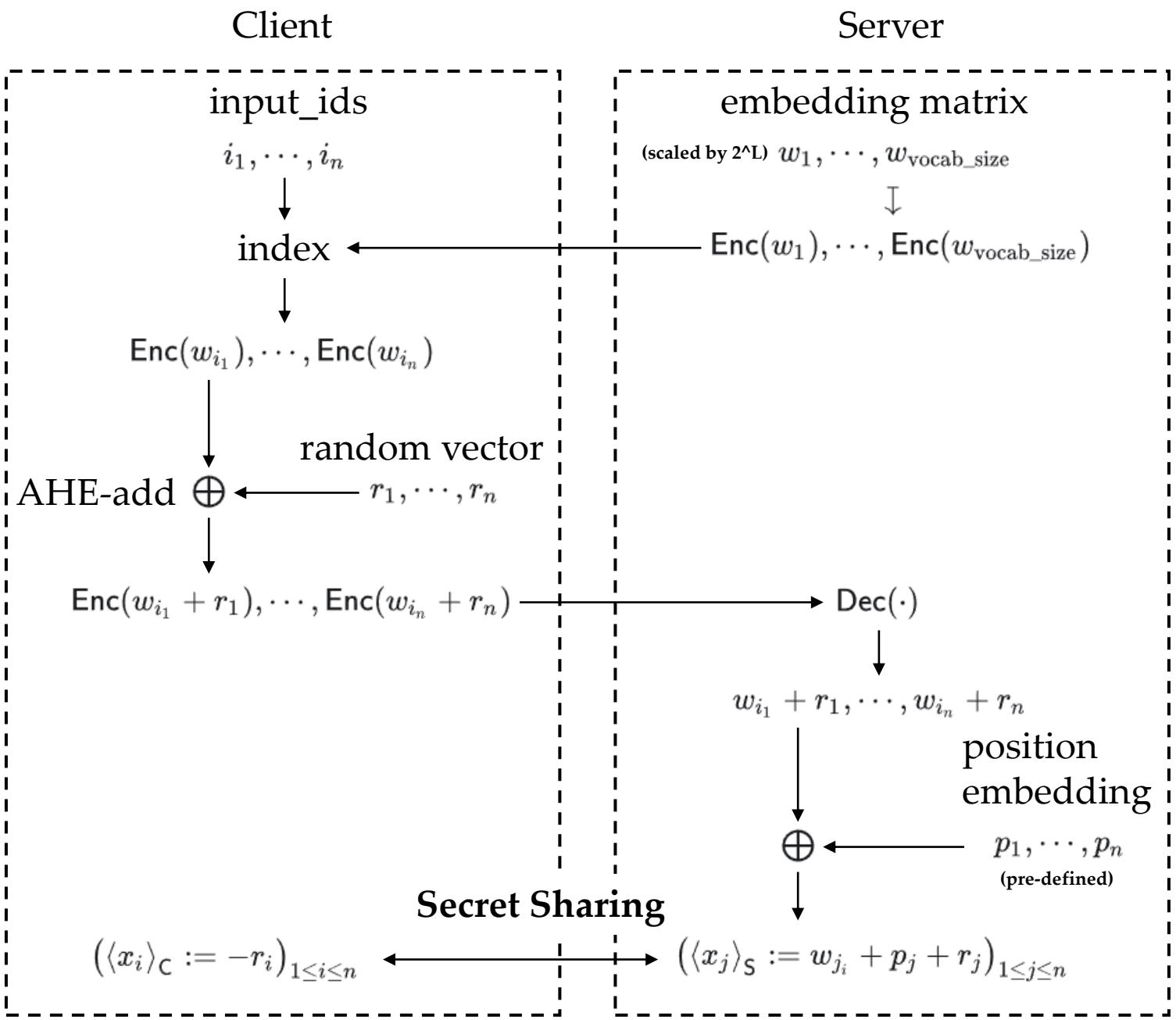
1. Embedding



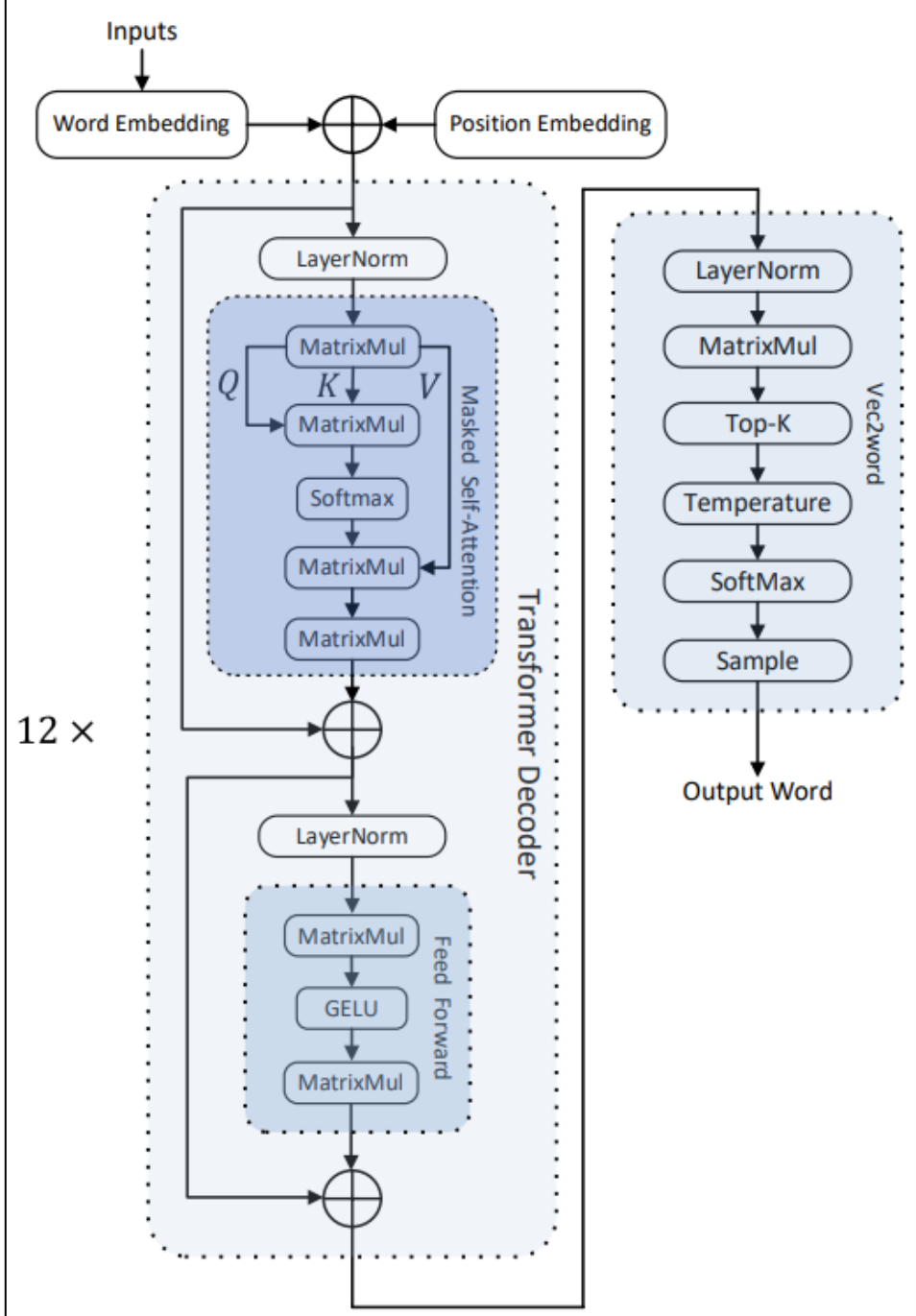
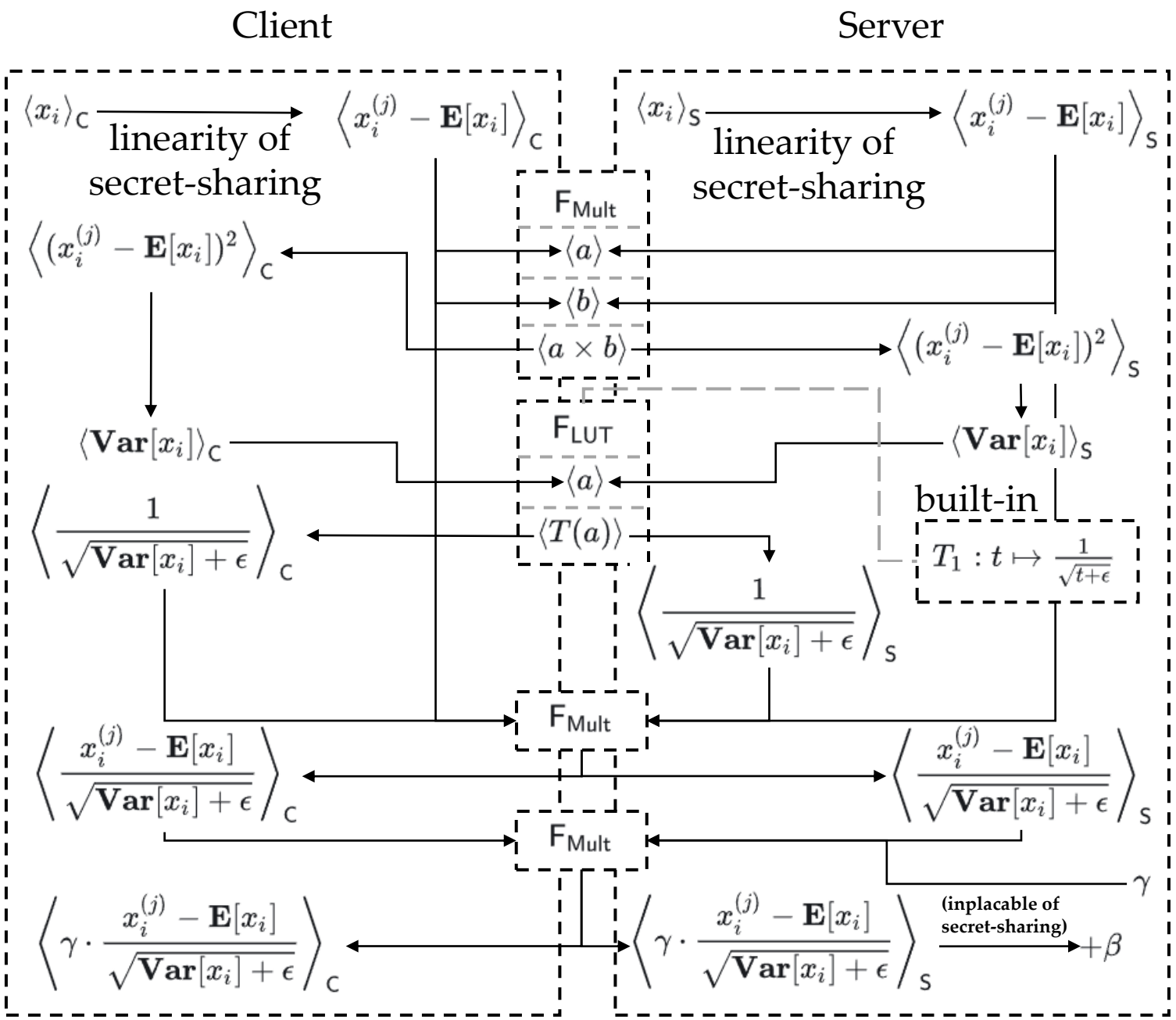


Figure 2: The architecture and workflow of GPT.

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



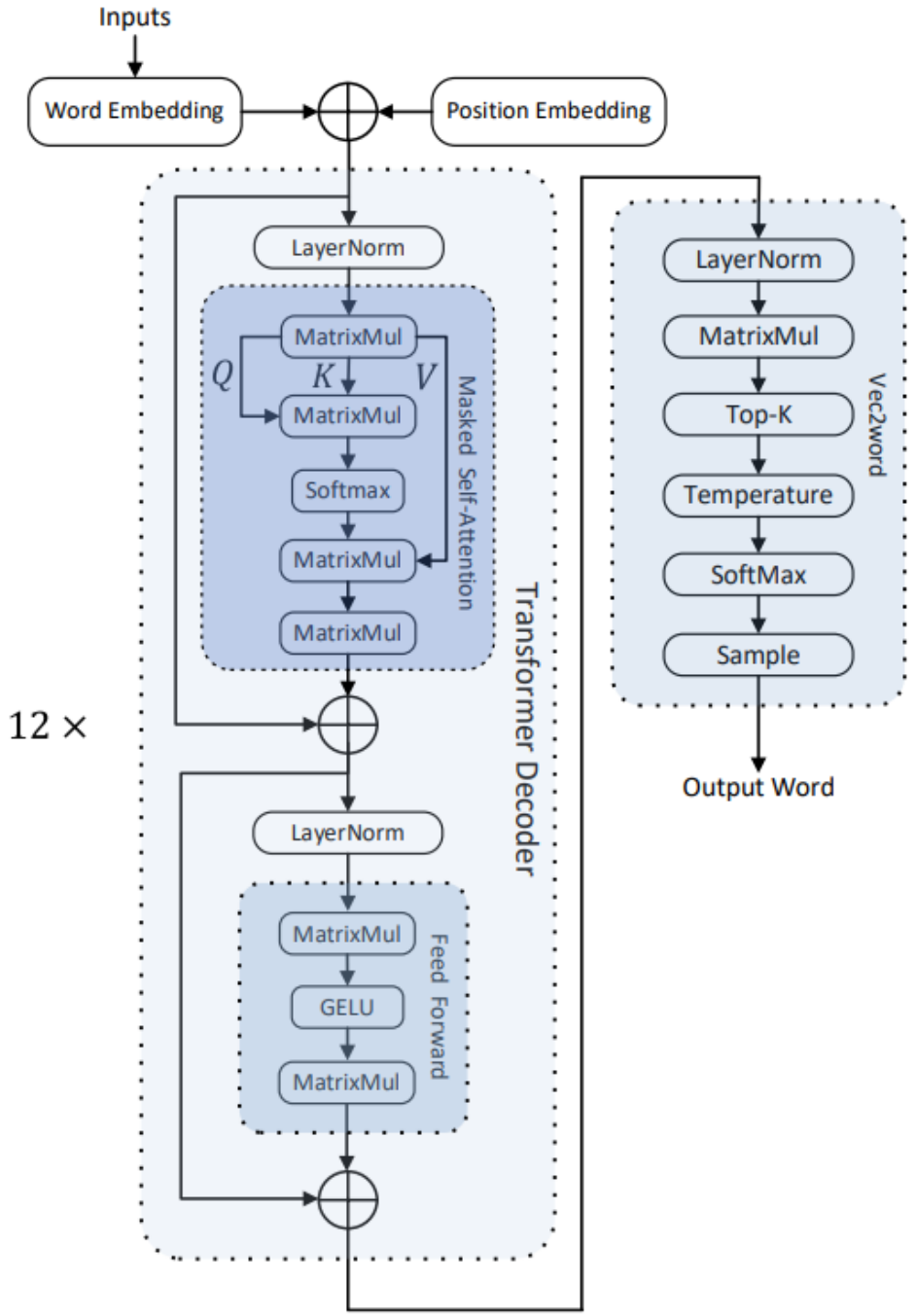
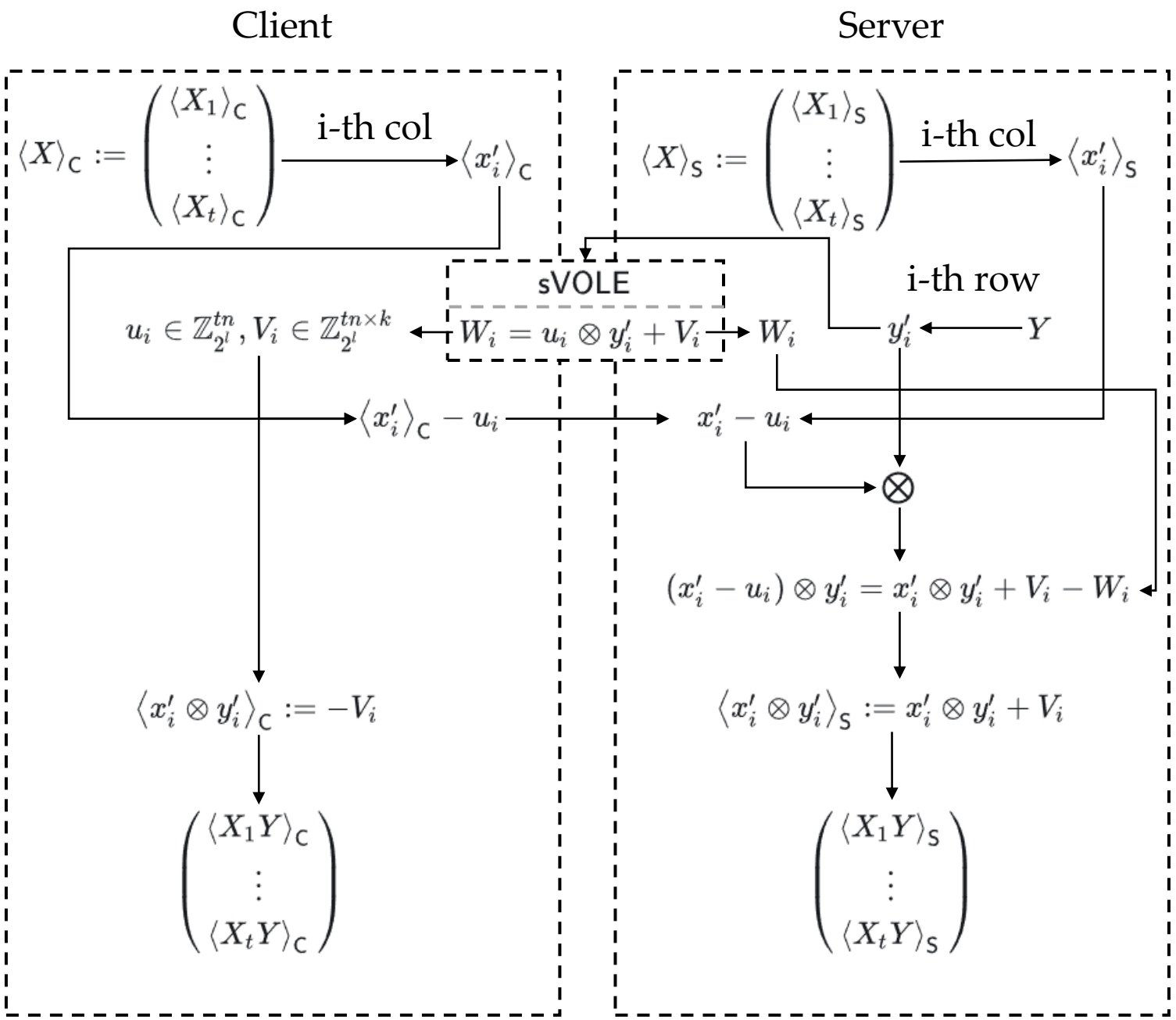


Figure 2: The architecture and workflow of GPT.

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

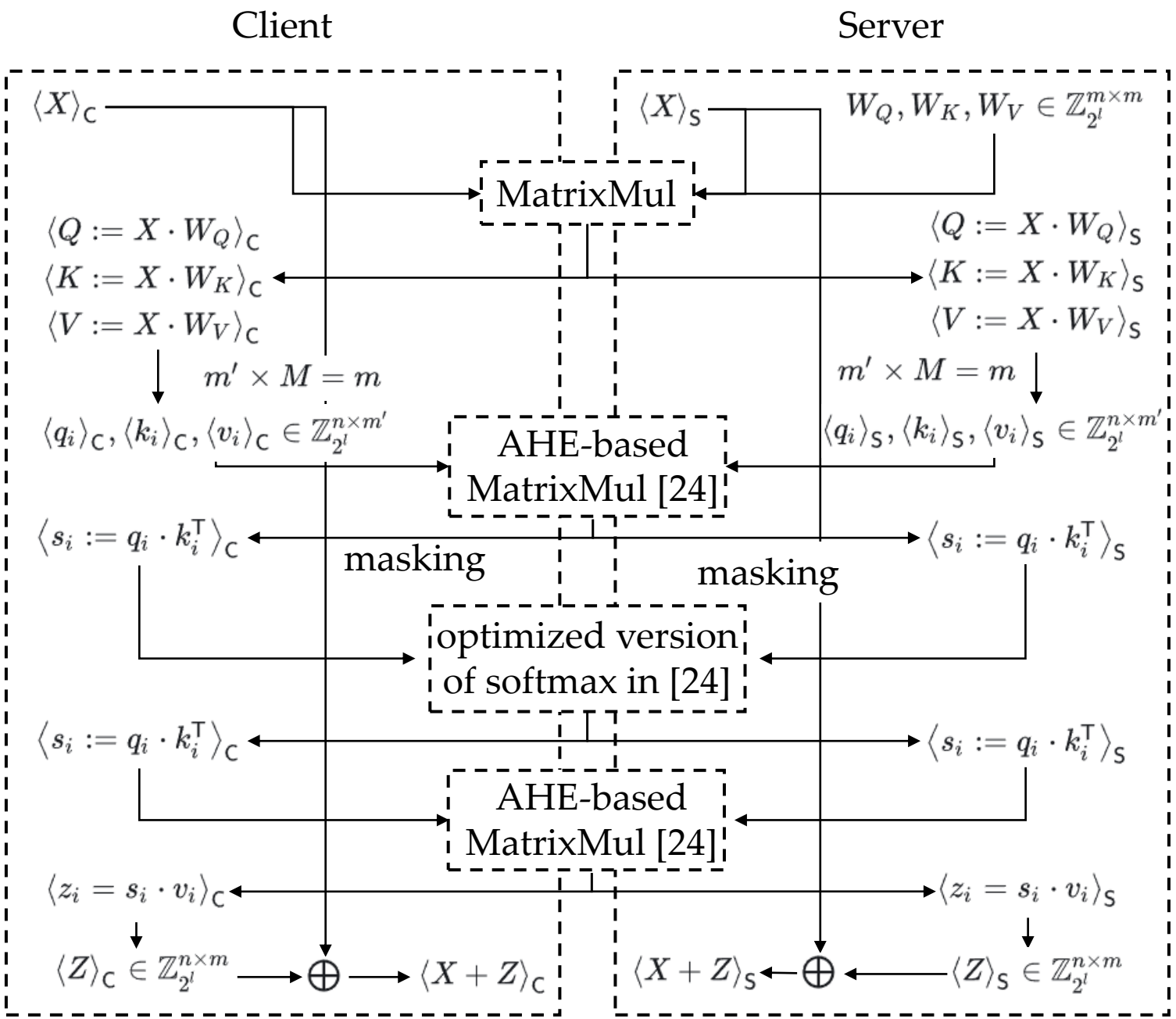


C has X_1, … , X_t and S has Y.
They want to get the secret-sharing of X_1 * Y, … , X_t * Y.

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

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

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



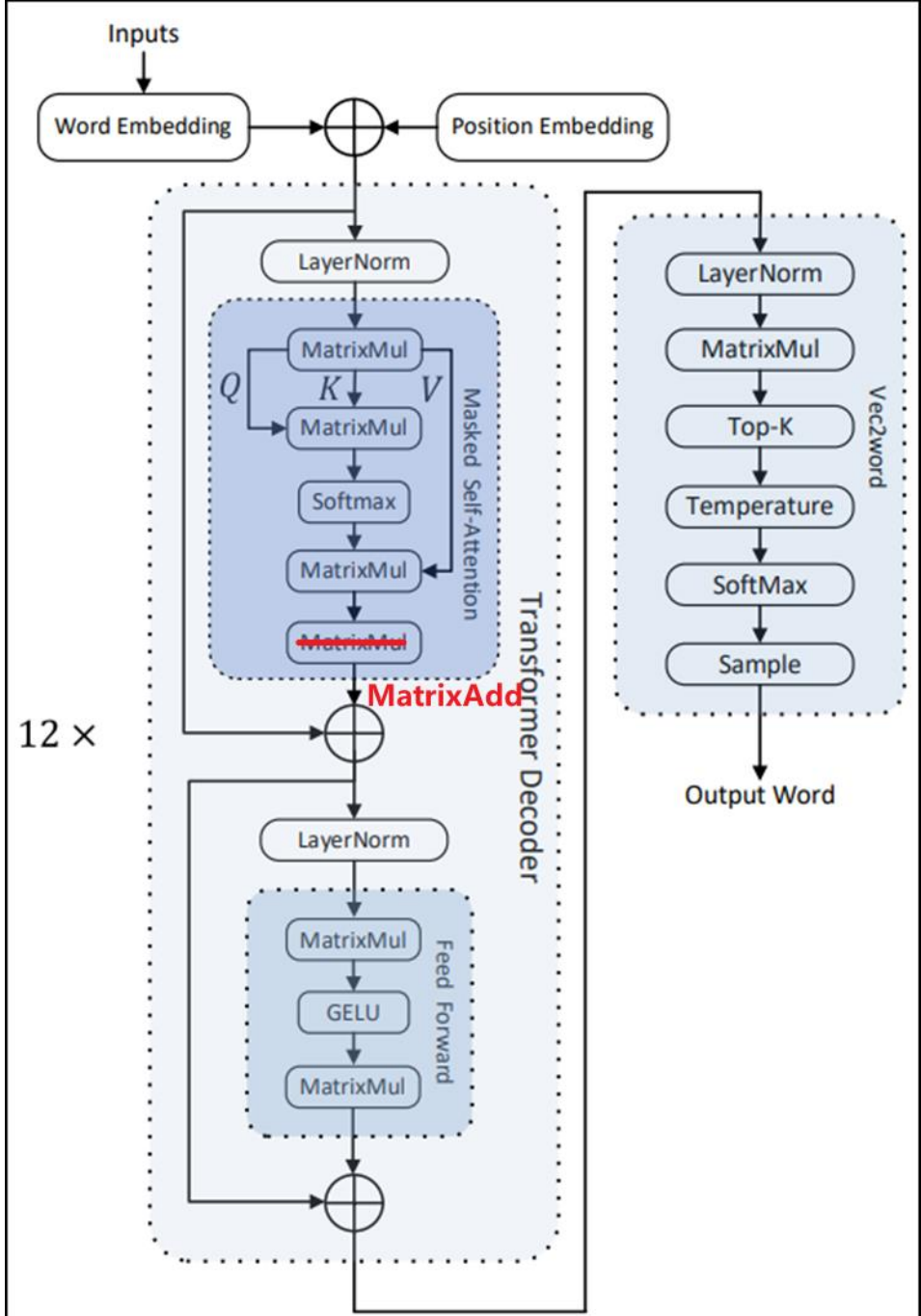


Figure 2: The architecture and workflow of GPT.

4. Feed Forward

Client

Server

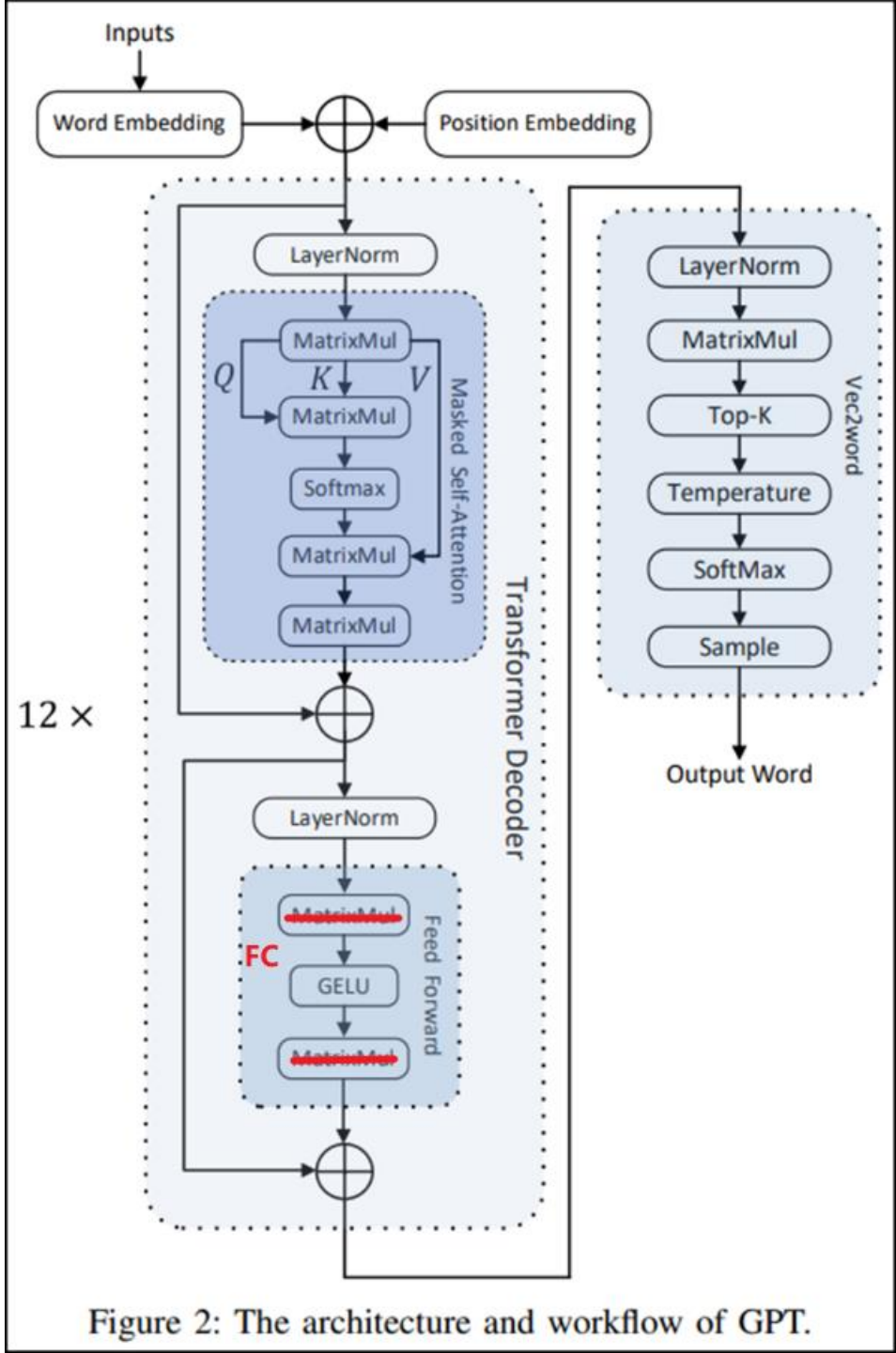
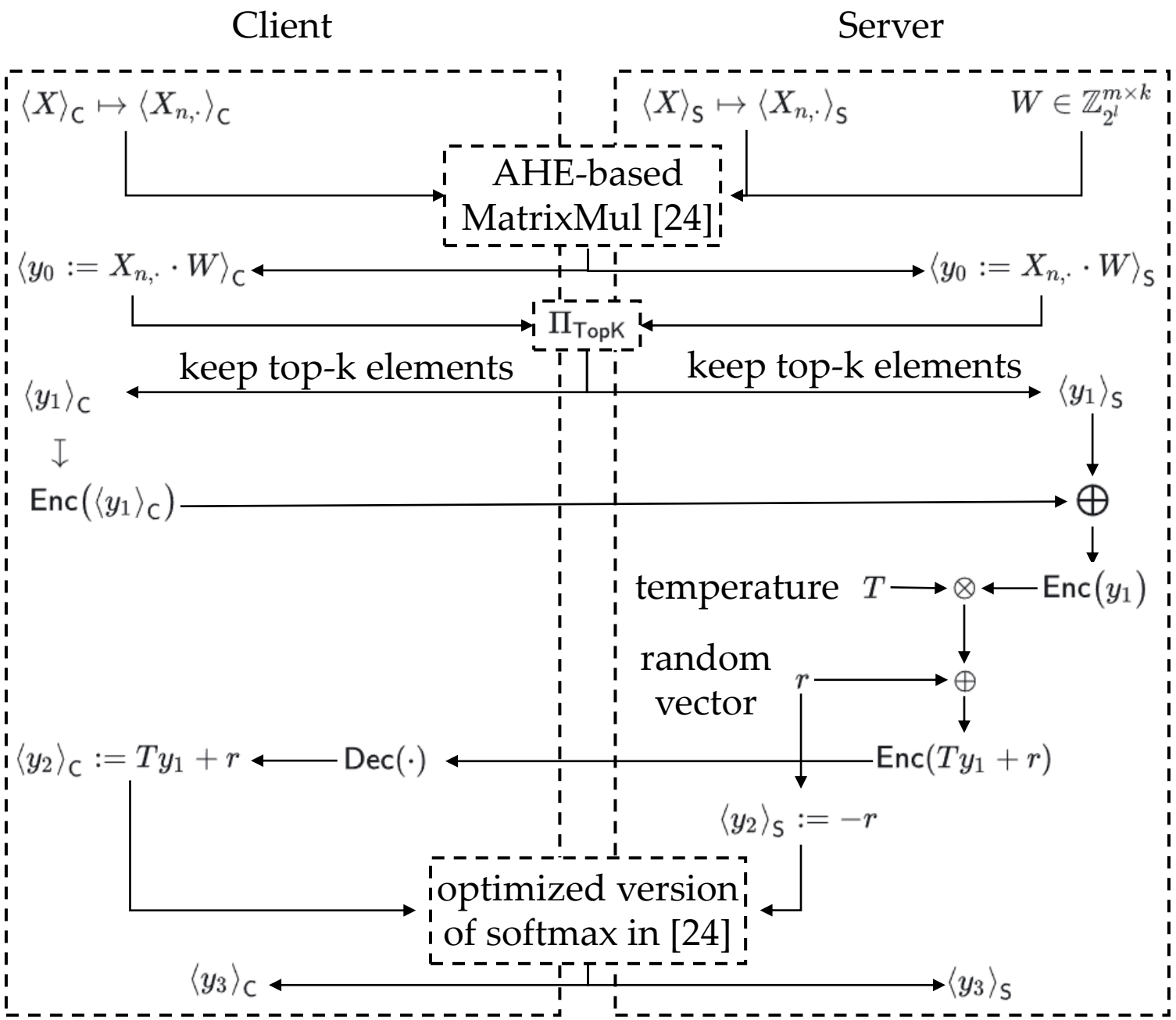


Figure 2: The architecture and workflow of GPT.

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



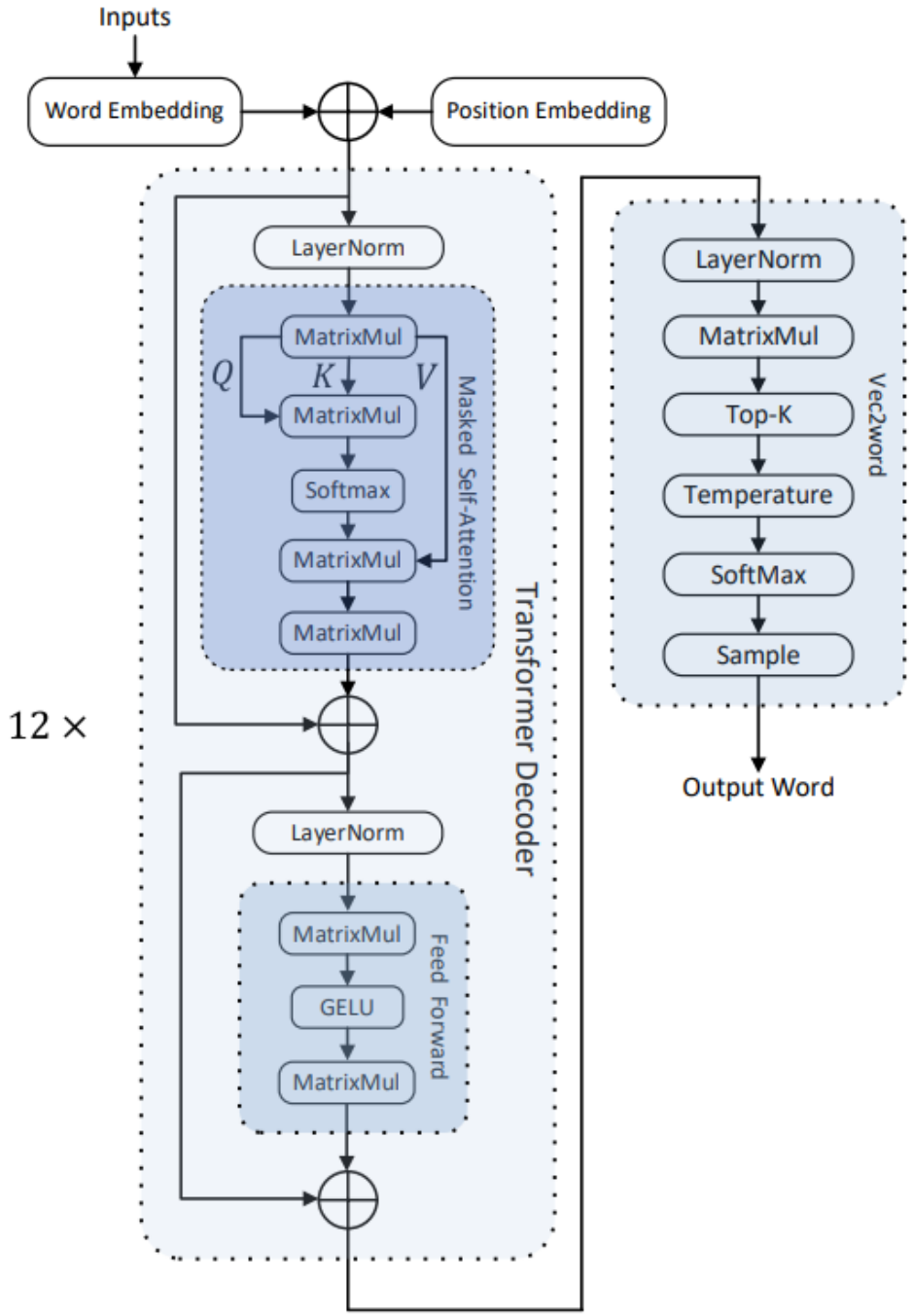


Figure 2: The architecture and workflow of GPT.