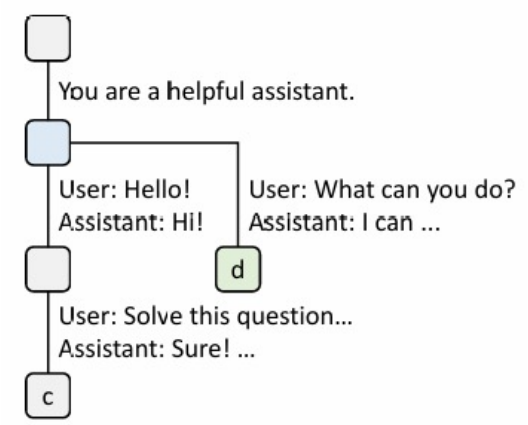
Possible Timing Side Channel of KV Cache

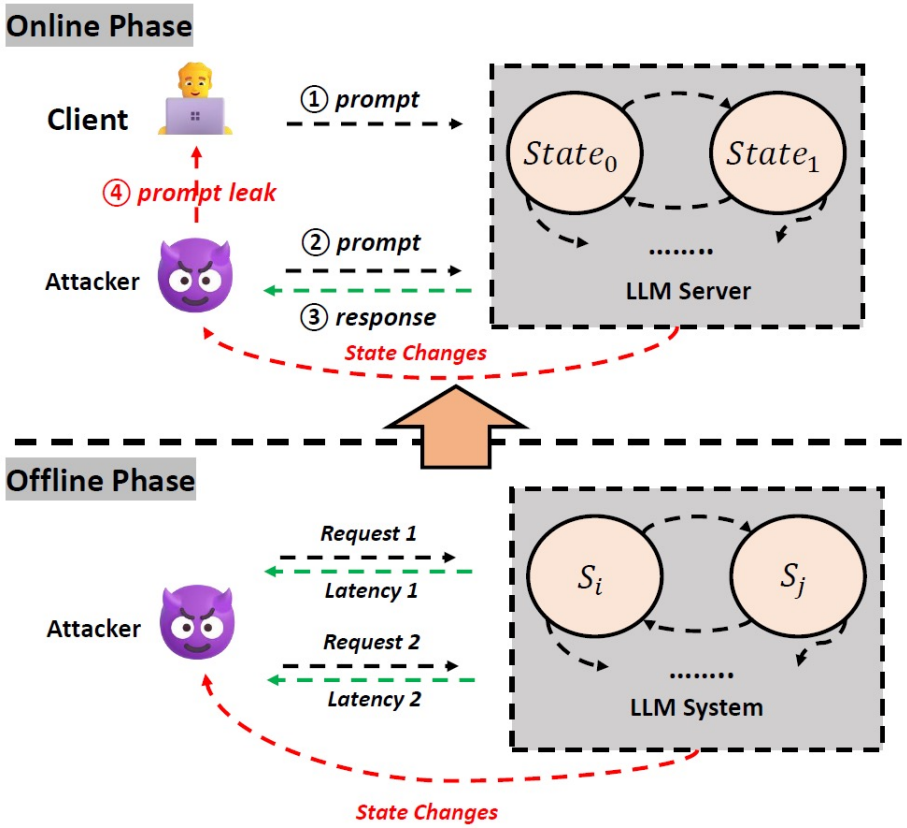
Linke Song
Institute of Information Engineering,
Chinese Academy of Sciences

Motivation: Shared Prefix

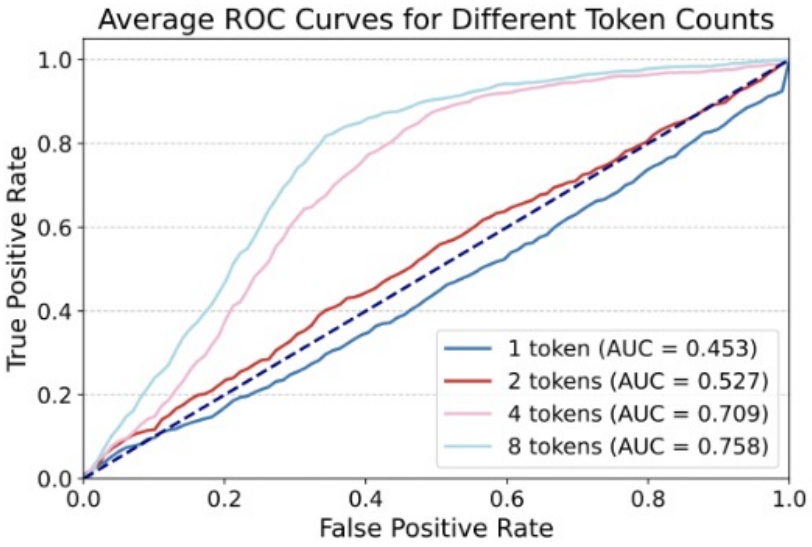
- When a new prompt comes, if the TokenKVPool has its prefix tokens info, the prefill process will be accelerated, which can be reflected in TTFT.



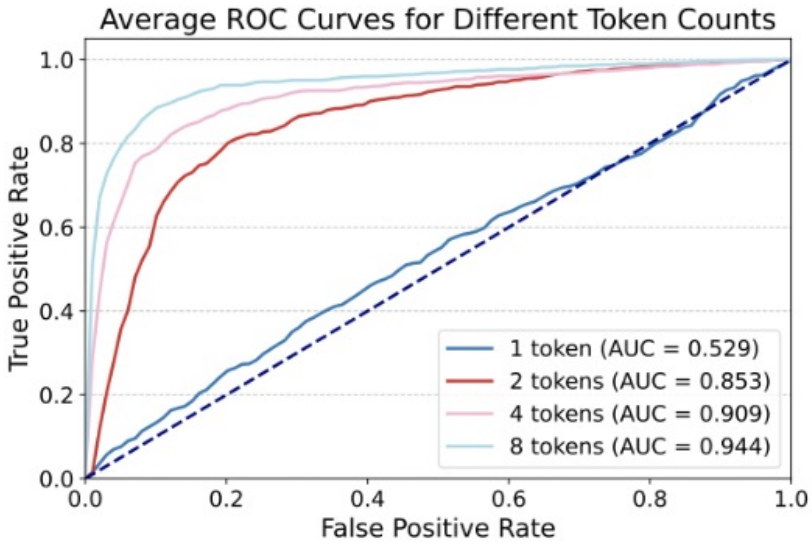
Threat Model



Leakage: TTFT differences SGlang v0.2.6



(a) llama2-13B



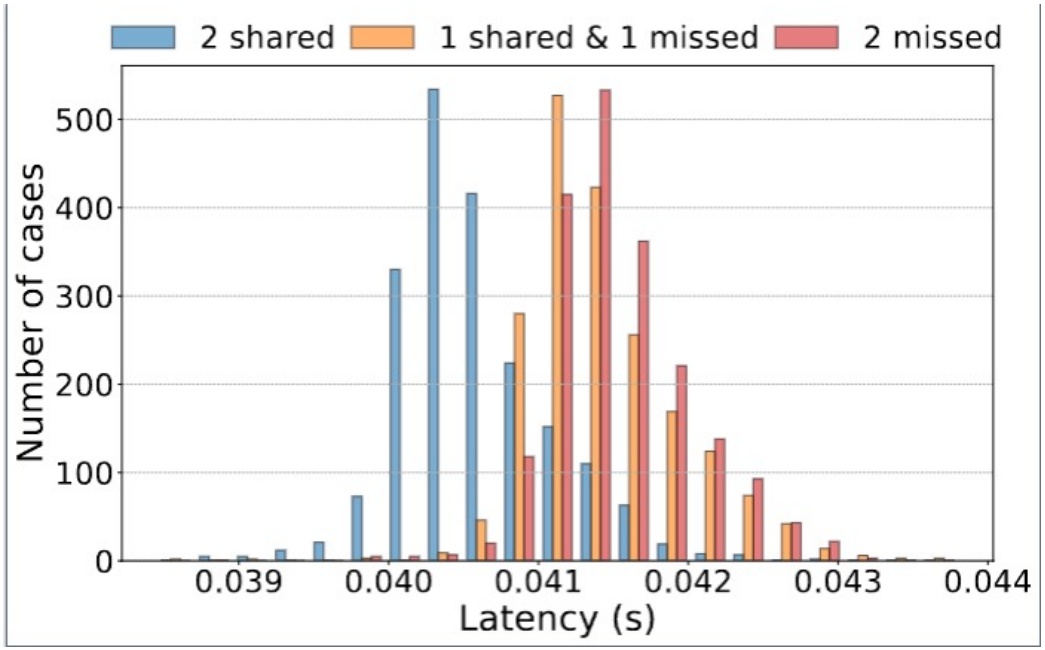
(b) llama2-70B-GPTQ

We keep the length of the sentence the same.

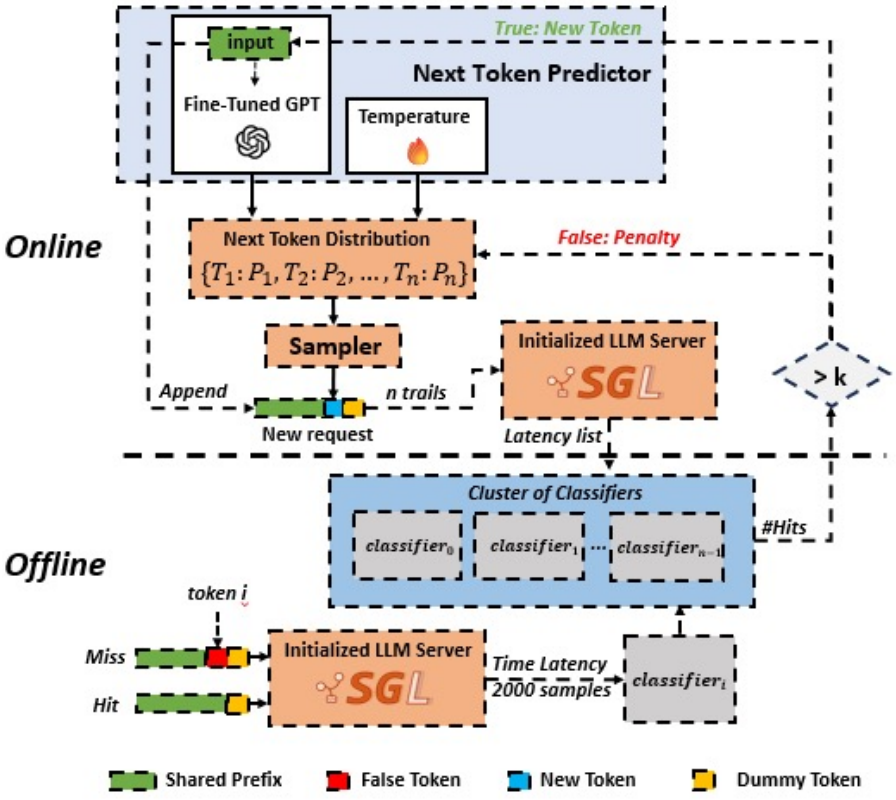
Compare with the prompts sharing no prefix at all.

Select one prompt, Change the first/second/third/fifth/nineth token.

Leakage: TTFT differences



Design



KEY FEATURES:

We assume A Fine-Tuned language model is obtained by the attacker.

Timing differences AUC: 0.529 -> 0.58

Flush Cache: Since when the guess is the same, the later prompt will be accelerated. Therefore, we can test in **More Trails** without interfering to improve the success rate.

Some Difficulties:

- **Not so stable:** one-token timing differences are still too **subtle** to be fingerprinted easily.
  - Some tasks like computing-intensive tasks of CPU may have some **even more subtle** interference to the TTFT

- **Length:** the length of the victim prompt affects the latency of attackers get.
  - You can have a xxxx(10 tokens)
  - You can have a xxxx(512 tokens)
  - **Attacker guess prompt:** You can have => Different TTFT.

Possible Ideas:

- **Not so stable:** if the prediction is not **token-by-token**, but **2tokens-by-2tokens** or **ntokens-by-ntokens**?
  - more times to guess (scaled searching space), but fewer trails (AUC)
  - We write a Simulator here.....

- **Length:** According to the latency, First get the length range of the victim prompt, then select the corresponding classifier.

Possible Mitigations:

- **Detect possible harmful behavior:**
  - consistently asking the same question, asking for flush_cache
  - Always set max_new_tokens=1

- **Isolate TokenKVPool for different users**

- **Increase the granularity of min shared tokens:**
  - As the searching space of attacker scales exponentially, it could cost the attacker forever to recover tokens.

Sum:

- **Leakage exists**

- **Thank u!**

- **Github Issue:** Possible timing side-channels caused by shared prefix · Issue #1504 · sgl-project/sglang