RSAConference™2024

San Francisco | May 6 – 9 | Moscone Center

THE ART OF POSSIBLE

SESSION ID: PDP-M06

IP Protection & Privacy in LLM: Leveraging Fully Homomorphic Encryption

#RSAC

Jordan Frery

Research Scientist
Zama
@JordanFrery

Benoit Chevallier-Mames

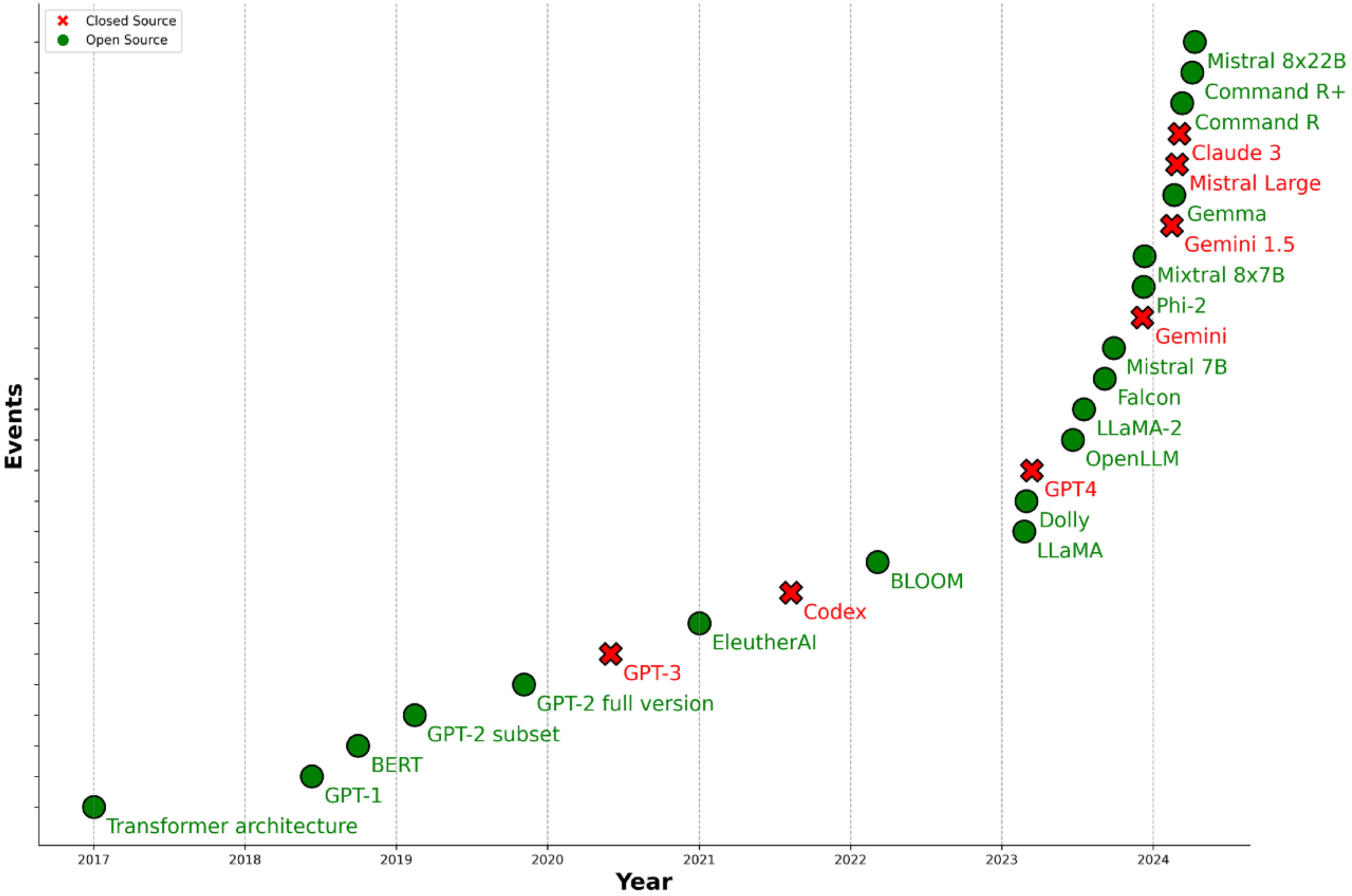
VP of Cloud and Machine Learning
Zama
@binoua_cml

Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference™ or any other co-sponsors. RSA Conference does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

The Rising Importance of Large Language Models (LLM)



ZAMA

Intellectual Property Privacy Concerns in LLM

- Accessing LLMs and Preserving IP
  - Serving LLM through a paid API (Claude, ChatGPT, Gemini, Copilot)
- This makes the **user's privacy at risk**

ZAMA

LLM SaaS Deployment: Convenience vs. Privacy

- Accessing LLM while preserving user's privacy:
  - 1. Build your own LLMs (too expensive)
  - 2. Building upon free for commercial use LLMs (not state of the art)
  - 3. Deploy proprietary LLM on-premise (leak of IP)

Rank	🤖 Model	License
1	GPT-4-1106-preview	Proprietary
2	GPT-4-0125-preview	Proprietary
3	Bard (Gemini Pro)	Proprietary
4	GPT-4-0314	Proprietary
5	GPT-4-0613	Proprietary
6	Mistral-Large-2402	Proprietary
7	Claude-1	Proprietary
8	Mistral Medium	Proprietary
9	Qwen1.5-72B-Chat	Qianwen LICENSE
10	Claude-2.0	Proprietary
11	Mistral-Next	Proprietary
12	Gemini Pro (Dev API)	Proprietary
13	Claude-2.1	Proprietary
14	Mixtral-8x7b-Instruct-v0.1	Apache 2.0
15	GPT-3.5-Turbo-0613	Proprietary

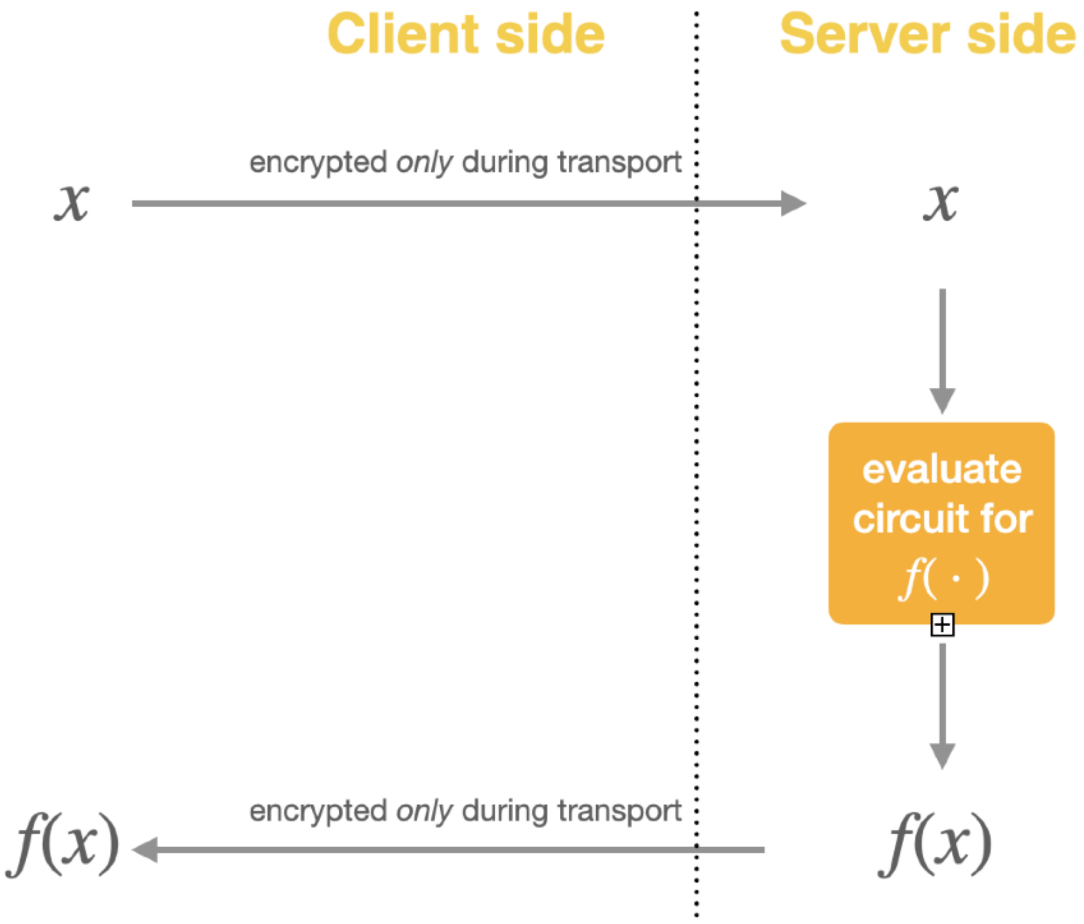
Our Goal

Use LLMs while preserving IP and user's privacy.

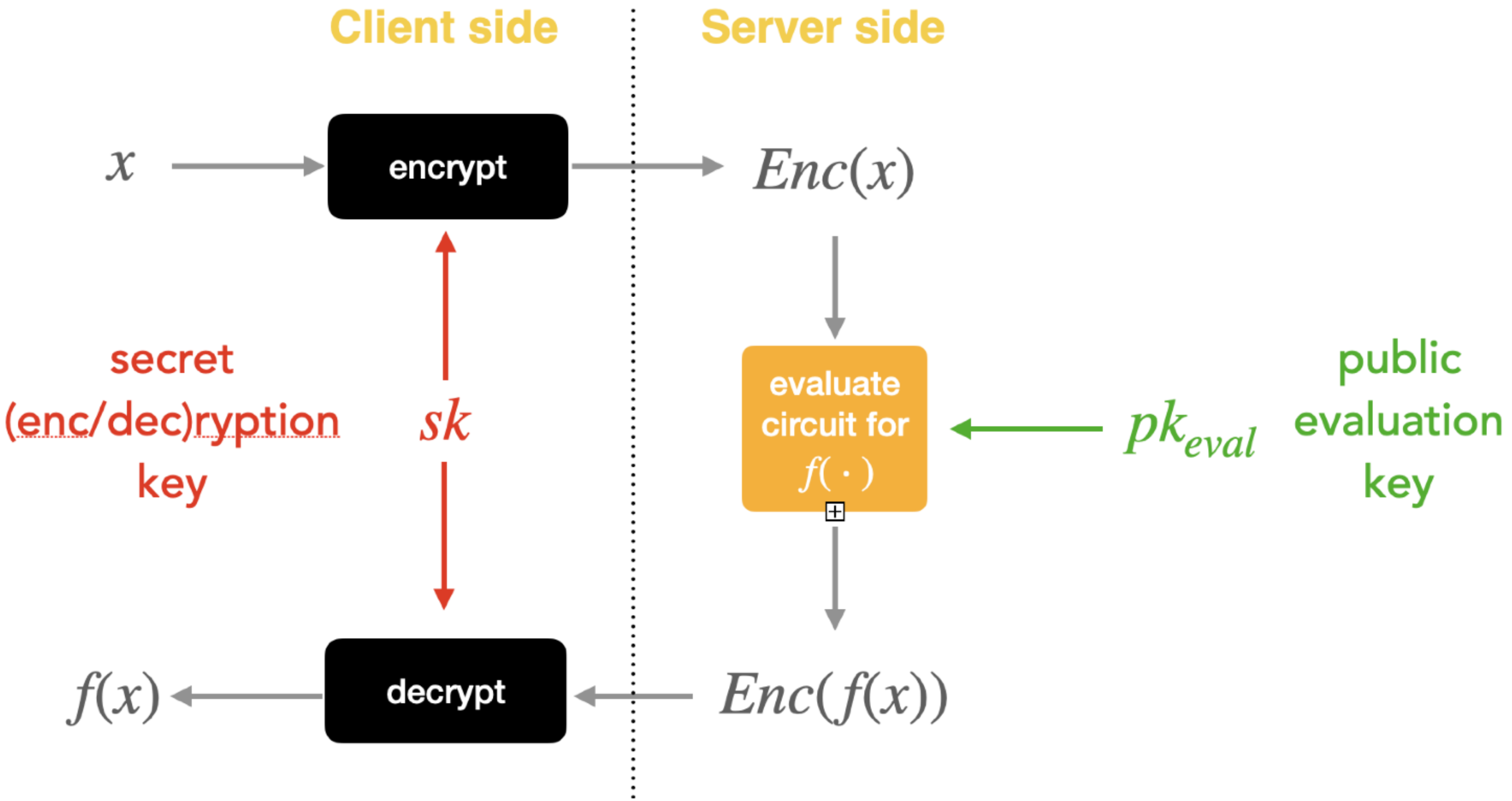
RSAConference™2024

Fully Homomorphic Encryption

Today Data is Only Encrypted During Transport



With FHE, Data is Encrypted Also While Processed



ZAMA

RSAConference™2024

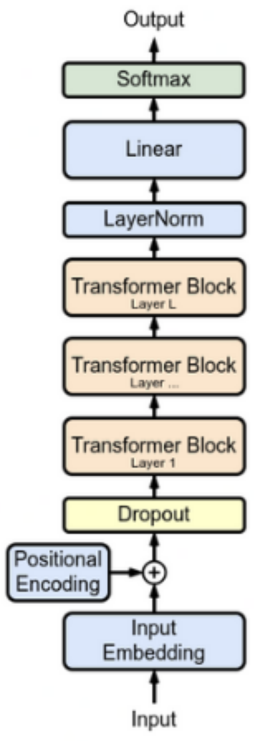
Leveraging FHE in LLMs

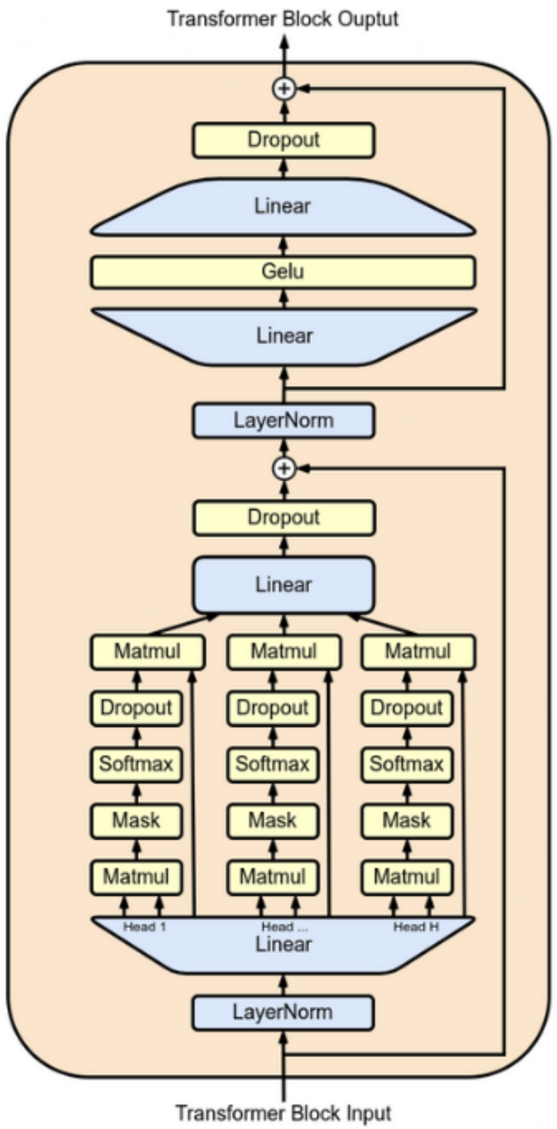
Our Open-Source Experiments

huggingface.co/blog/encrypted-llm

Transformers as Large Language Models

- Main operations:
  - Embedding
  - Feed-Forward
  - Attention

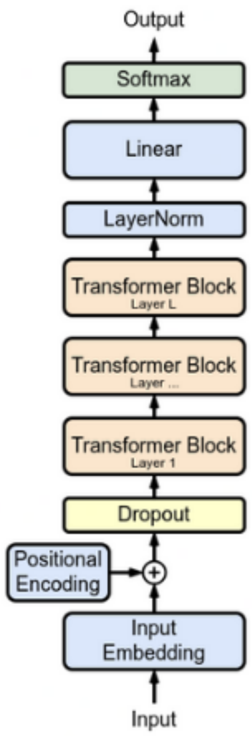


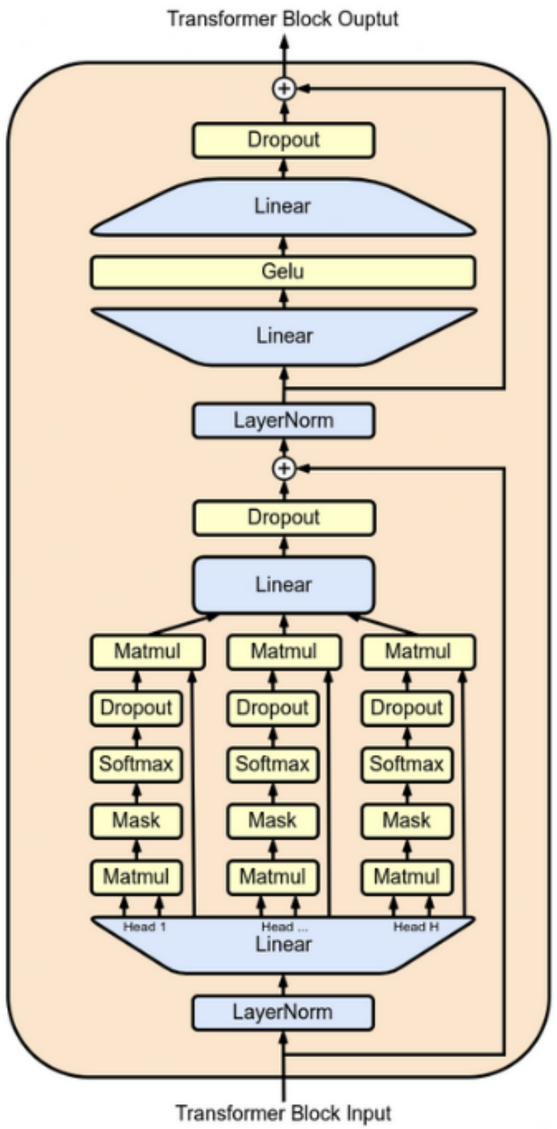


Transformers as Large Language Models

- Main operations:
  - **Embedding**
  - Feed-Forward
  - Attention

$$\mathrm{Embedding}_i = \mathrm{E}[i]$$



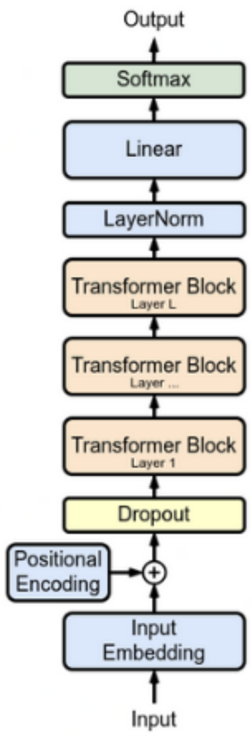


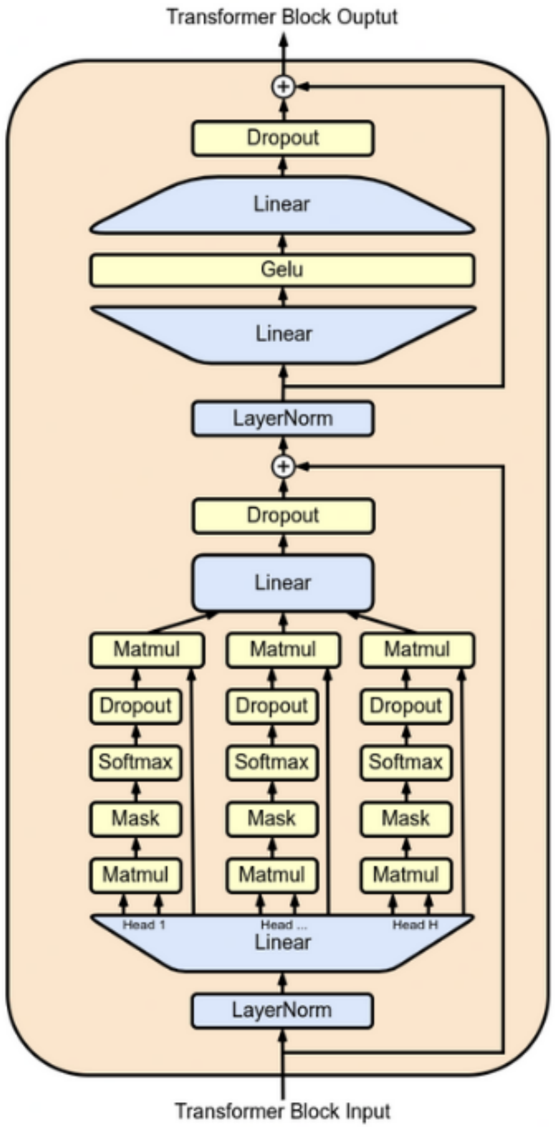
ZAMA

Transformers as Large Language Models

- Main operations:
  - **Embedding** ✓ FHE (ms)
  - Feed-Forward
  - Attention

$$\text{Embedding}_i = \text{E}[i]$$

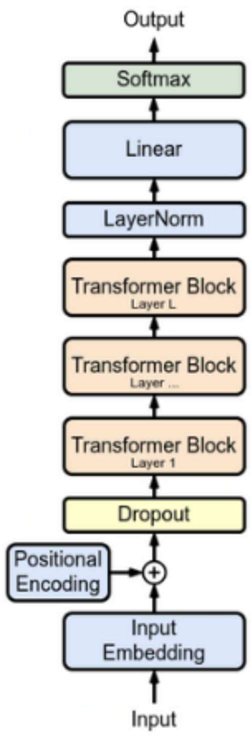


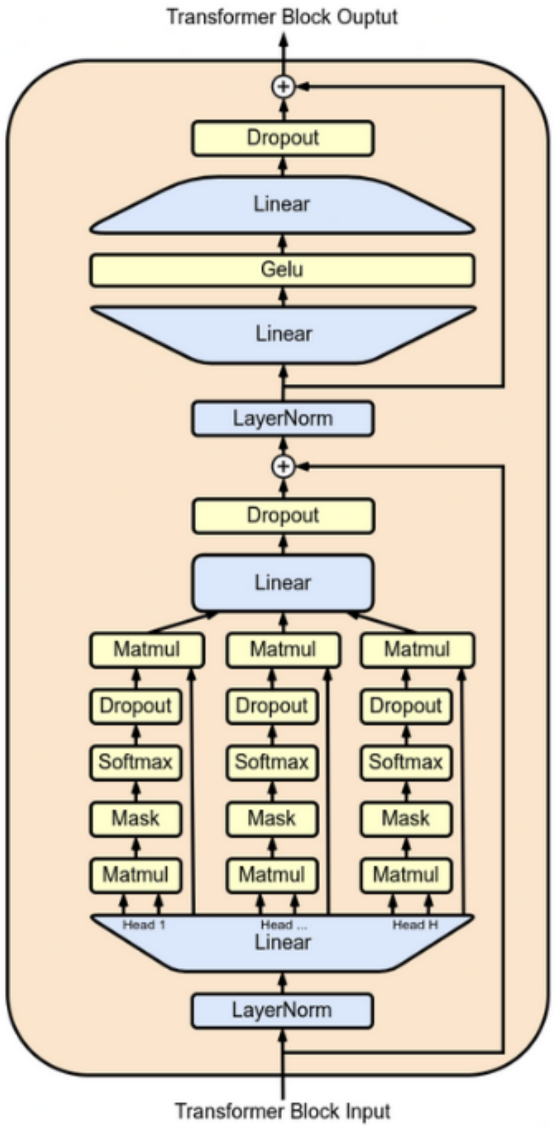


Transformers as Large Language Models

- Main operations:
  - Embedding ✓ FHE (ms)
  - **Feed-Forward**
  - Attention

$$\mathrm{FFN}(x) = F(xW_1 + b_1)W_2 + b_2$$



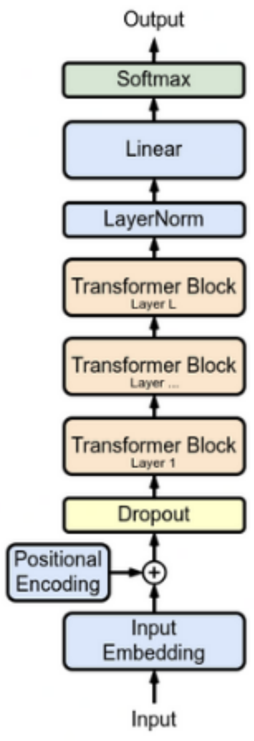


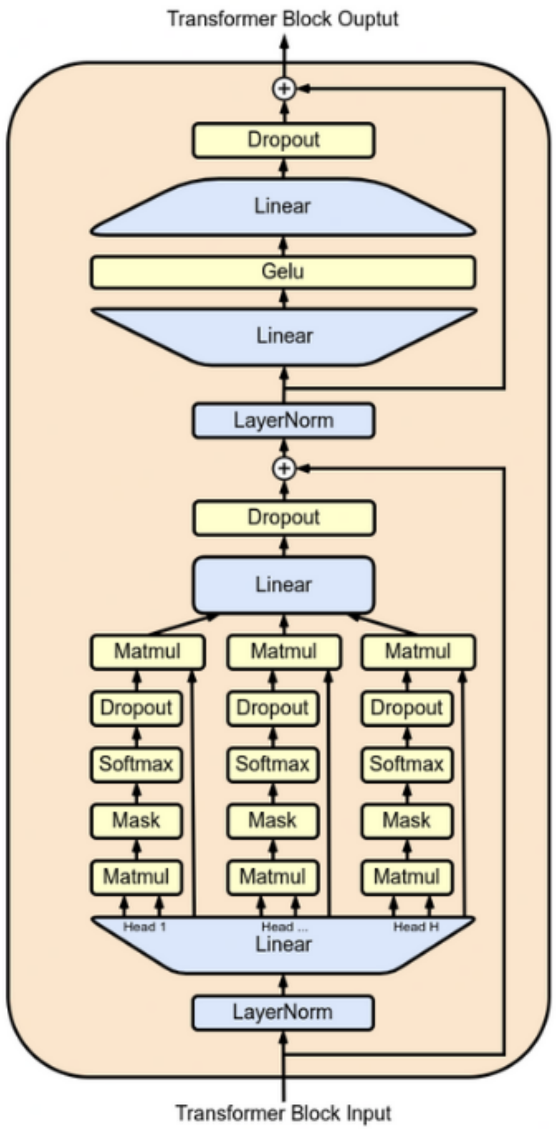
ZAMA

Transformers as Large Language Models

- Main operations:
  - Embedding ✓ FHE (ms)
  - **Feed-Forward** ✓ FHE (s)
  - Attention

$$\mathrm{FFN}(x) = F(xW_1 + b_1)W_2 + b_2$$

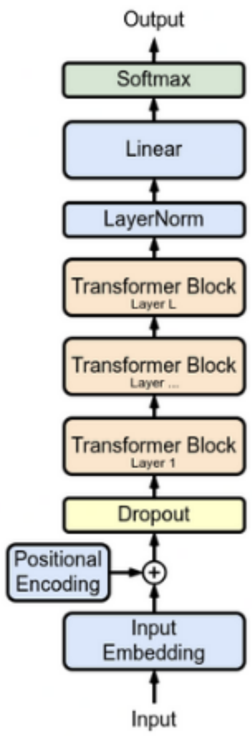


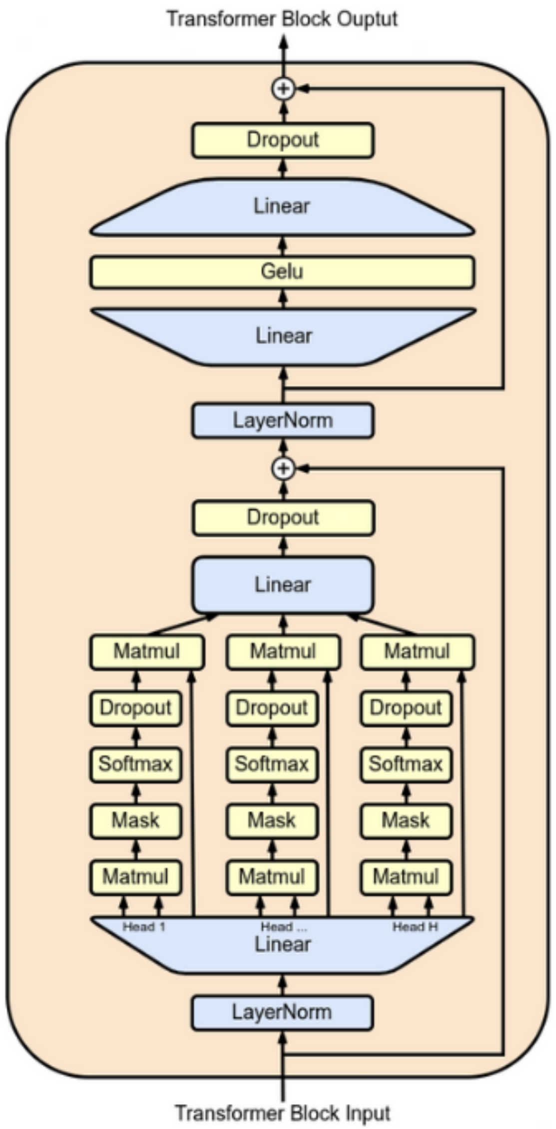


Transformers as Large Language Models

- Main operations:
  - Embedding ✓ FHE (ms)
  - Feed-Forward ✓ FHE (s)
  - **Attention**

$$Q = XW_Q$$
$$K = XW_K$$
$$V = XW_V$$
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$





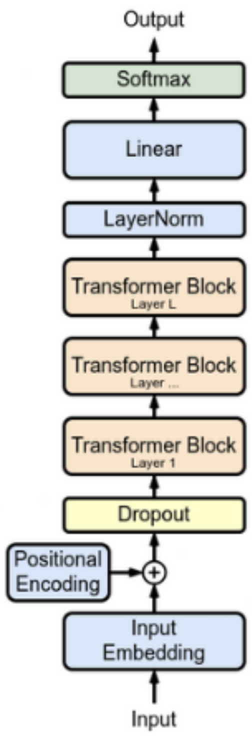
ZAMA

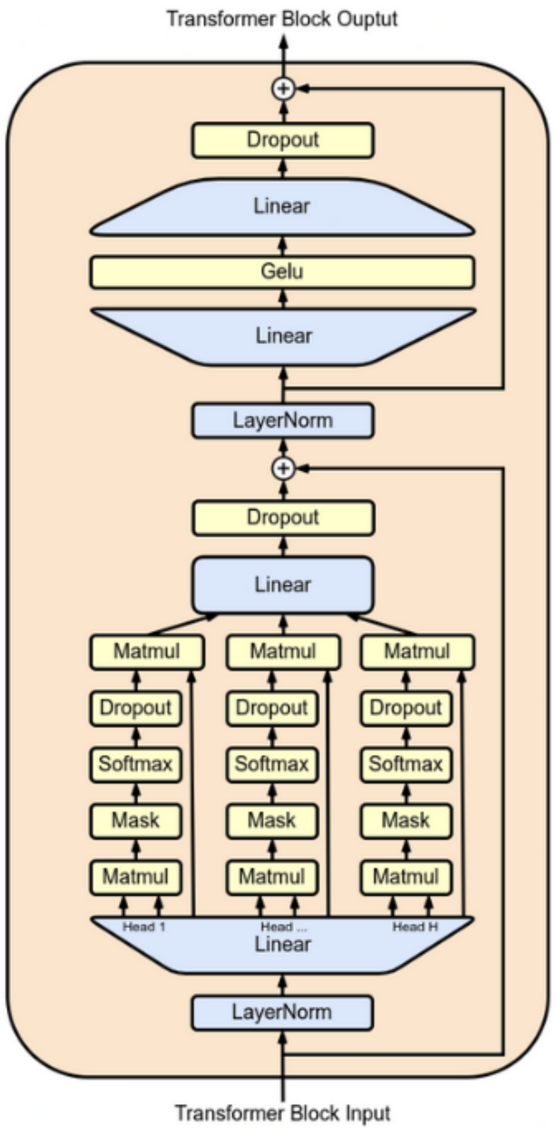
Transformers as Large Language Models

- Main operations:
  - Embedding ✓ FHE (ms)
  - Feed-Forward ✓ FHE (s)
  - **<u>Attention</u>** ✓ FHE (m/h)

$$Q = XW_Q$$
$$K = XW_K$$
$$V = XW_V$$

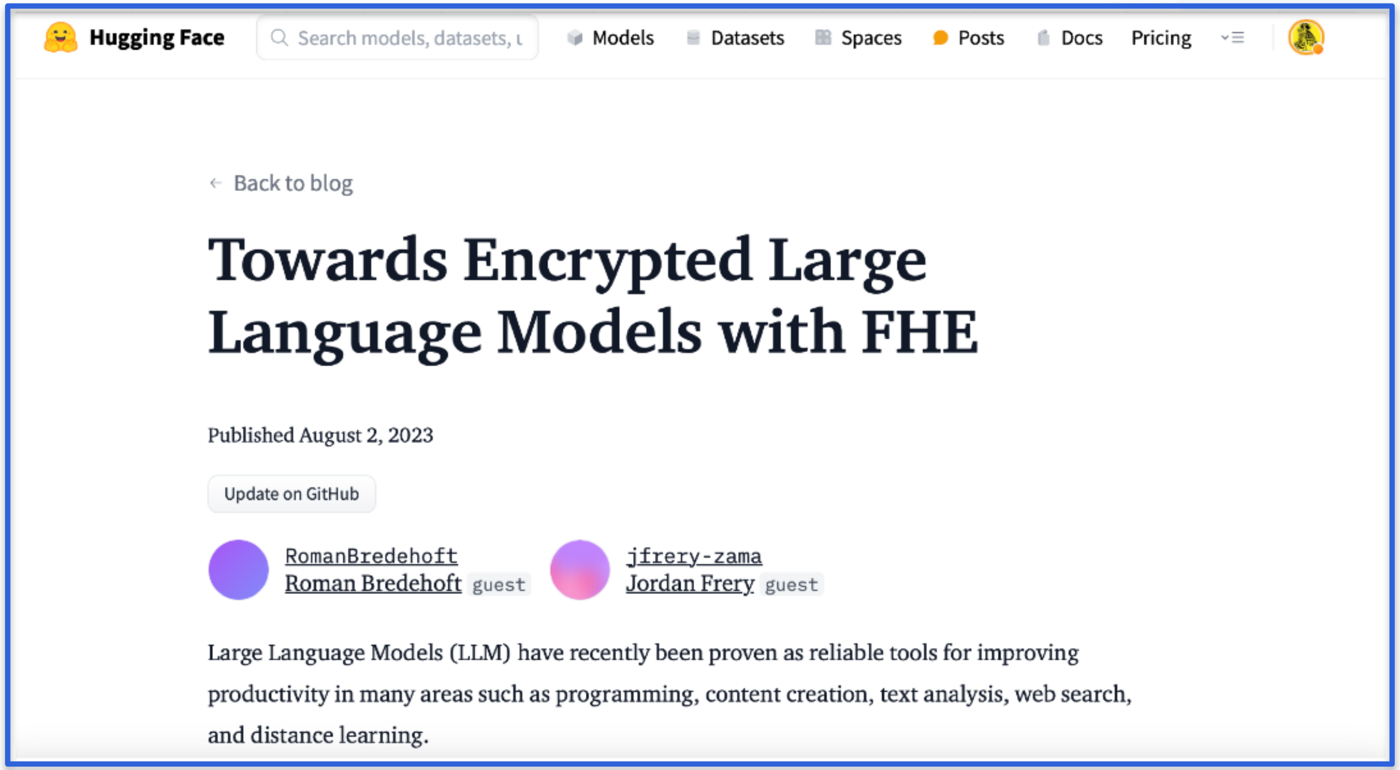
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$





Read the Full Blog Post and Open-Source Code

huggingface.co/blog/encrypted-llm



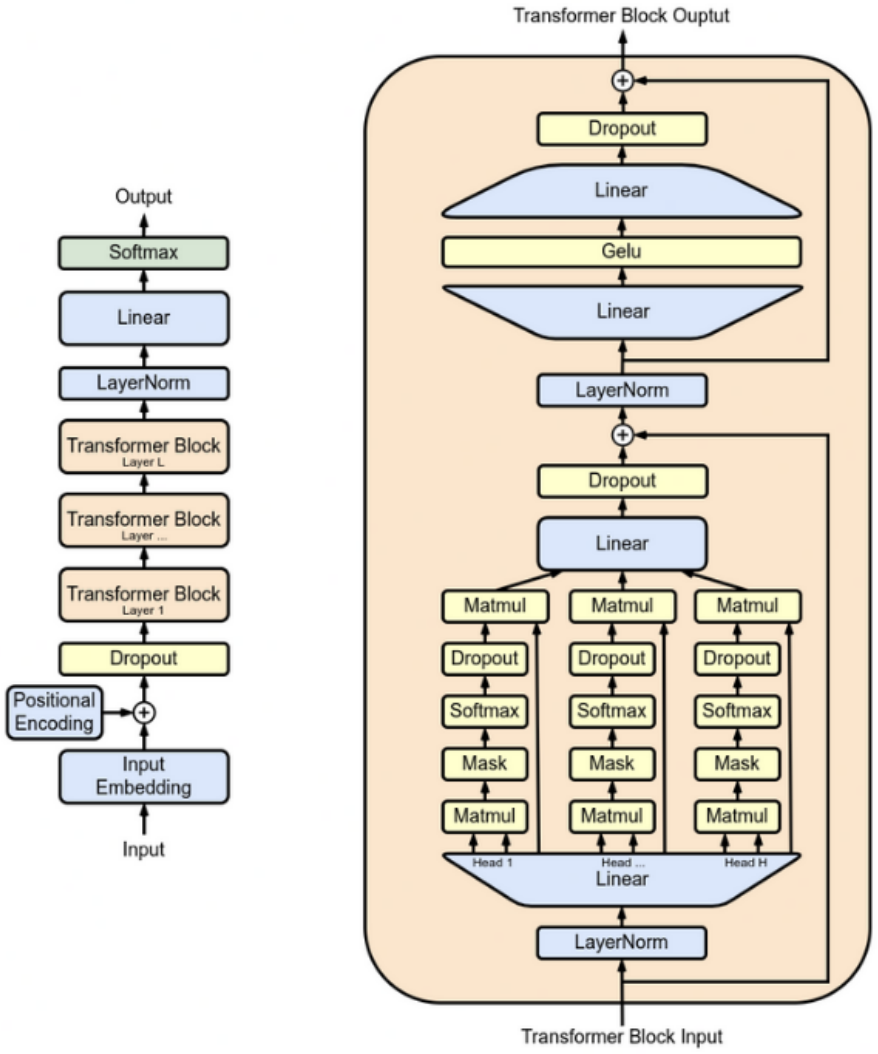
ZAMA

Hybrid Models

Faster Implementations

LLM User's Privacy And IP preserving

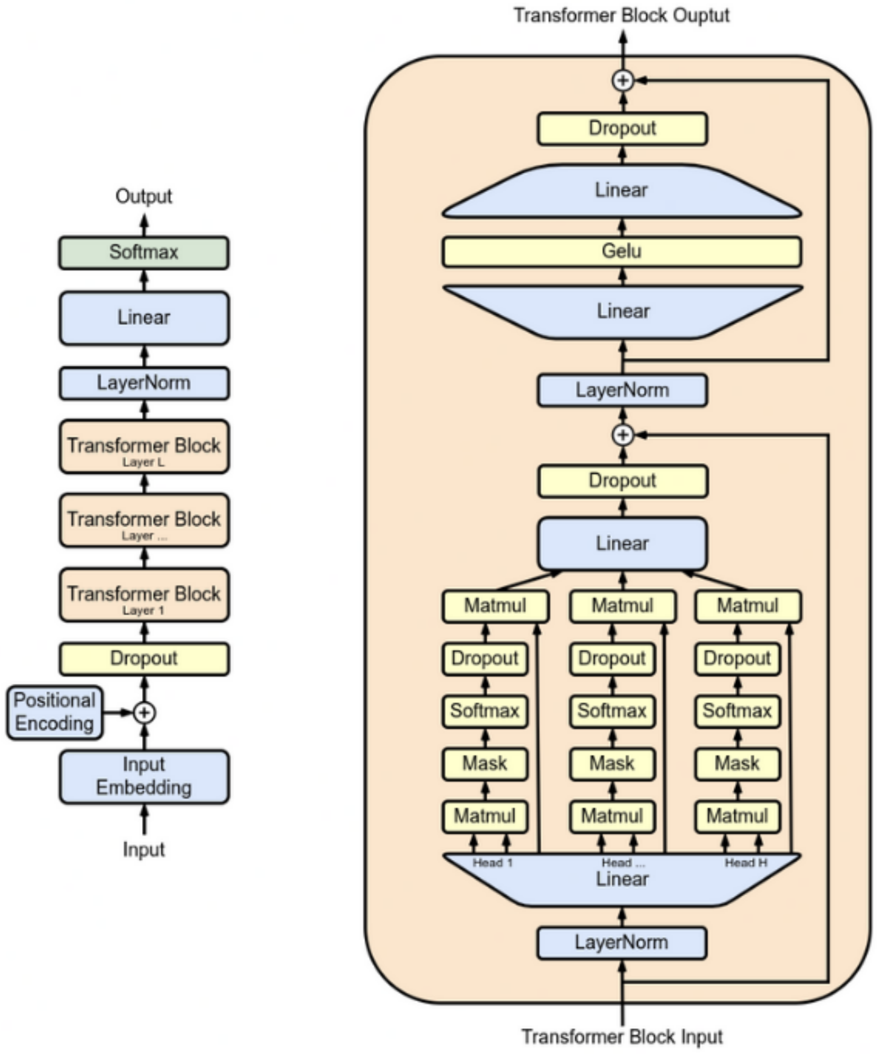
- Where is the Intellectual Property (IP) located?
  - Neural network parameters (weights/biases)
- Parameter distribution in google/gemma-7b:
  - Embedding -> 9.21%
  - Feed-Forward -> 71.55%
  - Attention -> 15.93%
  - Remaining -> 3.31%



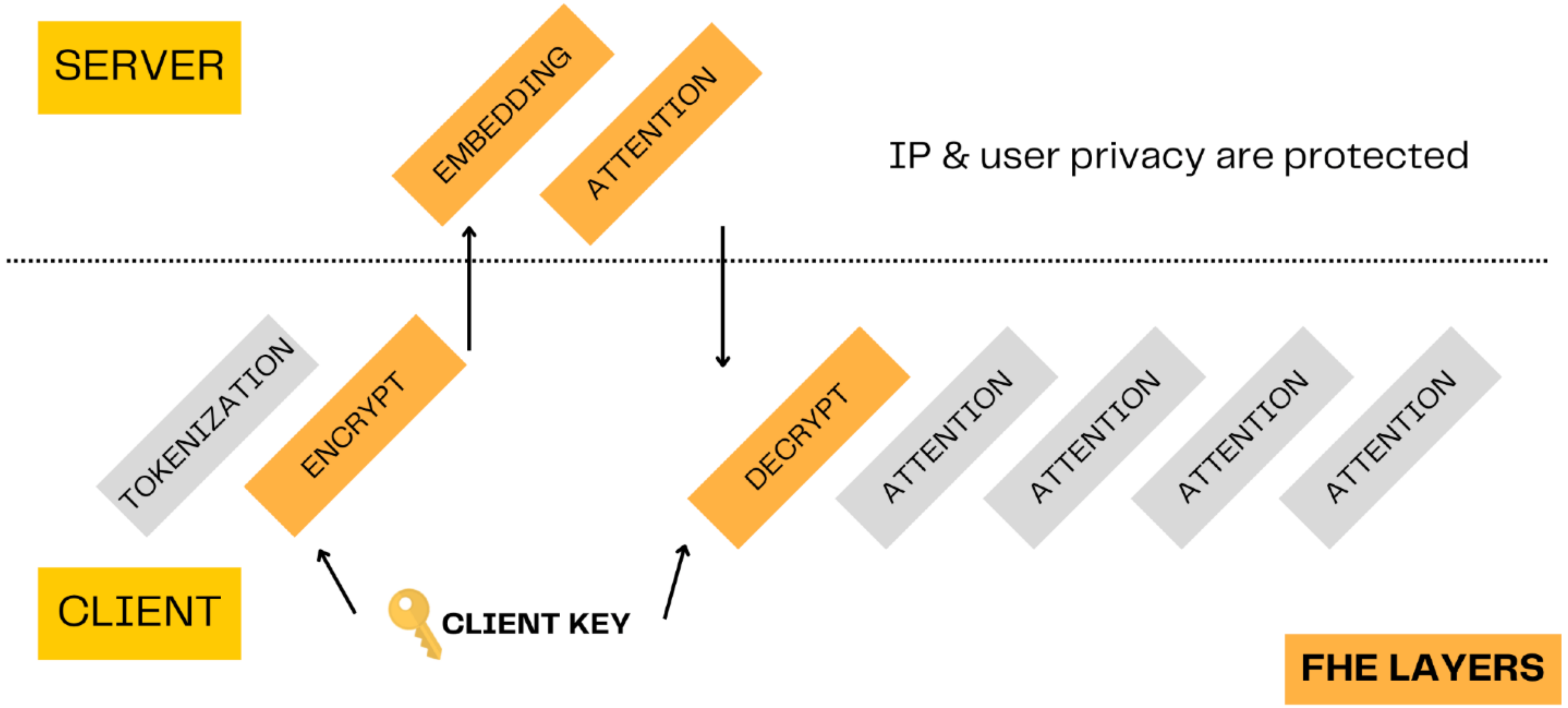
ZAMA

LLM User's Privacy And IP preserving

- Where is the Intellectual Property (IP) located?
  - Neural network parameters (weights/biases)
- Parameter distribution in google/ gemma-7b:
  - Embedding -> 9.21%
  - Feed-Forward -> 71.55%
  - Attention -> 15.93%
  - Remaining -> 3.31%



Hybrid Model Principle



RSAConference™2024

Demo

"Apply" Slide

- Next week you should:
  - Identify AI systems in your company who need privacy:
    - for your customers (their personal data)
    - for your company (your assets)
- In the first three months following this presentation you should:
  - Prototype protection one of these AI systems with FHE
  - Get help from discord.fhe.org
- Within six months you should:
  - Have moved your most critical AI systems to FHE (or other PETs)

RSAConference™2024

San Francisco | May 6 – 9 | Moscone Center

THE ART OF POSSIBLE

SESSION ID: PDP-M06

IP Protection & Privacy in LLM: Leveraging Fully Homomorphic Encryption

#RSAC

Jordan Frery

Research Scientist
Zama
@JordanFrery

Benoit Chevallier-Mames

VP of Cloud and Machine Learning
Zama
@binoua_cml

Appendix: Glossary

In the document:

- **FHE**: fully homomorphic encryption
- **ms**: milli seconds
- **s**: seconds
- **m/h**: minutes or hours

For LLM technical slides (Slides 11-17):

- Embedding, Feed-Forward, Attention and Softmax are described in Wikipedia or LLM technical papers
- F, W1, W2, Q, K, V, W_Q, W_K, W_V and X are matrices
- x, b_1, b_2 are vectors
- M^T stands for transposition of a matrix M
- d_k is the dimension of Q and K