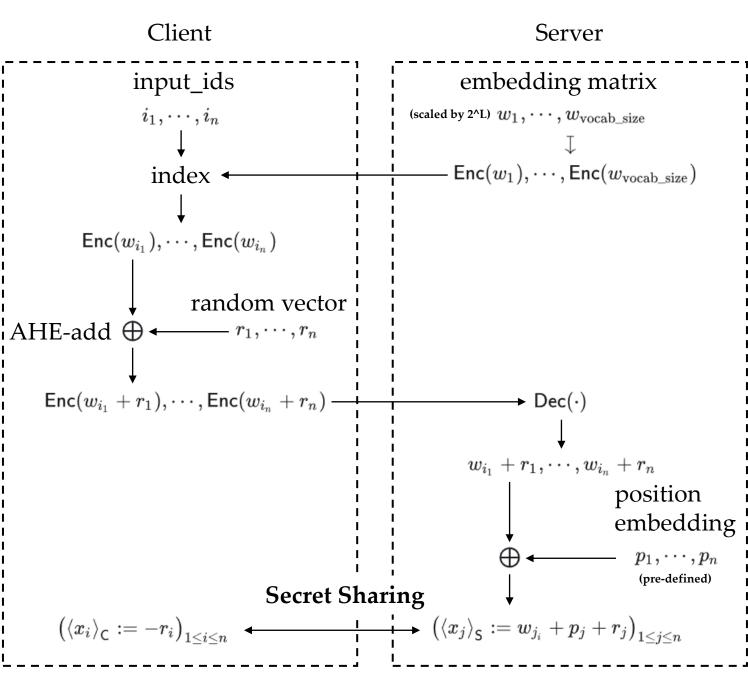
## 1. Embedding



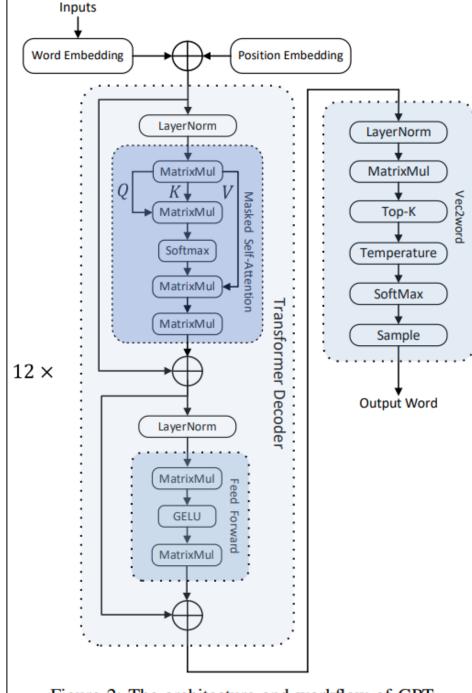


Figure 2: The architecture and workflow of GPT.

**2. LayerNorm**  $x_i^{(j)}$  denotes the j-th component of vector x\_i

Client Server  $\longrightarrow \left\langle (x_i^{(j)} - \mathbf{E}[x_i])^2 \right\rangle_{\mathsf{S}}$  $F_{LUT}$  $\langle \mathbf{Var}[x_i] \rangle_{\mathsf{C}}$  —  $-\langle \mathbf{Var}[x_i] 
angle_{\mathsf{S}}$ built-in |  $\sqrt{-} : t \mapsto rac{1}{\sqrt{t+\epsilon}}$  $\mathsf{F}_{\mathsf{Mult}}$  $\left\langle \frac{x_i^{(j)} - \mathbf{E}[x_i]}{\sqrt{\mathbf{Var}[x_i] + \epsilon}} \right
angle_{\mathsf{C}}$  $\left\langle rac{x_i^{(j)} - \mathbf{E}[x_i]}{\sqrt{\mathbf{Var}[x_i] + \epsilon}} 
ight
angle_{\mathsf{S}}$  $\left\langle \gamma \cdot \frac{x_i^{(j)} - \mathbf{E}[x_i]}{\sqrt{\mathbf{Var}[x_i] + \epsilon}} \right\rangle_{\mathsf{C}} \leftarrow \left\langle \gamma \cdot \frac{x_i^{(j)} - \mathbf{E}[x_i]}{\sqrt{\mathbf{Var}[x_i] + \epsilon}} \right\rangle_{\mathsf{S}} \xrightarrow{\text{(inplacable of secret-sharing)}} + \beta$ 

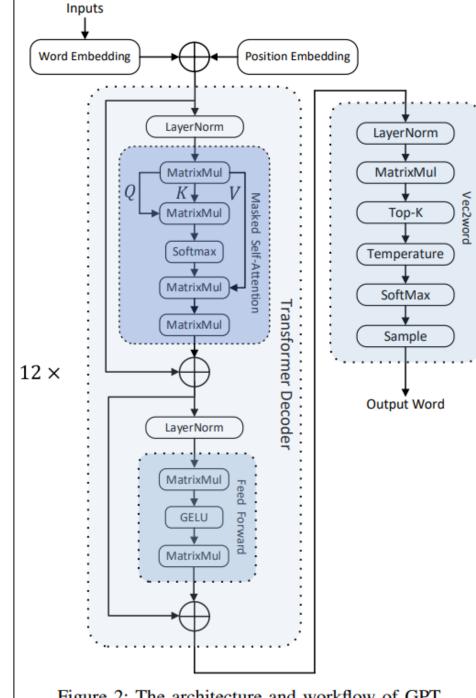
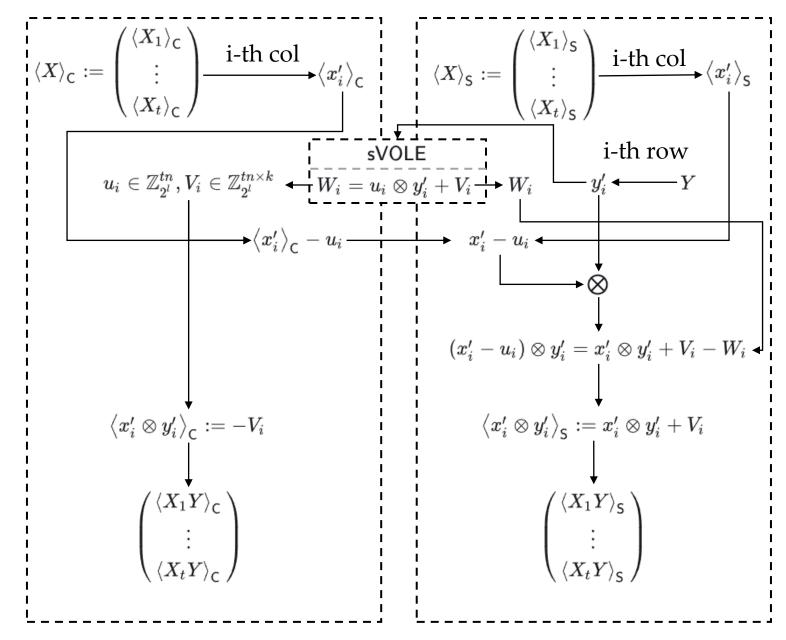


Figure 2: The architecture and workflow of GPT.

 $\textbf{0. MatrixMul} \hspace{0.5cm} \otimes : \mathbb{Z}_{2^l}^n \times \mathbb{Z}_{2^l}^k \ni (x,y) \mapsto x \otimes y := \left(x^{(j)} \cdot y^{(r)}\right)_{1 \leq j \leq n, 1 \leq r \leq k} \in \mathbb{Z}_{2^l}^{n \times k}$ 

Client

Server

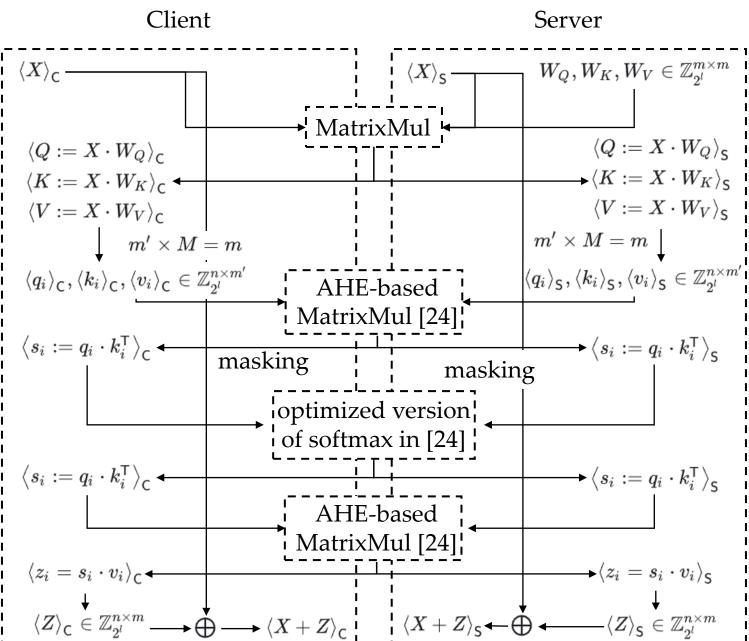


C has  $X_1, ..., X_t$  and S has Y. They want to get the secret-sharing of  $X_1 * Y, ..., X_t * Y$ .

 $\mathbf{Lemma}:X\in\mathbb{Z}_{2^{l}}^{n imes m},Y\in\mathbb{Z}_{2^{l}}^{m imes k}, ext{then}:$ 

$$X \cdot Y = \sum_{j=1}^m \underbrace{X_{\cdot,j}}_{i ext{-th col}} \otimes \underbrace{Y_{j,\cdot}}_{i ext{-th row}}$$

## 3. Self-Attention $X := (x_i^{(j)})_{1 \le i \le n, 1 \le j \le m}$



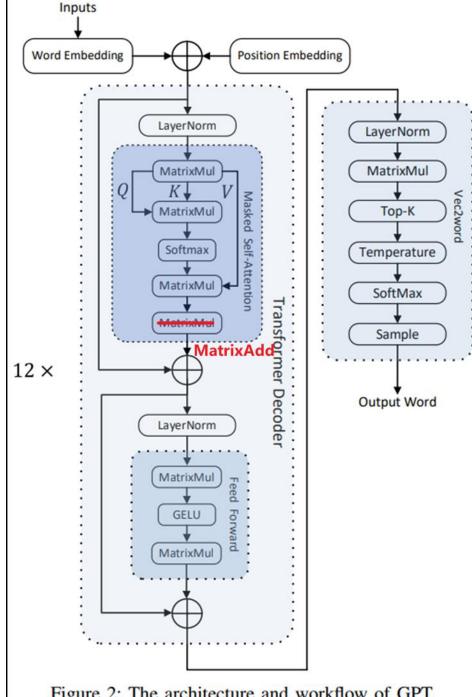
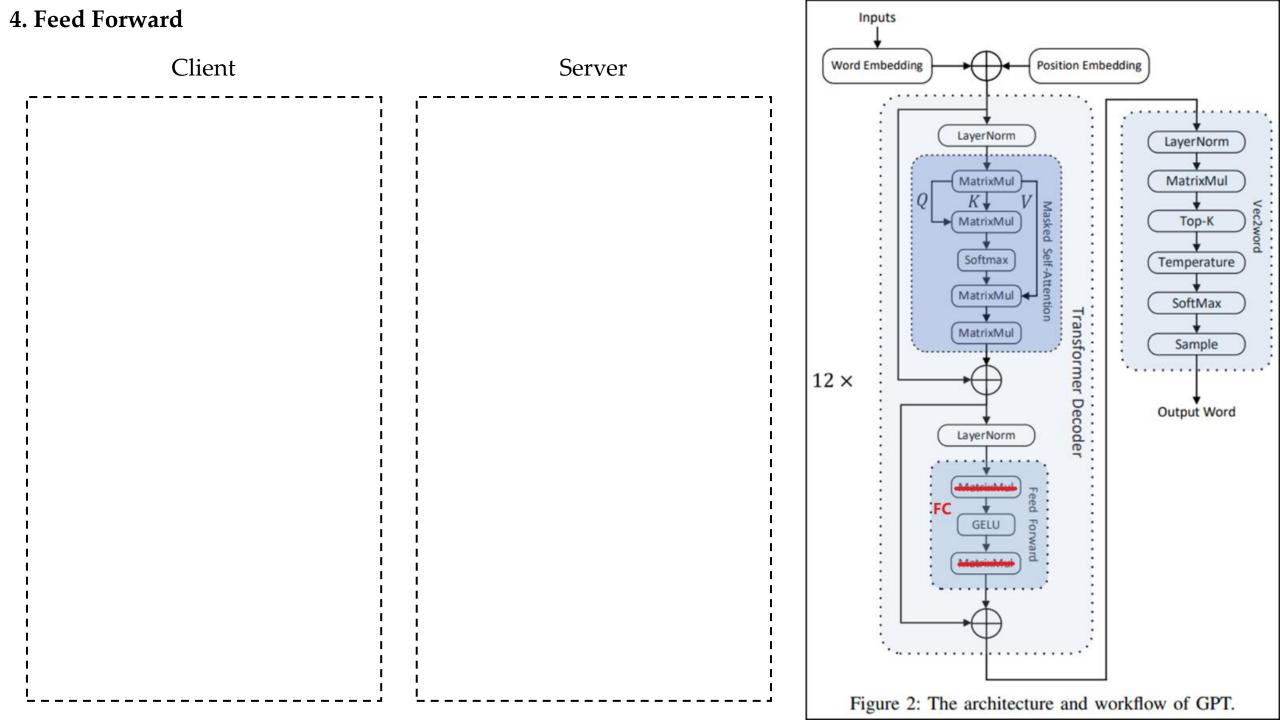


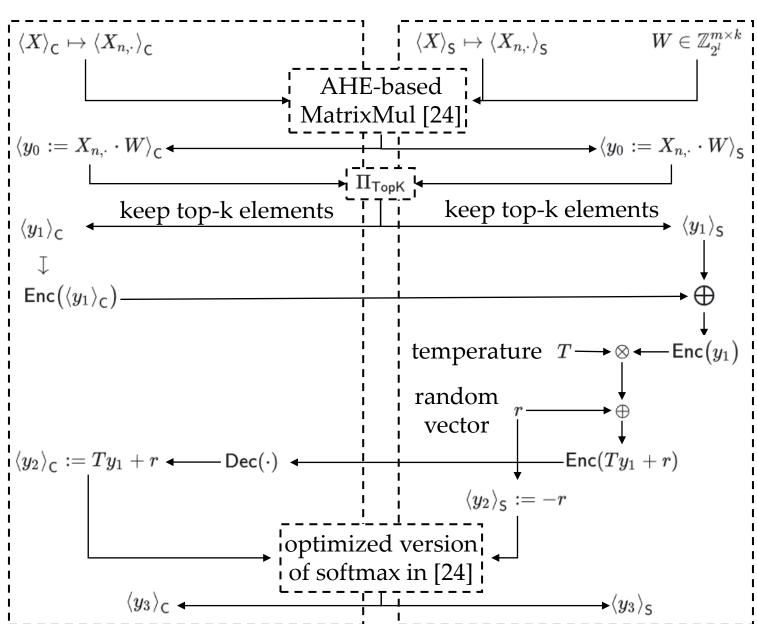
Figure 2: The architecture and workflow of GPT.



5. Vec2Word  $k := \text{vocab\_size}$ 

Client

Server



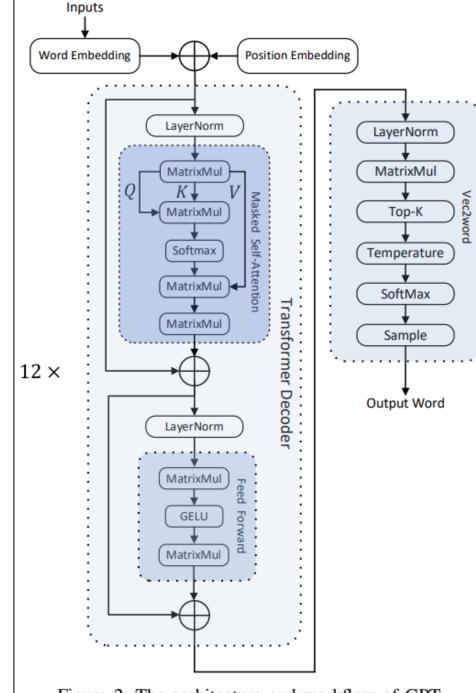


Figure 2: The architecture and workflow of GPT.