I Know What You Asked: Prompt Leakage via KV-Cache Sharing in Multi-Tenant LLM Serving

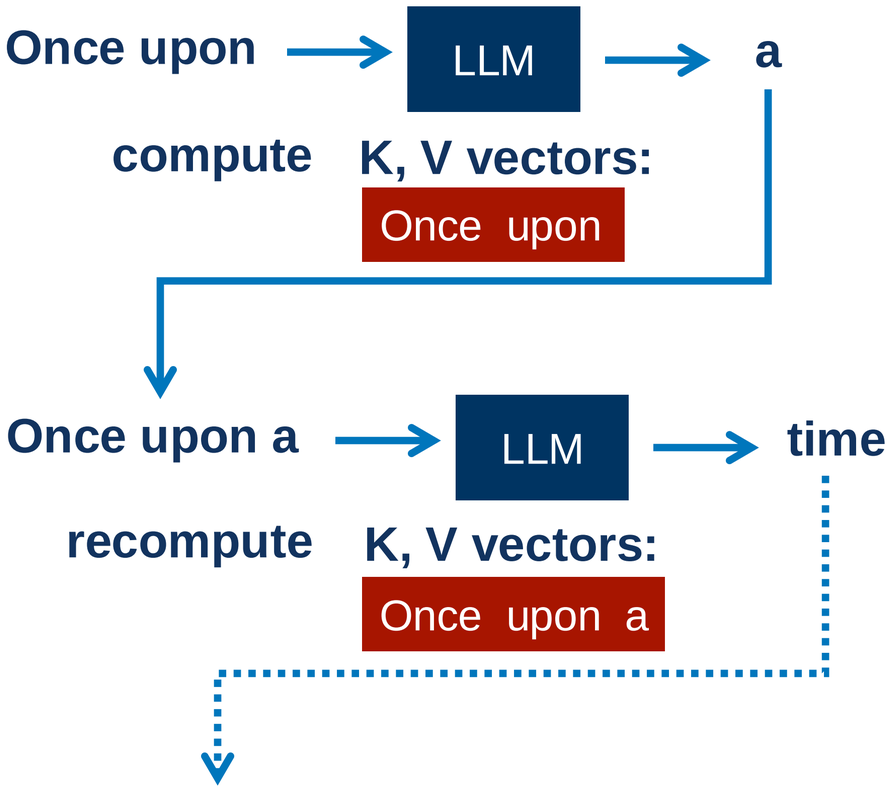
Guanlong Wu[1], Zheng Zhang[2], Yao Zhang[2], Weili Wang[1], Jianyu Niu[1], Ye Wu[2], Yinqian Zhang[1]

[1]Southern University of Science and Technology

[2]ByteDance Inc.

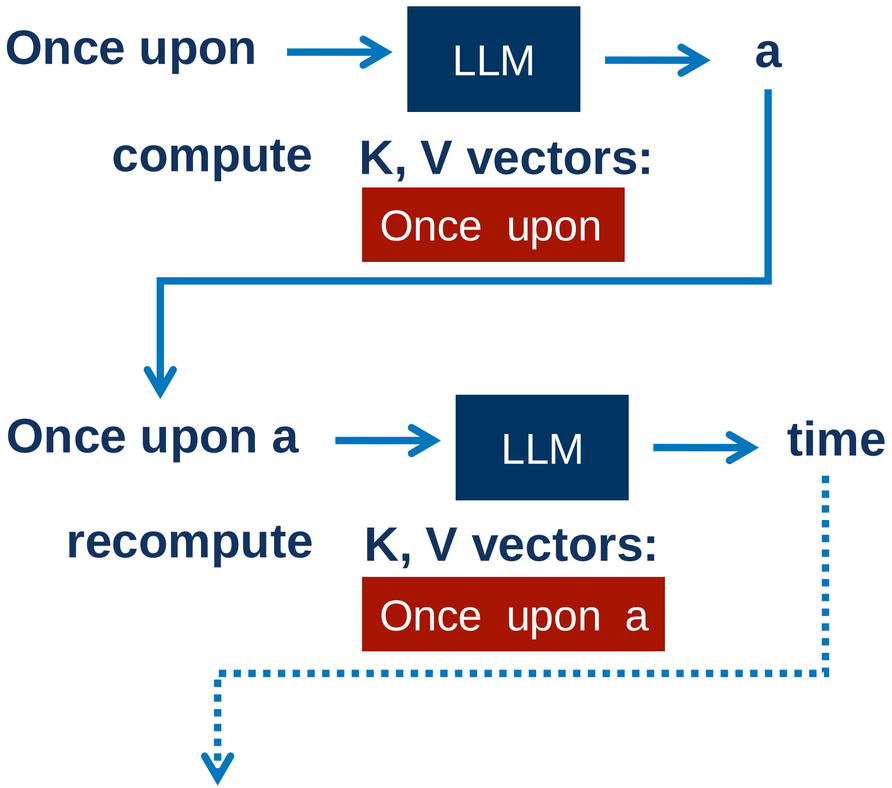
Large Language Model (LLM)

LLM works as a recursive process



Large Language Model (LLM)

LLM works as a recursive process



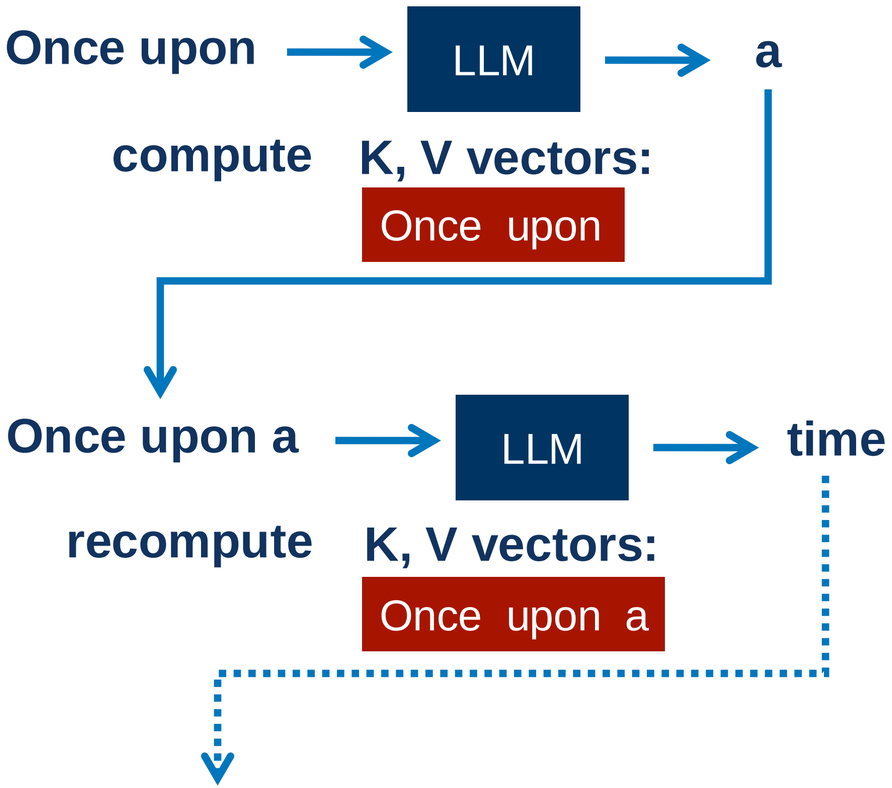
Key, Value (KV)

trade space for time

Large Language Model (LLM)

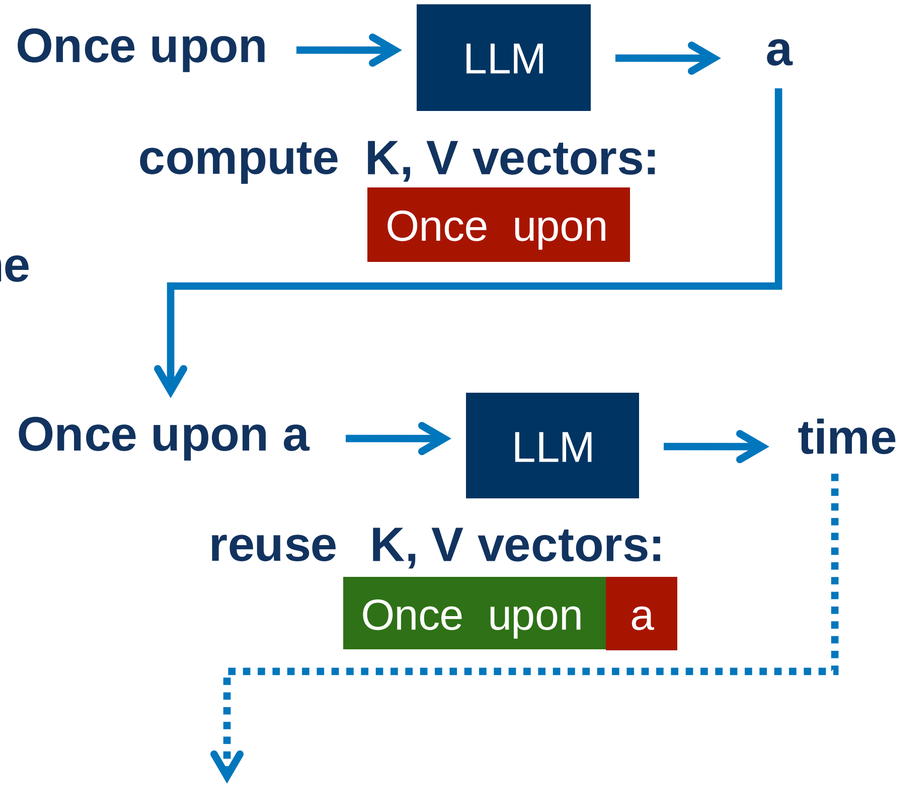
LLM works as a recursive process

***KV cache* becomes a key component for fast LLM serving**



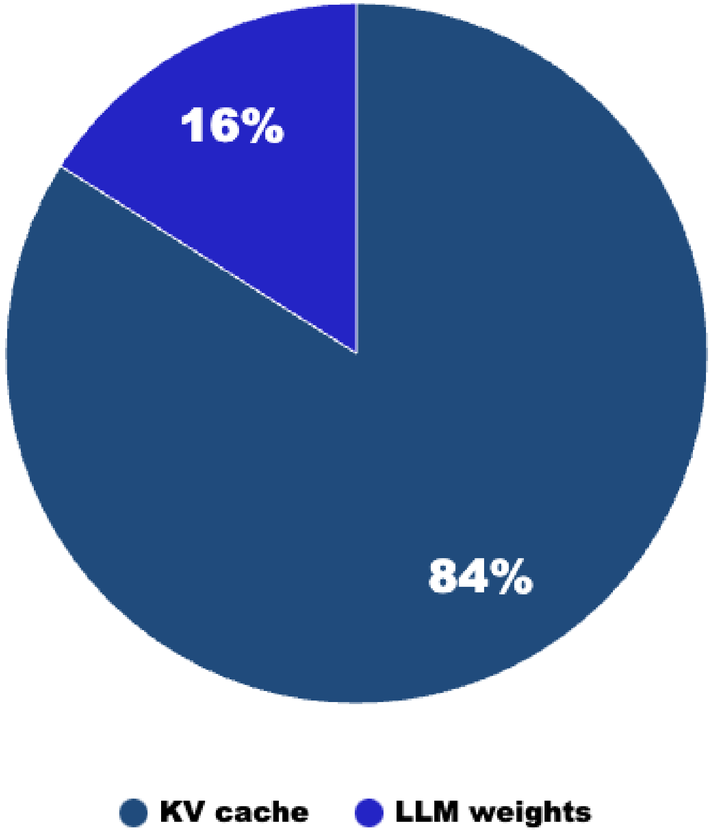
Key, Value (KV)

trade space for time



KV Cache as a Bottleneck

Llama2-7B model weights vs one 128K token prompt under this model

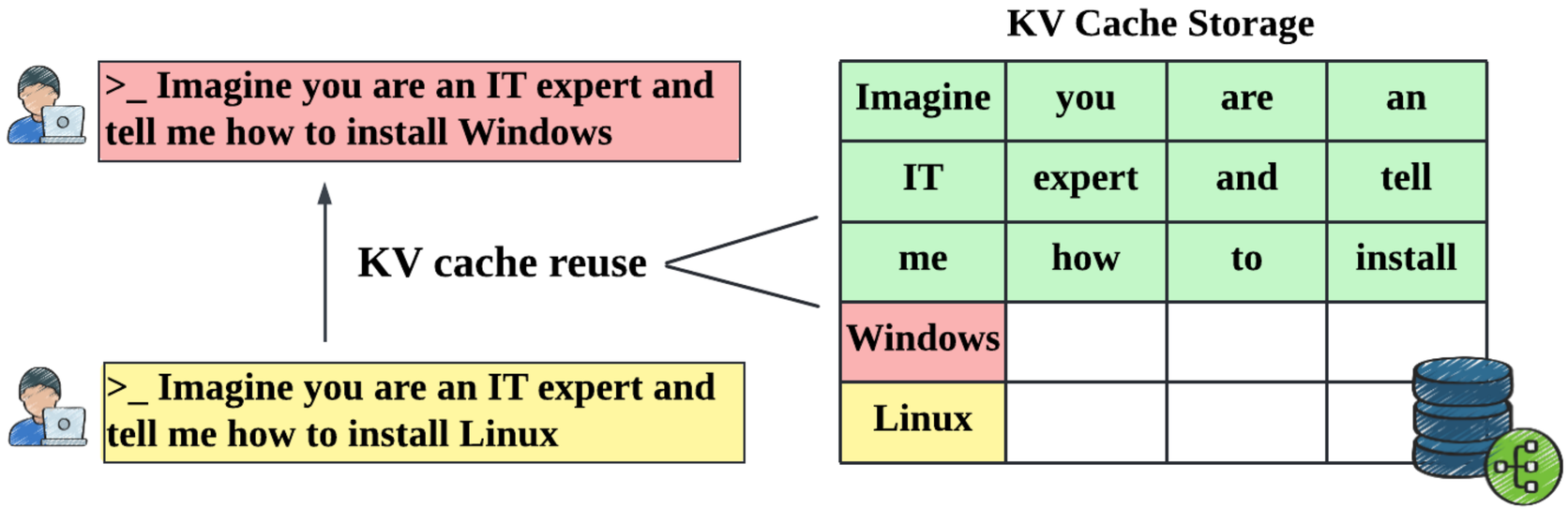


BS	Model	Model Size (GB)	fp16 KV Cache Size with Different Seq. Len. (GB)			
		16 → 2-bit	32K	128K	1M	10M (16 → 2-bit)
1	7B	12.6 → 1.6	16	64	512	4883 → 610
	13B	24.1 → 3.0	25	100	800	7629 → 954
	30B	60.3 → 7.5	49	195	1560	14877 → 1860
	65B	121.1 → 15.1	80	320	2560	24414 → 3052
4	7B	12.6 → 1.6	64	256	2048	19531 → 2441
	13B	24.1 → 3.0	100	400	3200	30518 → 3815
	30B	60.3 → 7.5	195	780	6240	59509 → 7439
	65B	121.1 → 15.1	320	1280	10240	97656 → 12207

Source: Hooper, Coleman, et al., 2024

KV Cache Sharing

Reusing KV cache across users significantly reduces memory consumption



KV Cache Sharing

KV cache can only be reused if all preceding tokens match

Request 1:
Imagine you are an IT expert and tell me how to install Windows

Request 2:
Imagine you are an IT expert and tell me how to install Linux

Request 1:
Imagine you are an IT expert and tell me how to install Windows

Request 2:
Please imagine you are an IT expert and tell me how to install Windows

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
ByteDance 字节跳动

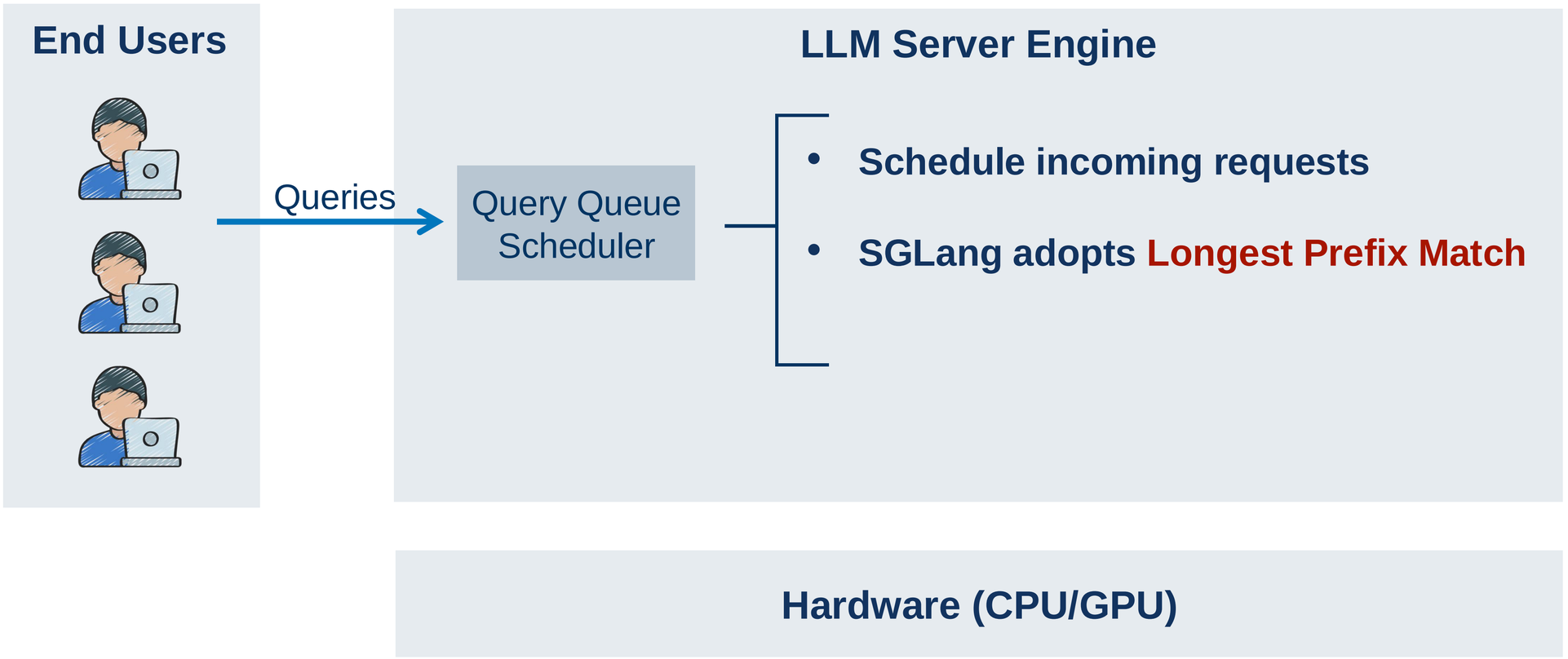
LLM Inference Systems

End Users

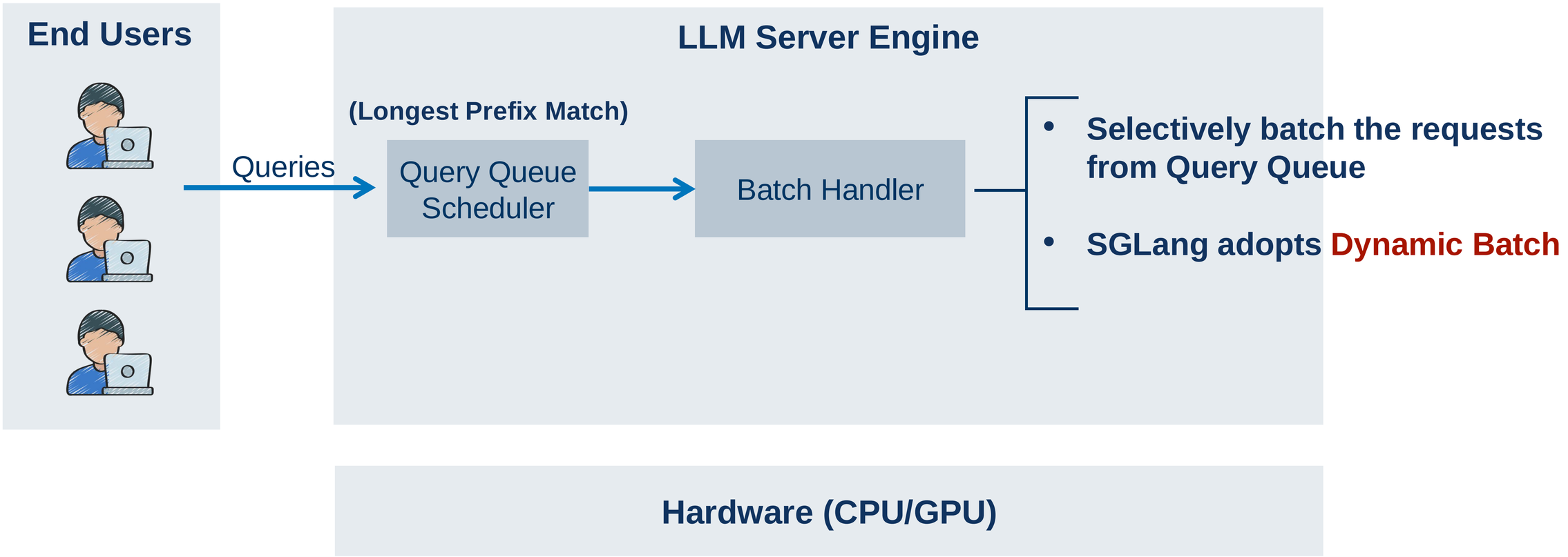
LLM Server Engine

Hardware (CPU/GPU)

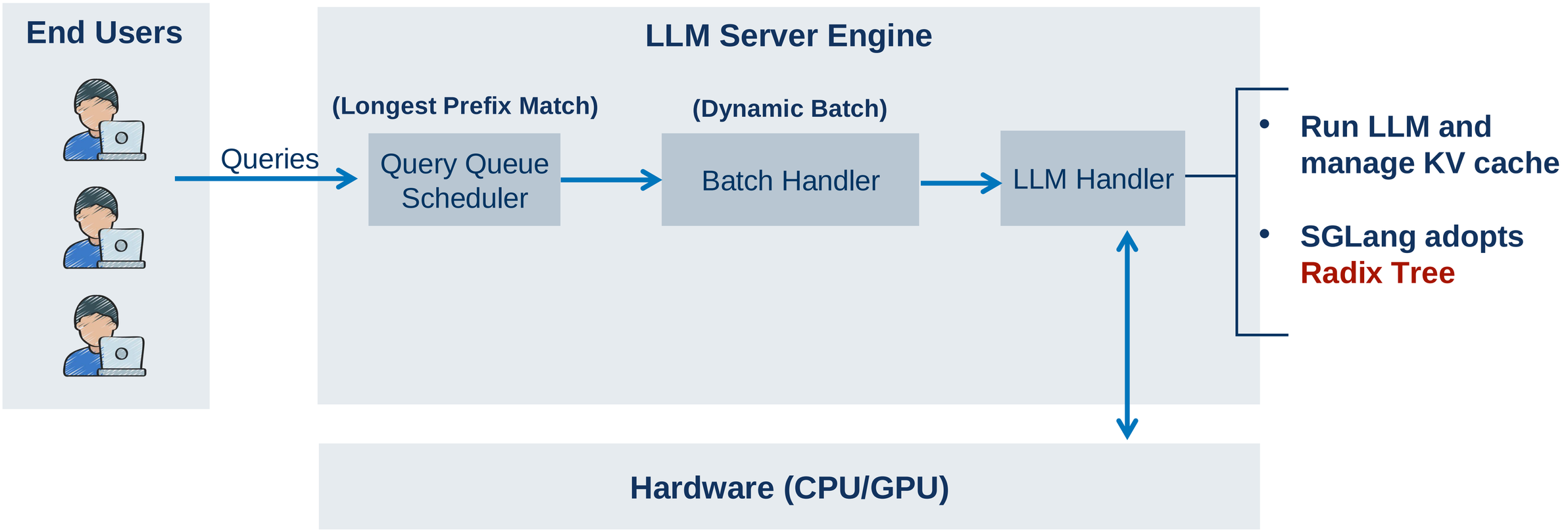
LLM Inference Systems



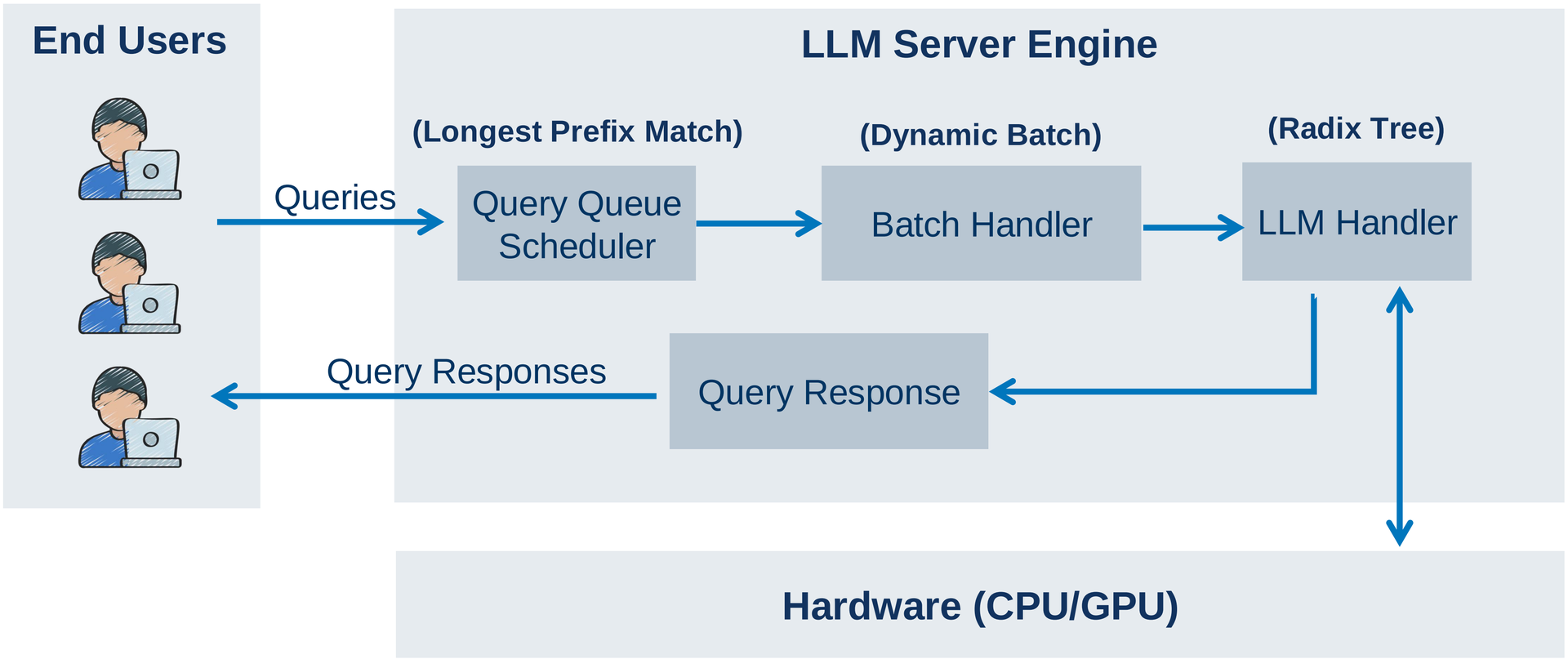
LLM Inference Systems



LLM Inference Systems



LLM Inference Systems



Our Attack

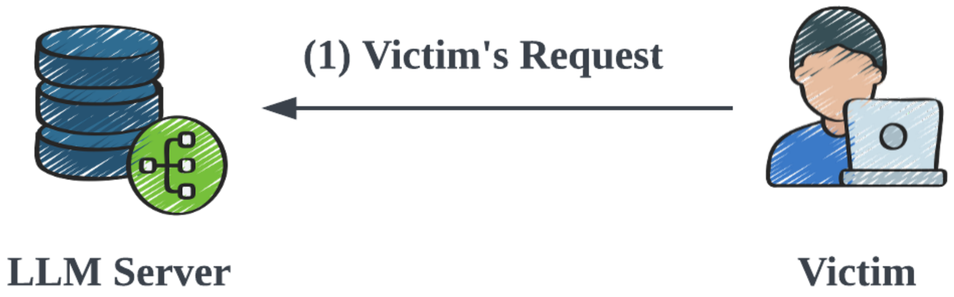

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
Teecert
Labs
ByteDance
字节跳动

Attack Overview

<u>Attack Core:</u> The adversary can detect if its request matches a previous one by observing whether KV cache sharing is triggered.

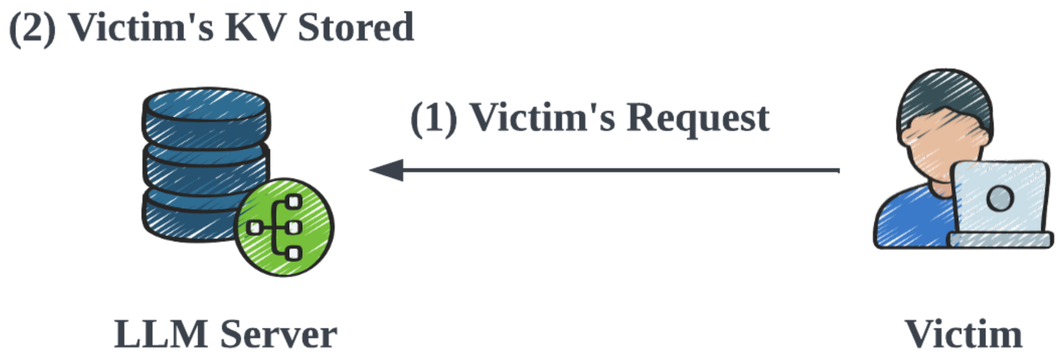
Attack Overview

<u>Attack Core:</u> The adversary can detect if its request matches a previous one by observing whether KV cache sharing is triggered.



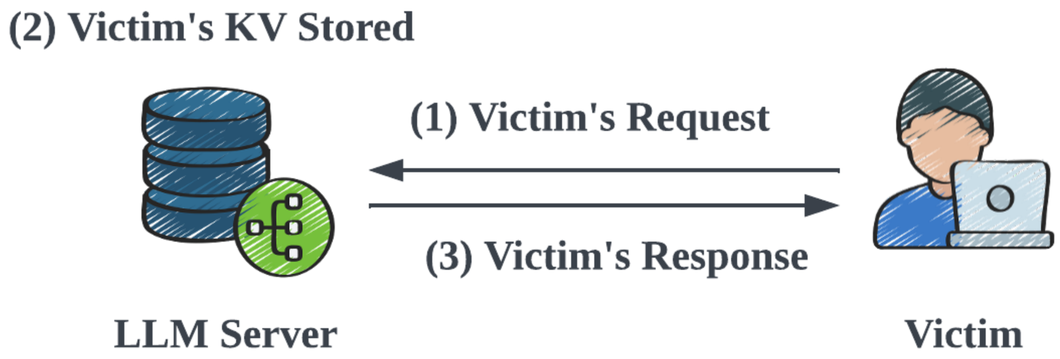
Attack Overview

<u>Attack Core:</u> The adversary can detect if its request matches a previous one by observing whether KV cache sharing is triggered.



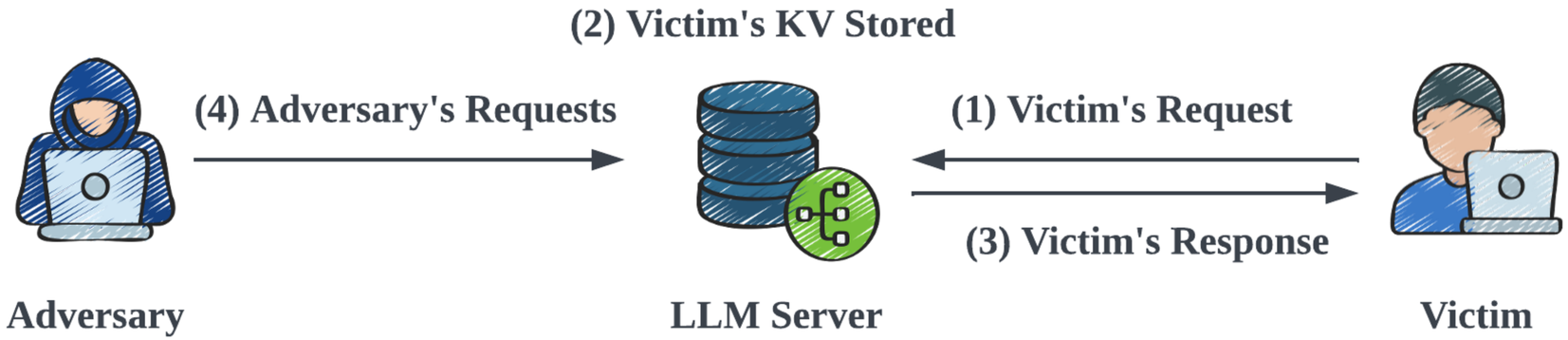
Attack Overview

<u>Attack Core:</u> The adversary can detect if its request matches a previous one by observing whether KV cache sharing is triggered.



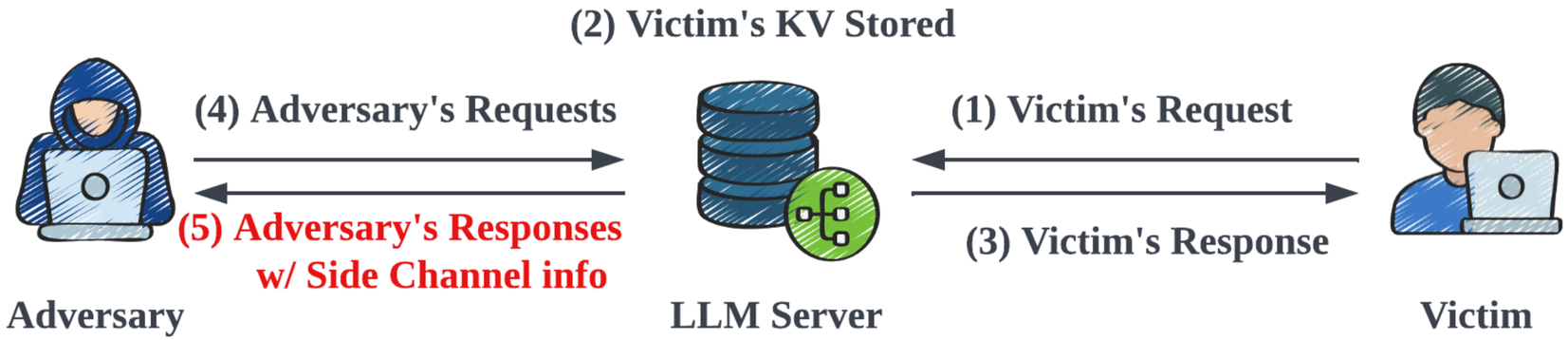
Attack Overview

<u>Attack Core:</u> The adversary can detect if its request matches a previous one by observing whether KV cache sharing is triggered.



Attack Overview

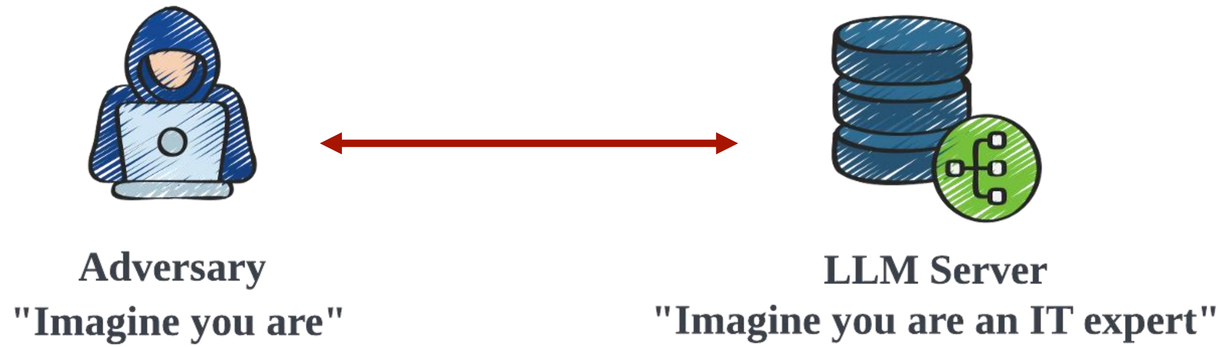
<u>Attack Core:</u> The adversary can detect if its request matches a previous one by observing whether KV cache sharing is triggered.



Token-by-token Extraction

Assume a previously served request: “Imagine you are an IT expert”

The adversary has already extracted: “Imagine you are”

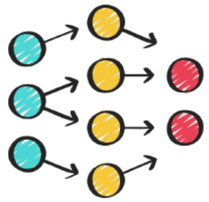


Token-by-token Extraction

Use a local LLM to predict possible tokens

Token-by-token Extraction

Also generate a dummy token for side channel effect



Local LLM

(1) Trigger local LLM to guess
"Imagine you are"

Adversary
"Imagine you are"

LLM Server
"Imagine you are an IT expert"

(2) Generate candidates:

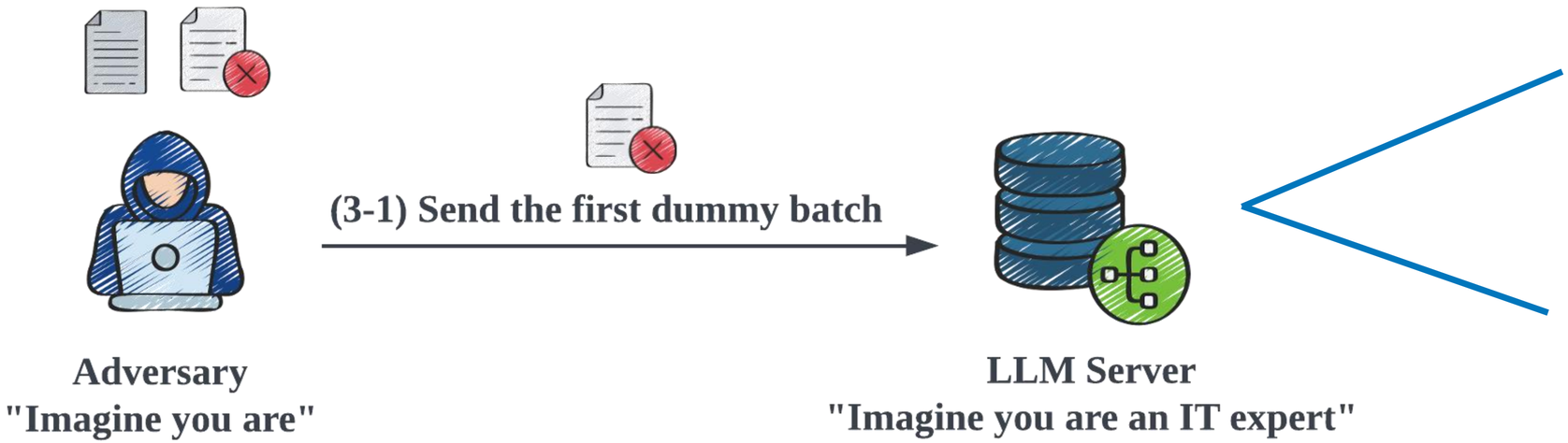
- Imagine you are an
- Imagine you are a
- Imagine you are the
- ...

Also generate dummy:

- Imagine you are %

Token-by-token Extraction

Send three batches of requests in turn



Query Queue

Imagine you are %

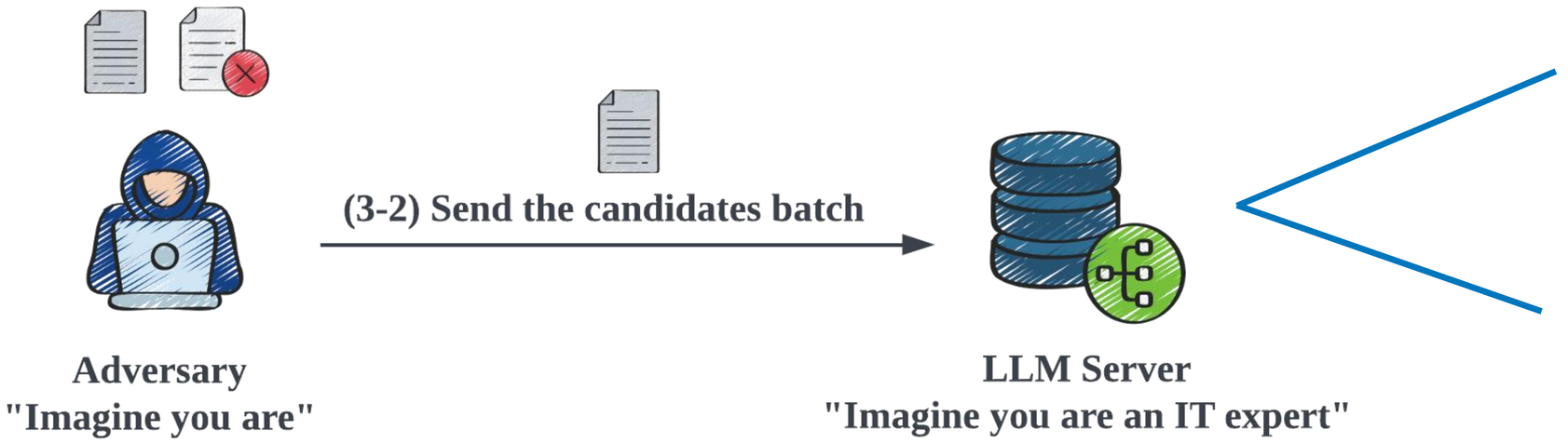
Imagine you are %

Imagine you are %

...

Token-by-token Extraction

Send three batches of requests in turn



Query Queue

Imagine you are %

Imagine you are %

Imagine you are %

Imagine you are a

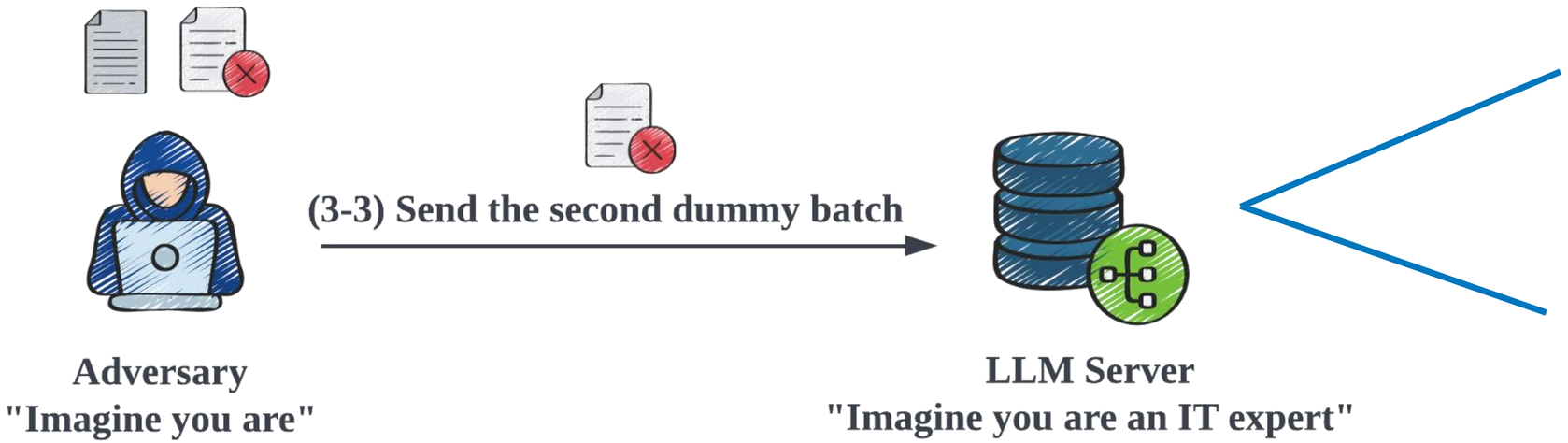
Imagine you are an

Imagine you are the

...

Token-by-token Extraction

Send three batches of requests in turn



Query Queue

Imagine you are %

Imagine you are %

Imagine you are %

Imagine you are a

Imagine you are an

Imagine you are the

Imagine you are %

Imagine you are %

Imagine you are %

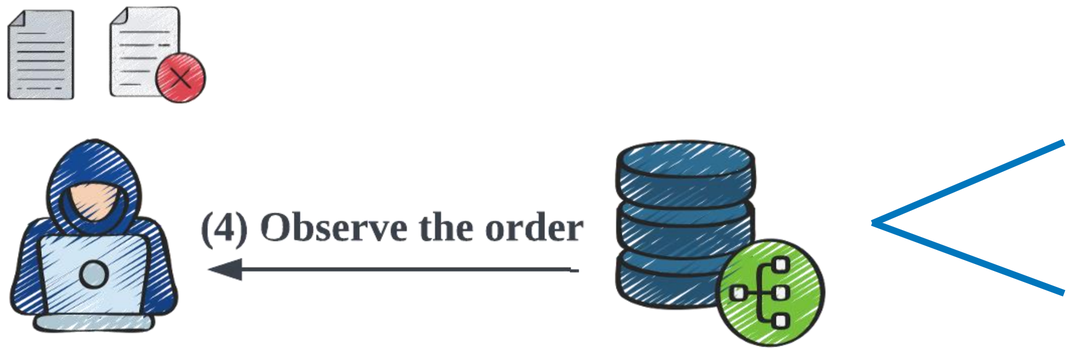
...

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

ByteDance 字节跳动

Token-by-token Extraction

We leverage serving order as a side-channel effect, as the longer token matches can be served first



Query Queue

Imagine you are %
Imagine you are %
Imagine you are %
Imagine you are a
Imagine you are an
Imagine you are the
Imagine you are %
Imagine you are %
Imagine you are %
...

After LPM

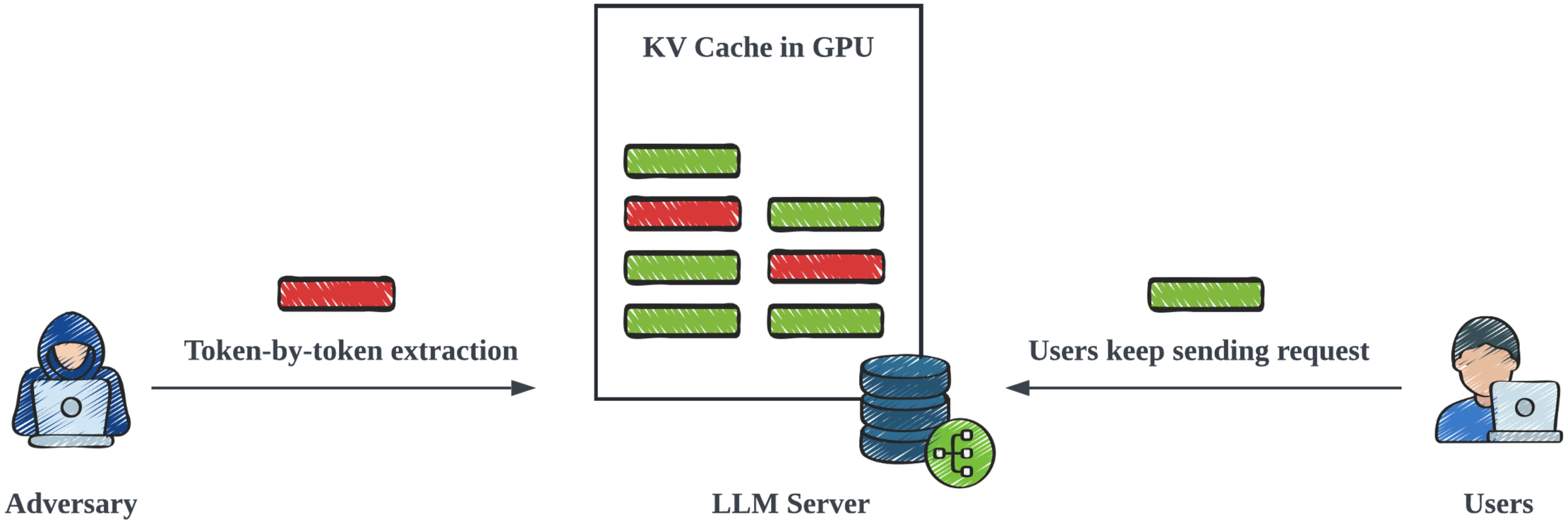
Query Queue

Imagine you are %
Imagine you are %
Imagine you are %
Imagine you are an
Imagine you are %
Imagine you are %
Imagine you are %
Imagine you are a
Imagine you are the
...

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
ByteDance 字节跳动

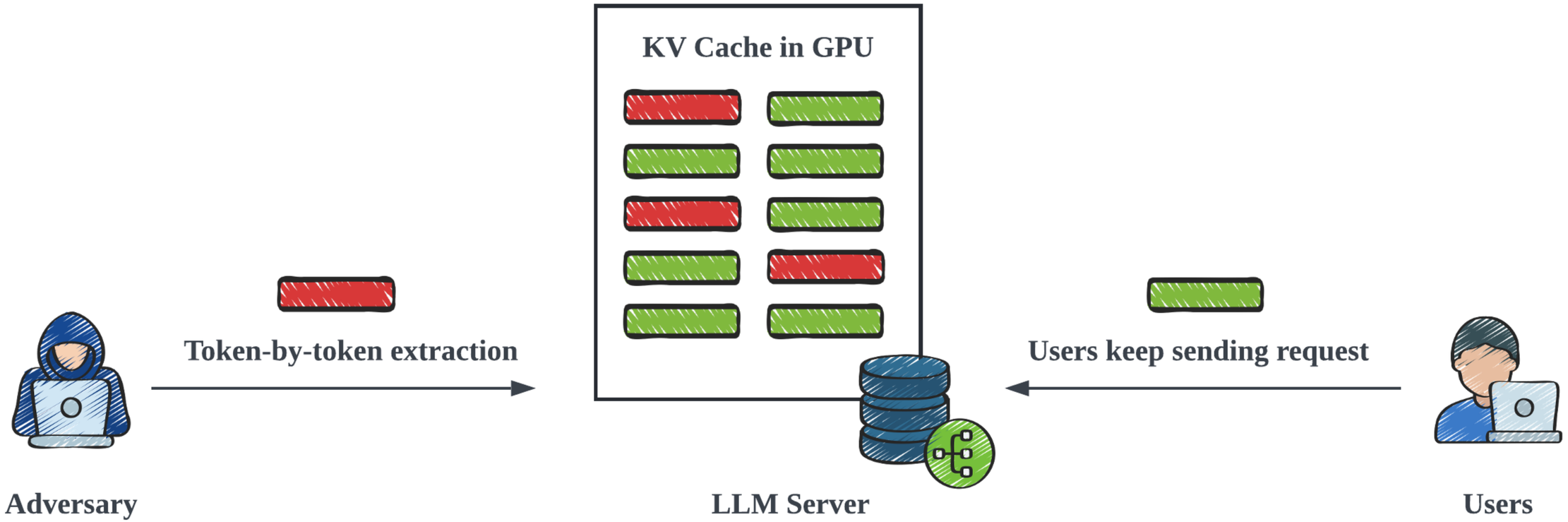
Complete Attack Flow

The adversary tracks a random prompt and use token-by-token extraction



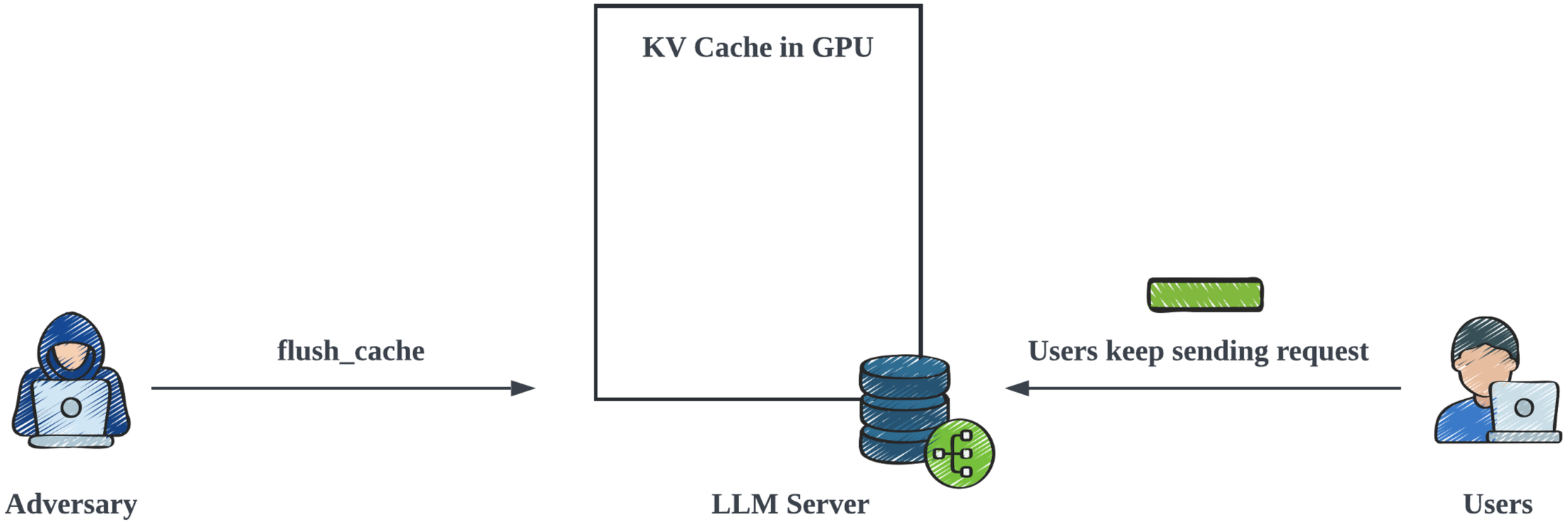
Complete Attack Flow

The adversary switches to another prompt if the tracked prompt is evicted or the next token is hard to guess



Complete Attack Flow

The adversary uses *flush_cache* in SGLang to clear KV storage and switch prompts from a clean slate (our paper adopts a more complex alternative when *flush_cache* is unavailable)

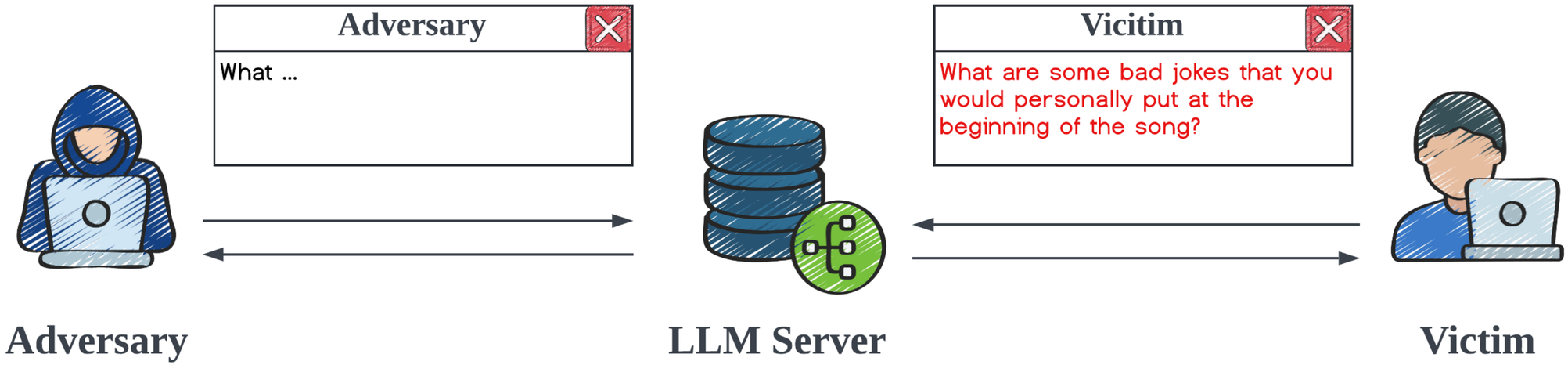


南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
ByteDance 字节跳动

Attack Scenarios

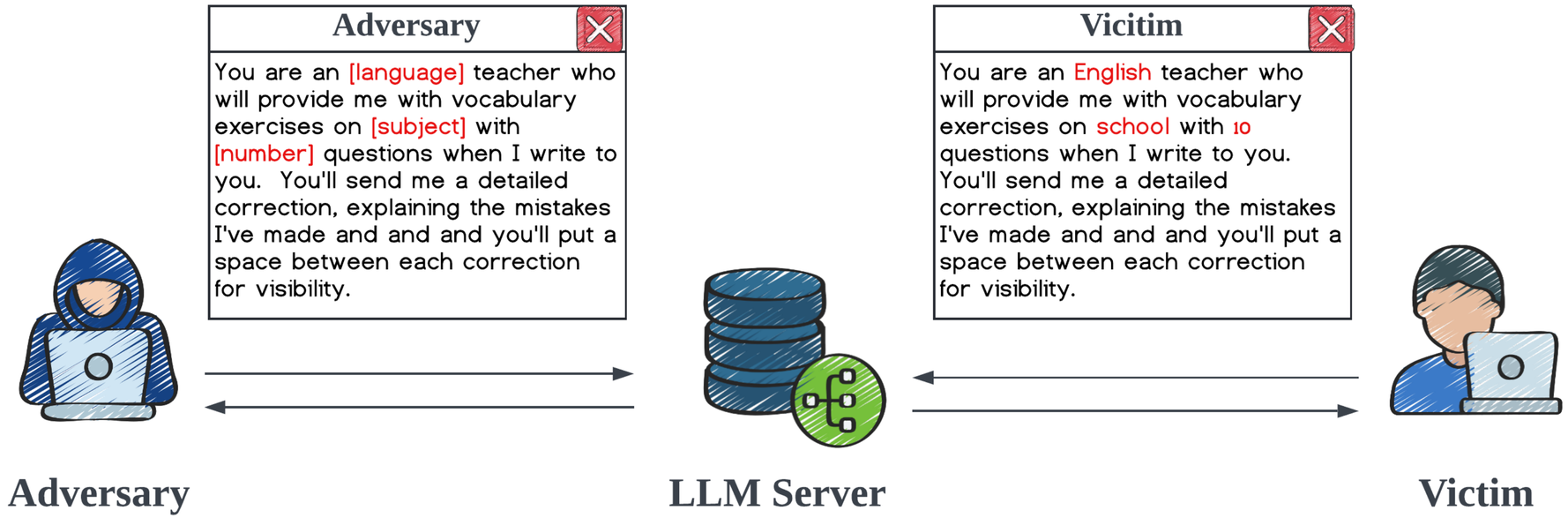
Scenario 1

The adversary has *no background knowledge* and extracts all tokens to reverse the full prompt from another user



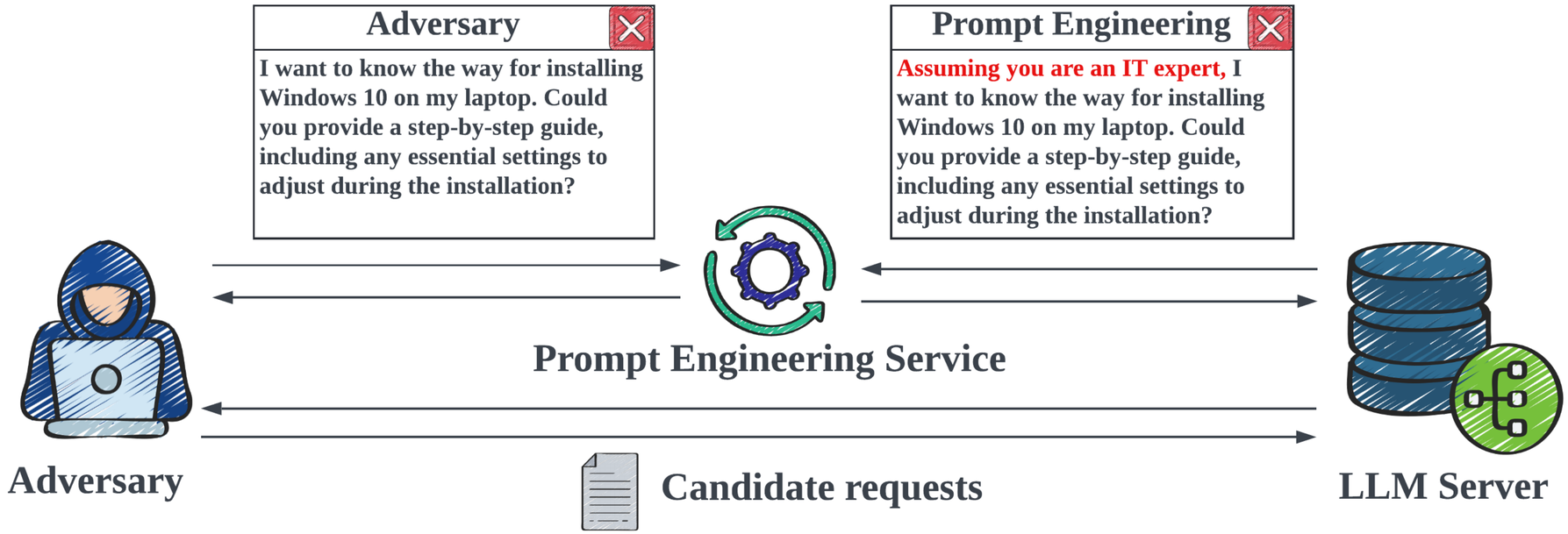
Scenario 2

The adversary knows *the prompt template* and extracts only a few key tokens to steal sensitive information from another user



Scenario 3

The adversary knows *the prompt input* and aims to steal prompt template (valuable in today's LLM application)



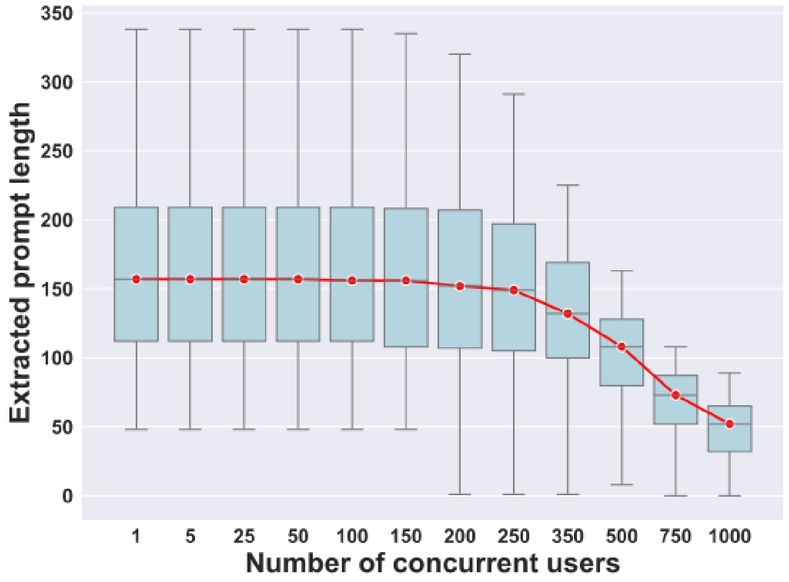
Evaluations

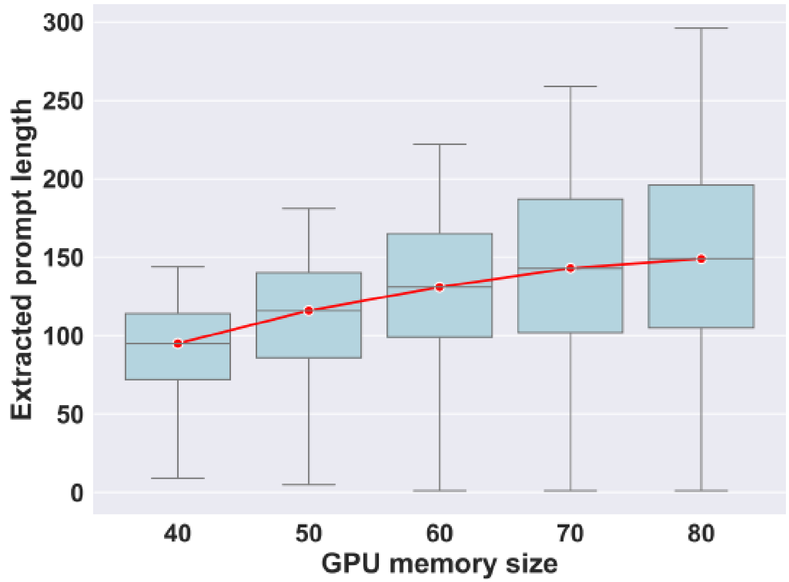
Evaluation Setup

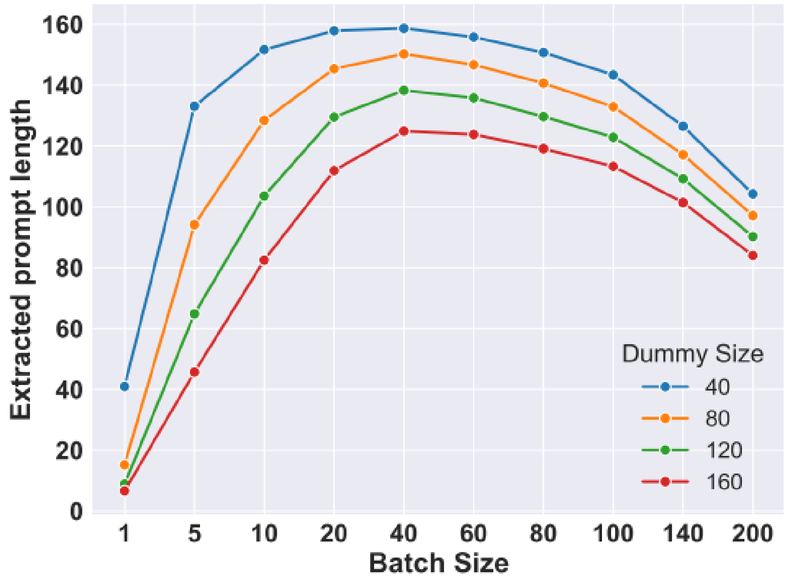
- LLM server configuration: Llama2-13B, Llama3-8B-GQA
- User configuration: 40 requests every 3 hours per user (OpenAI)
- Four datasets: ultrachat, PromptBase, awesome-chatgpt, alpacca
- Three scenarios: whole prompt reconstruction, input reconstruction, template reconstruction
- Two research questions:
  - How **effective** is the attack?
  - How much **cost** of the attack?

How effective is the attack?

Three decisive factors: memory capacity, concurrent users' requests, attack strategy





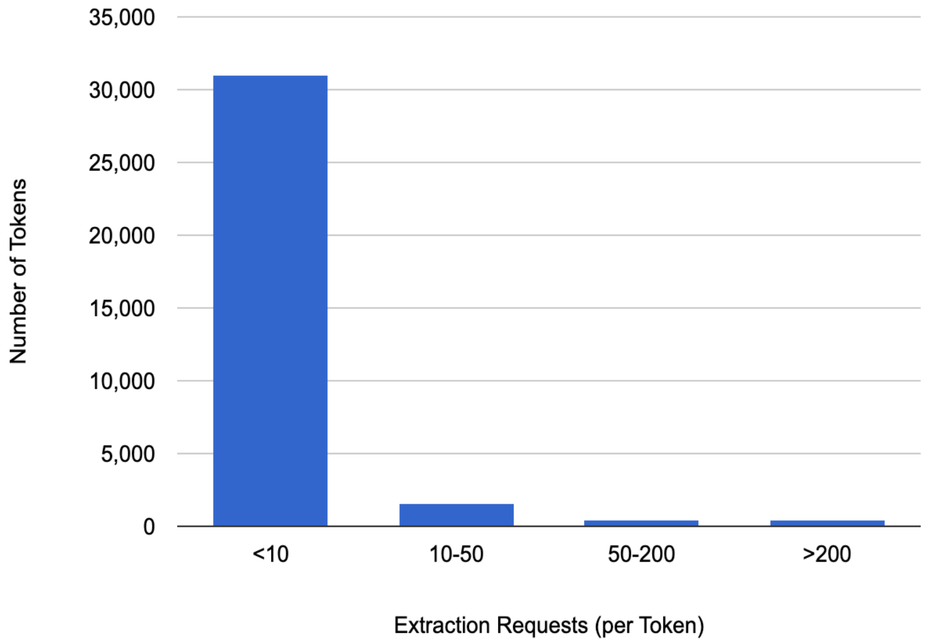


南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

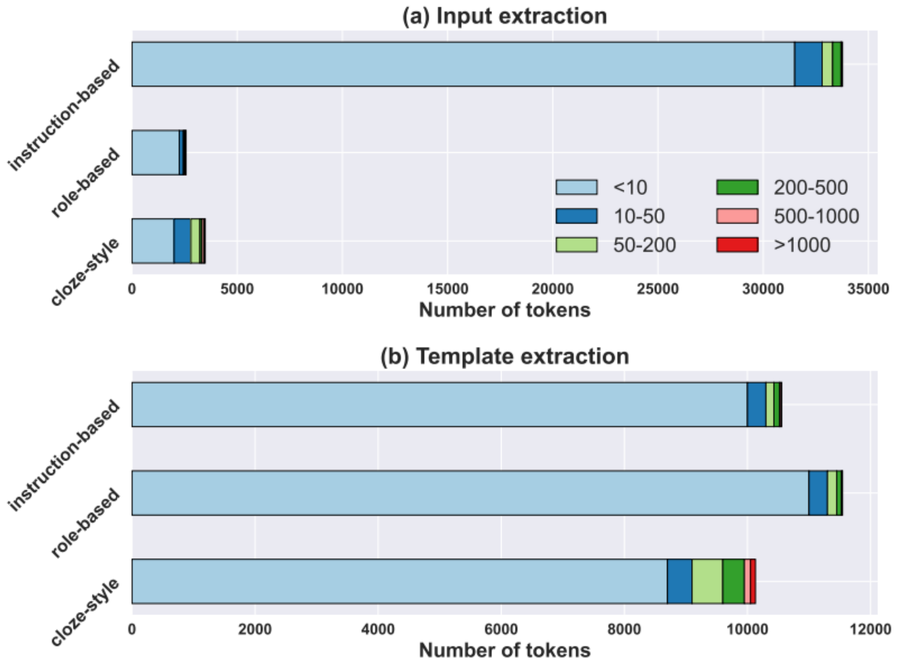
ByteDance 字节跳动

How much cost of the attack?

Most tokens can be reversed with less than 10 guesses



Scenario 1



Scenario 2 & Scenario 3

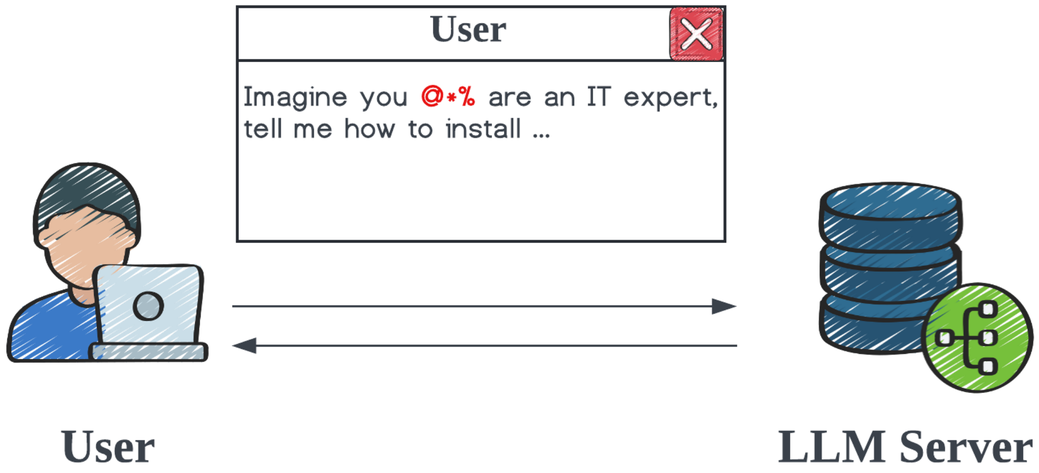
More evaluation on all three scenarios can be found in our paper

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

ByteDance 字节跳动

Countermeasures

- Prioritizing requests with multiple matched tokens instead of one, which significantly raises attack cost while preserving performance.
- Adding rare tokens to the prompt to disrupt the token-by-token attack



Conclusions

- We point out that resource sharing in multi-tenant LLM systems introduces a new attack surface for LLM security
- We propose an attack targeting the KV cache sharing mechanism to extract prompts from other users
- We outline the necessary attack conditions for resource sharing in multi-tenant LLM systems, offering guidance to framework designers and service providers for more secure design

Thanks!

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

ByteDance
字节跳动