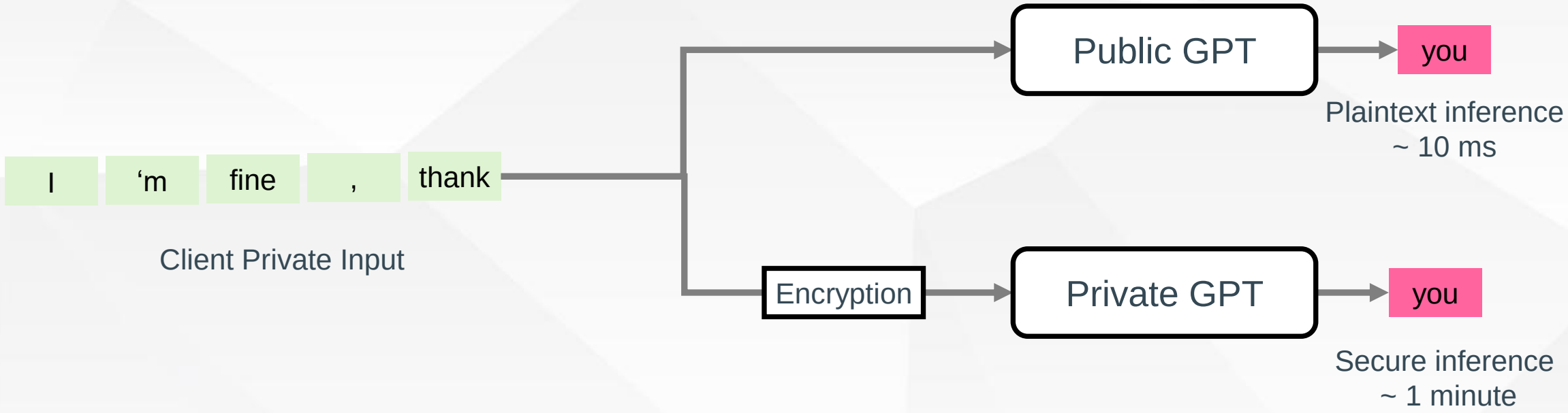
An Efficient Private GPT Never Autoregressively Decodes

Zhengyi Li, Yue Guan, Kang Yang, Yu Feng, Ning Liu, Yu Yu, Jingwen Leng, Minyi Guo

hobbit@sjtu.edu.cn

2025.05

Public GPT shares knowledge with the private GPT



- Existing secure decoding of GPTs puts all computation in secure world.
  - Guarantees the client only learns the final response while the server learns nothing.
- However, public GPTs share knowledge with private GPTs, and not all information embedded in the private model is prohibitive.

How can we integrate the public GPT into the secure decoding without affecting privacy and decoding quality?

Latency insensitivity to the input length

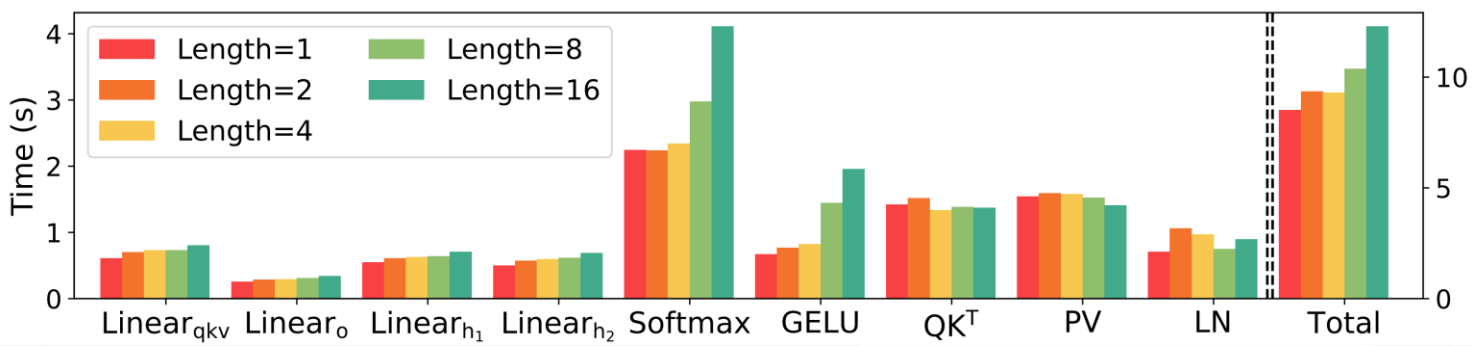


Figure 1: The latency of securely executed layers against varying input lengths.

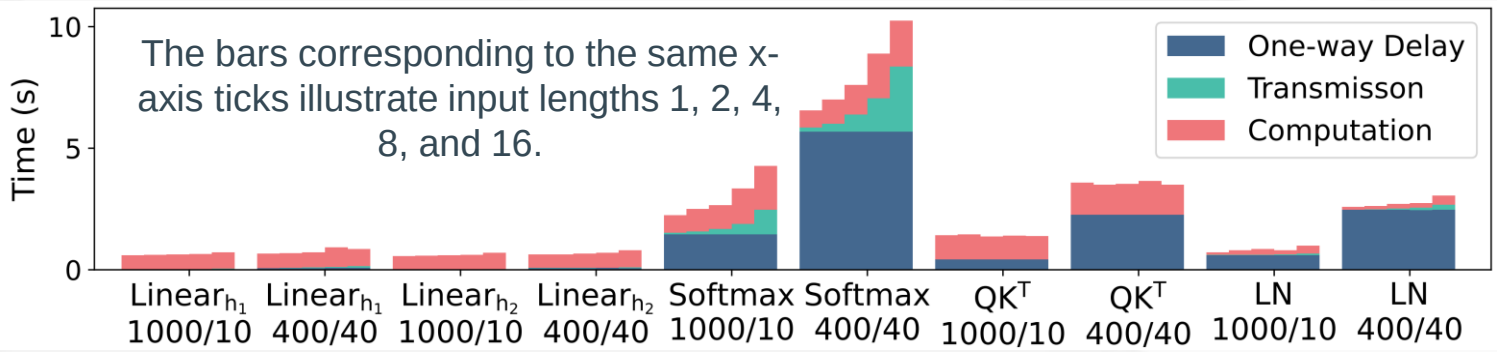


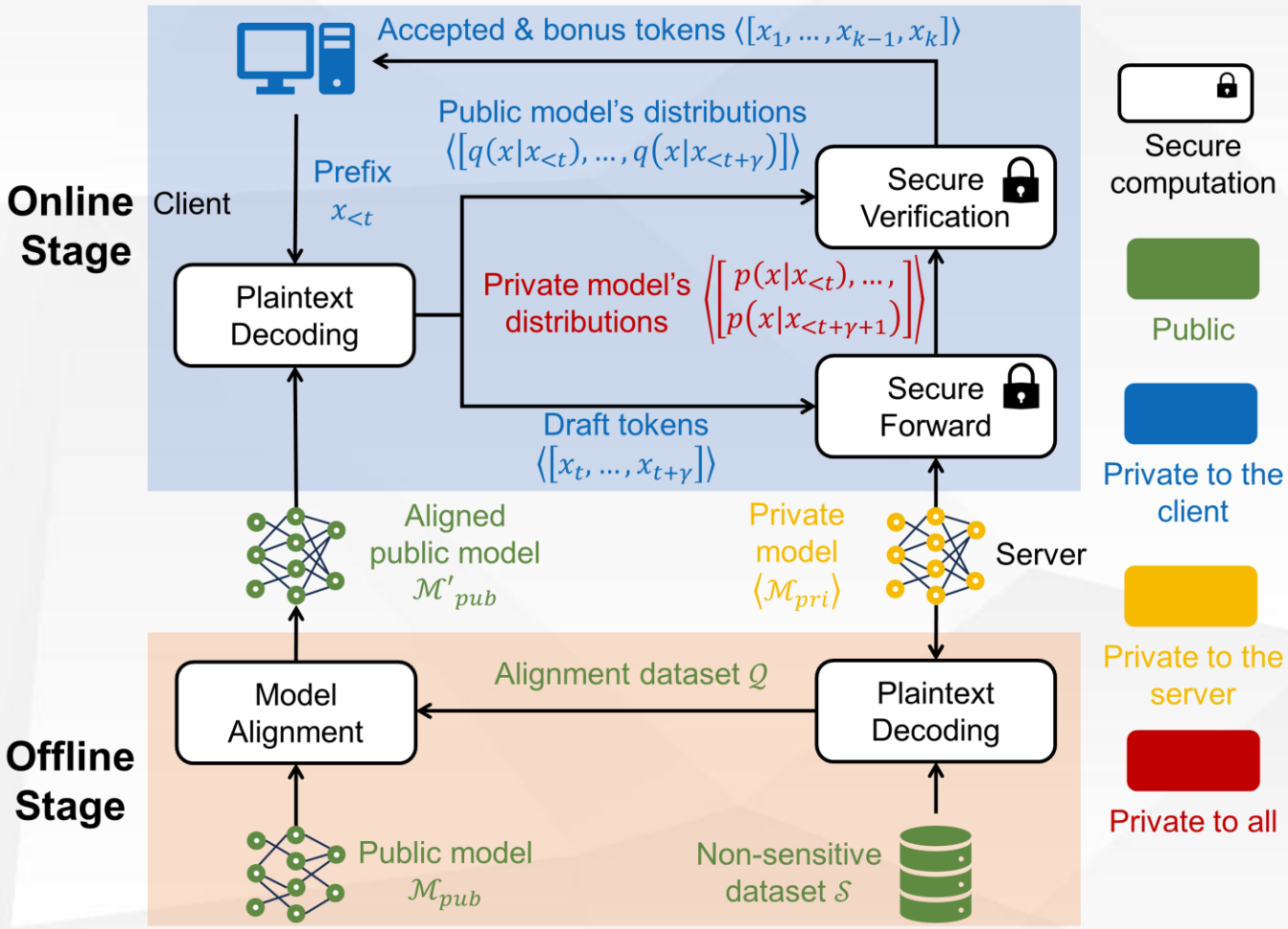
Figure 2: The latency breakdown of some layers.

Bandwidth and one-way delay

- Latencies of secure execution of the GPT are **insensitive** to the input length due to

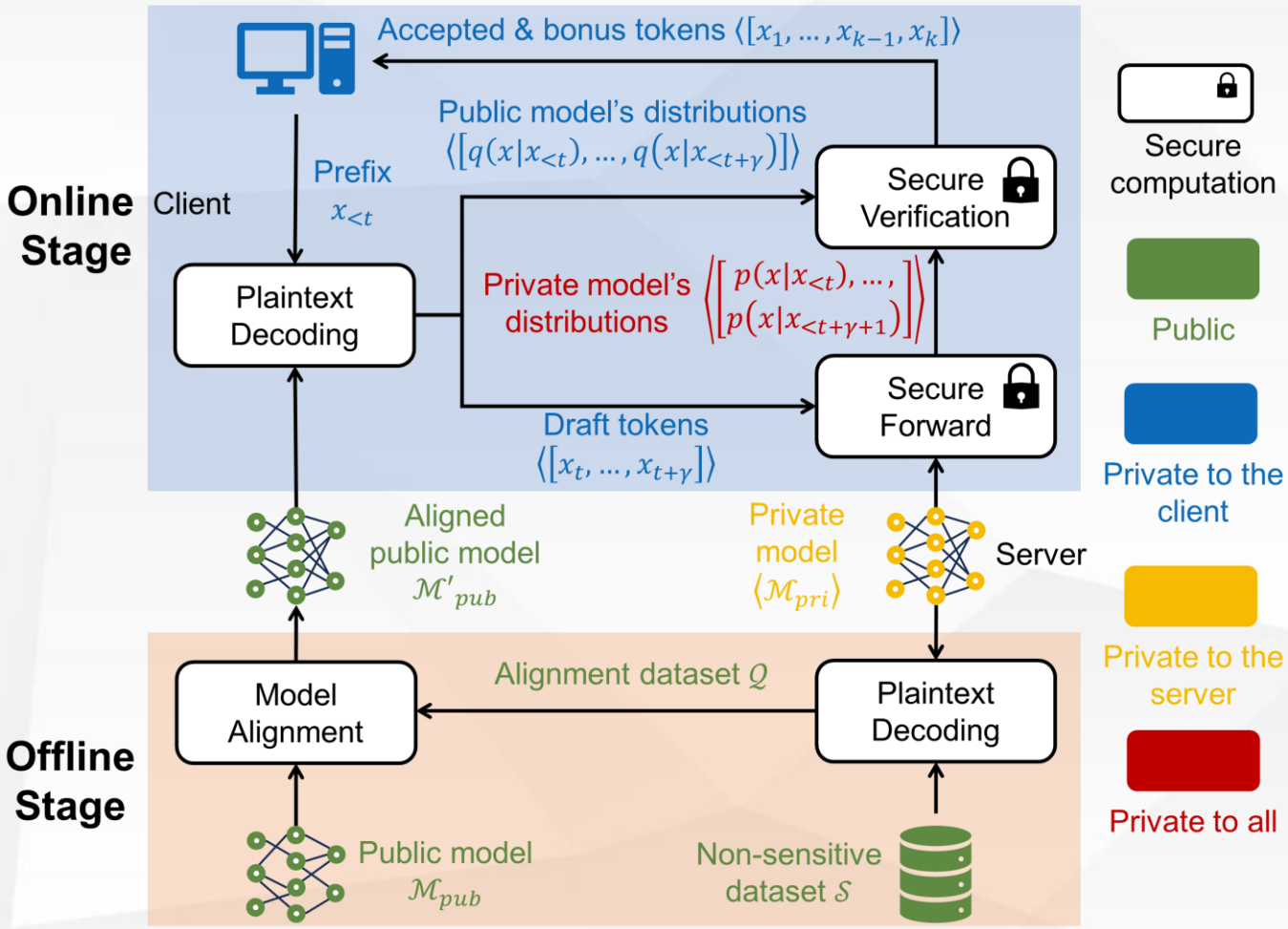
1. Unchanged one-way delay
2. Sub-linear increase in computation latency
3. Linearly increased transmission latency, but not the bottleneck.

Public decOding and Secure verificaTion (POST)



- Public decOding and Secure verificaTion (POST)
  - First generates multiple draft tokens through a public model.
  - Then securely forward multiple draft tokens.
- Similar forward cost but may accept more than one token in a single decoding step.

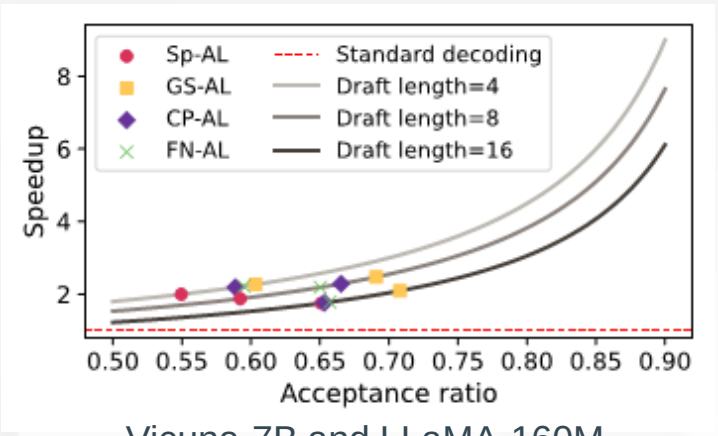
Public decOding and Secure verificaTion (POST)



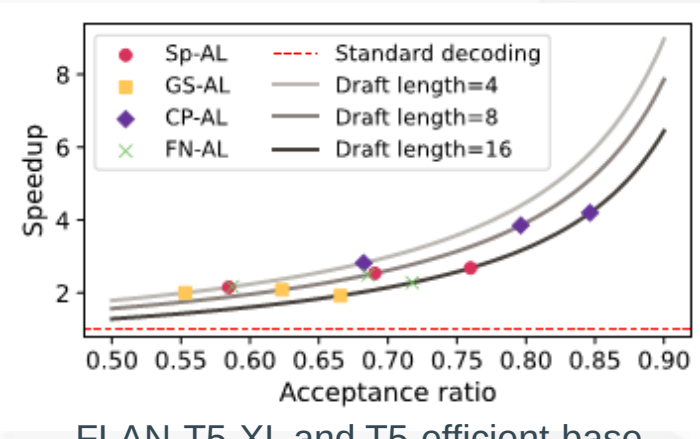
- Further improve the acceptance of draft tokens through:
  - Secure verification: An optimized speculative sampling protocol for 2PC protocol.
  - Model alignment: Knowledge distillation to align the public model's output to the private model.

Experiments

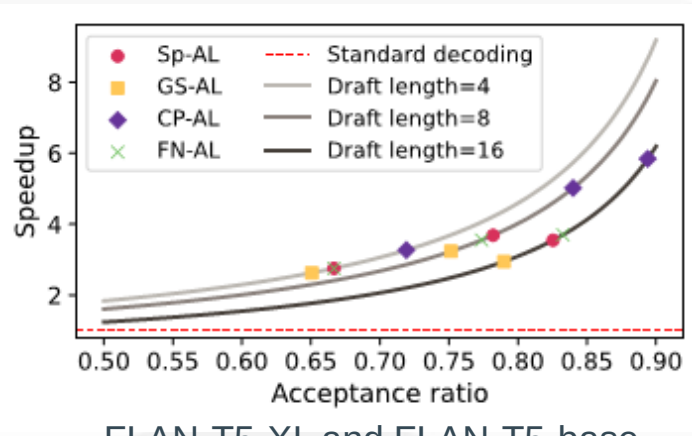
- On three pairs of public and private models
- 1000Mbps bandwidth and 10 ms one-way delay.
- Four tasks: Spider, Gsm8k, Code-search-Python, Alpaca-finance.



Vicuna-7B and LLaMA-160M



FLAN-T5-XL and T5-efficient-base



FLAN-T5-XL and FLAN-T5-base

- 2X~5X speedup across different settings.

Thanks!

饮 水 思 源　爱 国 荣 校