Building Trustworthy LLM

Shuai WANG
HKUST CSE
2024-04-05

Education

2013 – 2018	Penn State Univ.	IST	Ph.D.
2008 – 2012	Peking Univ.	EE	B.S.

Appointment

2019 – now	HKUST	CSE	Assistant prof.
2018 – 2019	ETH Zurich	CSE	Postdoc

Research Interest

Software security | LLM security | Reverse engineering | System security

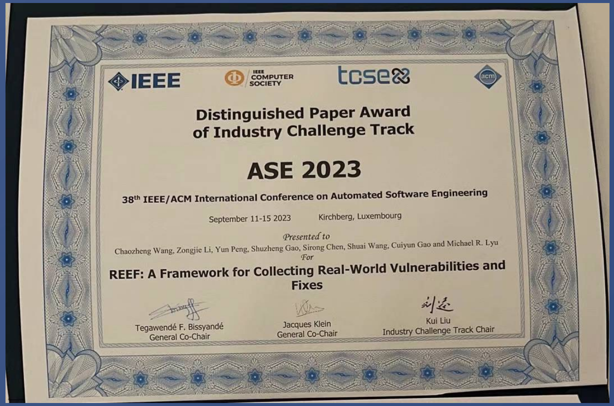
Red Team

- Offensive Security
- Ethical Hacking
- Exploiting Vulnerabilities
- Penetration Tests
- Black Box Testing
- Social Engineering
- Web App Scanning

Blue Team

- Defensive Security
- Infrastructure Protection
- Damage Control
- Incident Response
- Operational Security
- Threat Hunting
- Digital Forensics

9-10 researchers (two Postdocs and 7-8 Ph.D. students) in this LLM security line of research



IEEE IEEE COMPUTER SOCIETY tcse acm

Distinguished Paper Award
of Industry Challenge Track

ASE 2023

38th IEEE/ACM International Conference on Automated Software Engineering

September 11-15 2023 Kirchberg, Luxembourg

Presented to

Chaozheng Wang, Zongjie Li, Yun Peng, Shuzheng Gao, Sirong Chen, Shuai Wang, Cuiyun Gao and Michael R. Lyu

For

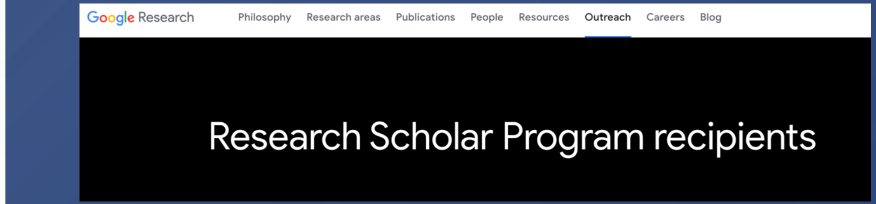
REEF: A Framework for Collecting Real-World Vulnerabilities and Fixes

Tegawendé F. Bissyandé
General Co-Chair

Jacques Klein
General Co-Chair

Kui Liu
Industry Challenge Track Chair

ACM SIGSOFT Distinguished Paper



Research Scholar Program recipients

2023 Google Award

2023 BlackHat USA

Roadmap – Building Trustworthy LLM

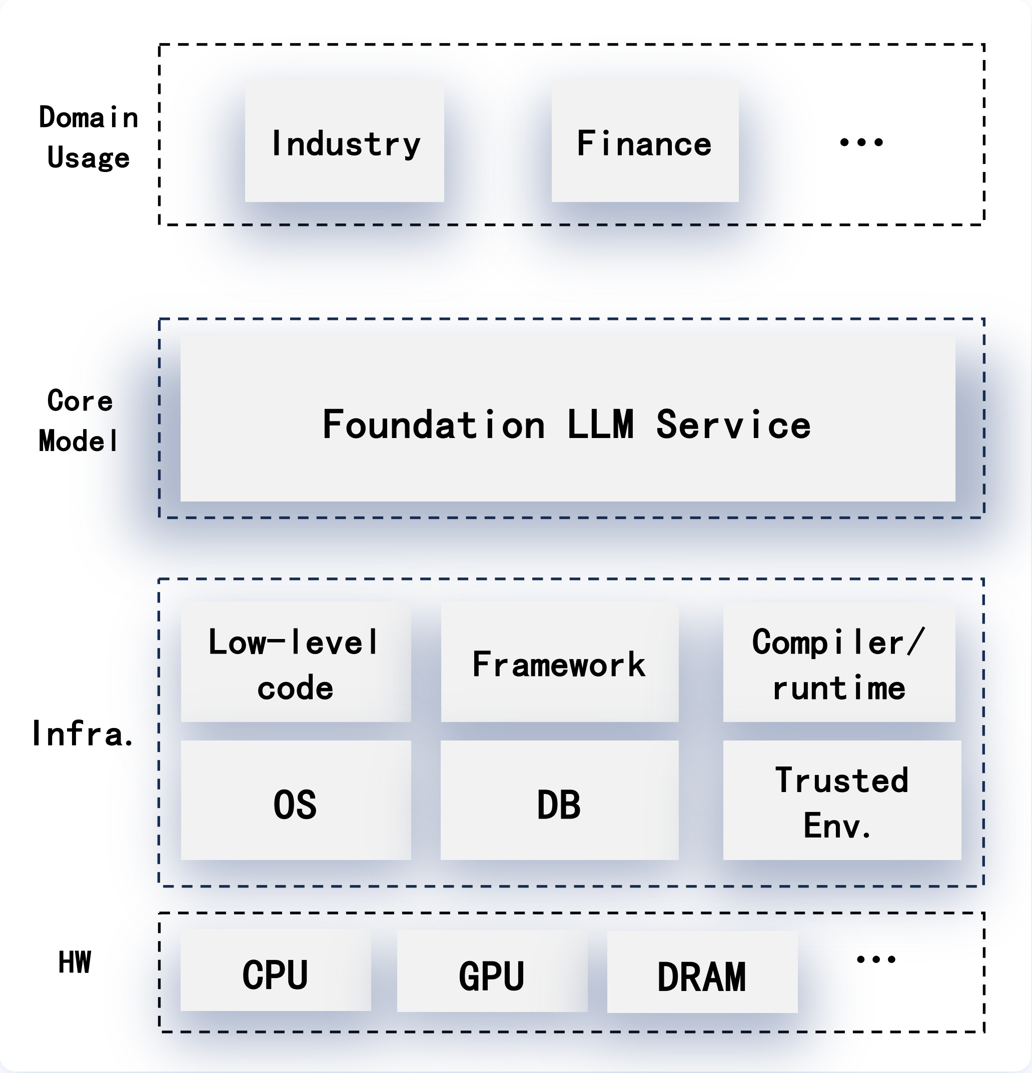
🟢 deliveries from our group
(papers, released datasets, other IPs)

🔵 joint effort with industry partners
(papers, patents, tool dissemination)

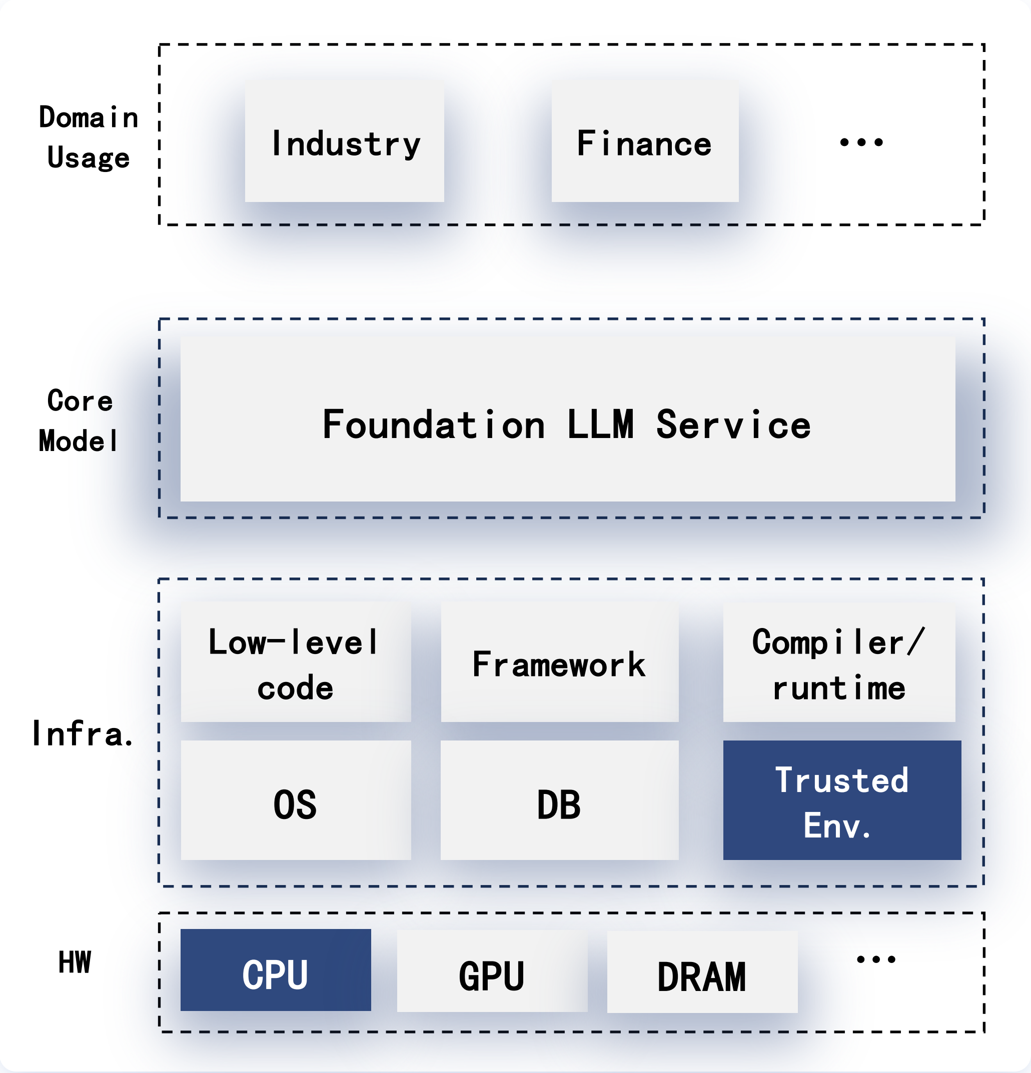
🔴 ongoing/planned

Object	Impact Stage	Category	Cases (simplified for presentation purposes)						
Foundation LLM	Offline Training	**"red team"**	training data poisoning	fine-tuning data poisoning	quantization risk 🔴	privacy-enhancing risk 🟢🔴	pre-trained data leakage	correction model training	
		"blue team"	data cleaning	secure SFT	RLHF	model alignment 🟢🔴	Pre-trained data protection	aligned dataset auto generation	
	Online Deployment	**"red team"**	jailbreak 🔴	model theft 🟢	data leakage 🔴	hallucination 🔴🔵	adversarial examples 🟢🔴	LLM agent misbehavior 🔴	hardware fault/side channels 🟢🔴
		"blue team"	jailbreak protection 🟢🔴	watermark 🔵	moderation layer	system-level obfuscation 🟢🔵	compiler-level protection 🟢	agent planning smoothing 🔴	trusted computing 🟢🔴
Domain LLM	Online Deployment	**"red team"**	RAG risk assessment 🔵🔴	financial scenario risk 🔵🔴	manufacturing scenario risk 🔵🔴	software dev risk 🟢🔴	Insecure plugin design (market)	medical scenario risk	cybersecurity scenario risk
		"blue team"	RAG protection 🔵🔴	model customization 🔵🔴	decision OTF repairing 🔵🔴	dev assist enhancing 🔴🟢	causal-based prompt opt. 🟢🔴		
Human Alignment	Social Impact		ethical suggestion 🔴🔵	political sensitivity	criminality	physical/mental health	discrimination	security/privacy awareness	bias in "LLM as a judger" 🔵

Case Study – Red Teaming



Case Study – Red Teaming



Domain Usage

Industry

Finance

…

Core Model

Foundation LLM Service

Infra.

Low-level code

Framework

Compiler/ runtime

OS

DB

Trusted Env.

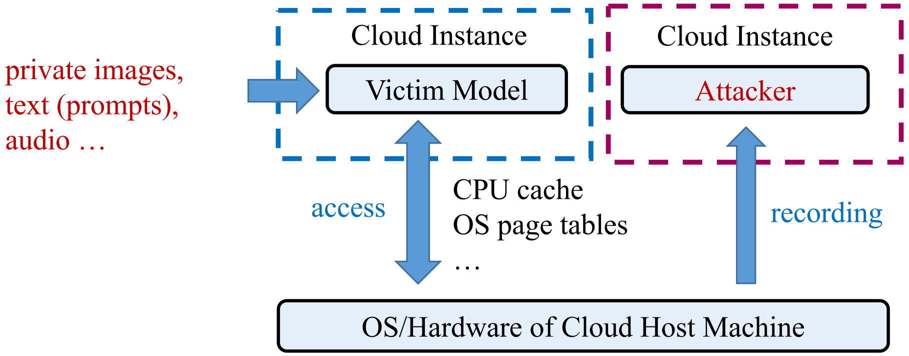
HW

CPU

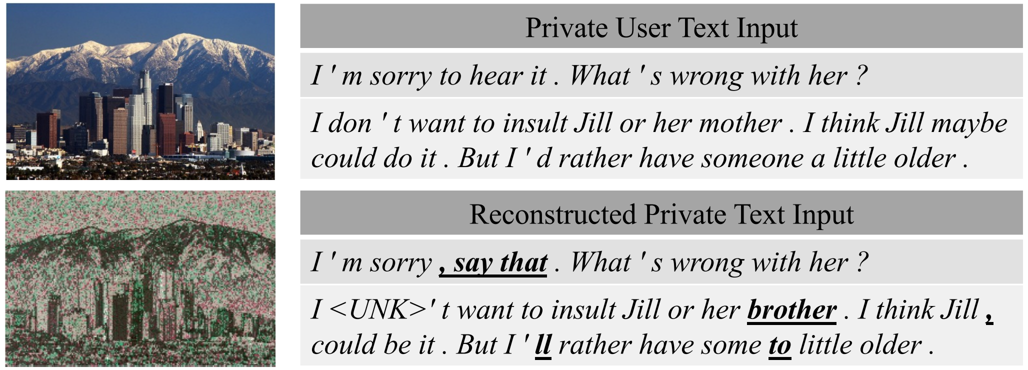
GPU

DRAM

…



(a) SCA toward cloud AI platforms to recover private images.



Private User Text Input

I ' m sorry to hear it . What ' s wrong with her ?

I don ' t want to insult Jill or her mother . I think Jill maybe could do it . But I ' d rather have someone a little older .

Reconstructed Private Text Input

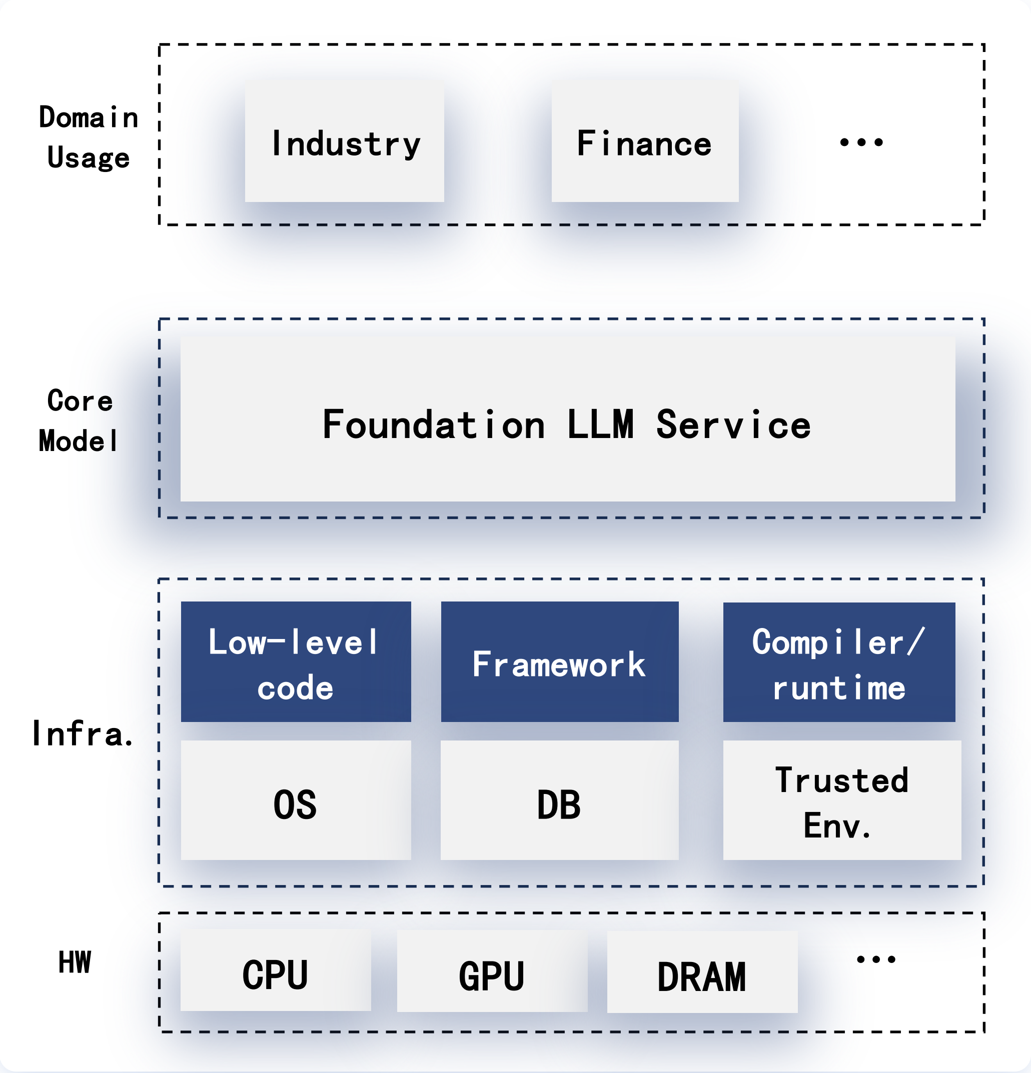
*I ' m sorry **, say that** . What ' s wrong with her ?*

*I <UNK> ' t want to insult Jill or her **brother** . I think Jill **,** could be it . But I ' **ll** rather have some **to** little older .*

(b) Private input images and text vs. the reconstructed inputs via side channel analysis.

User private inputs can be leaked via side channels.

Case Study – Red Teaming



Domain Usage

Industry

Finance

…

Core Model

Foundation LLM Service

Infra.

Low-level code

Framework

Compiler/ runtime

OS

DB

Trusted Env.

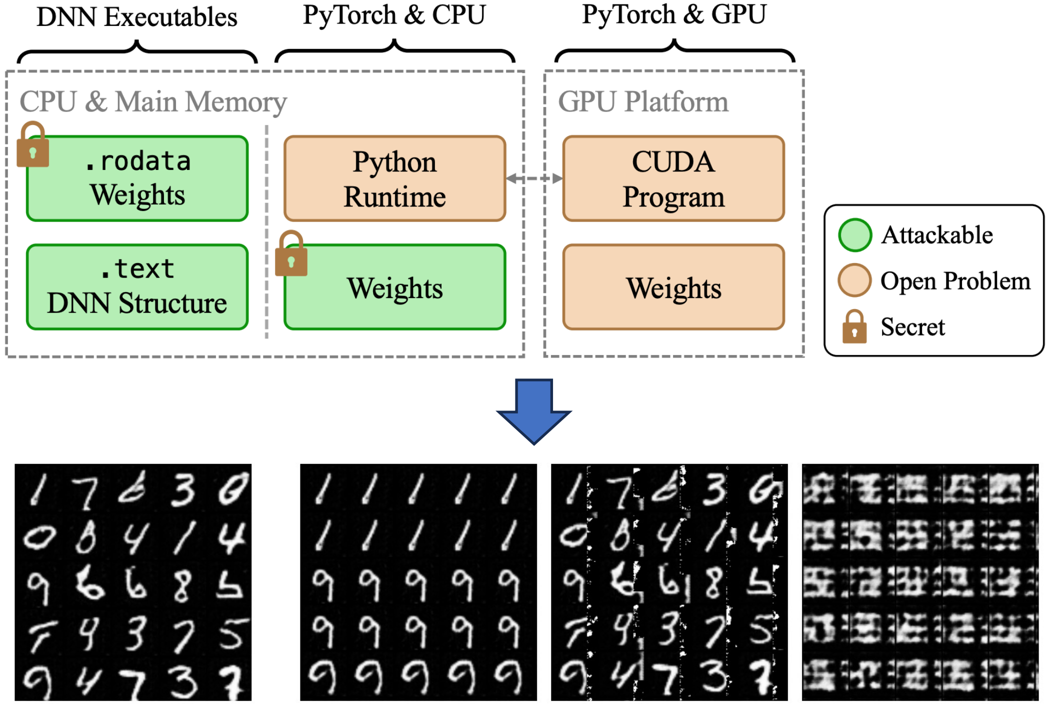
HW

CPU

GPU

DRAM

…

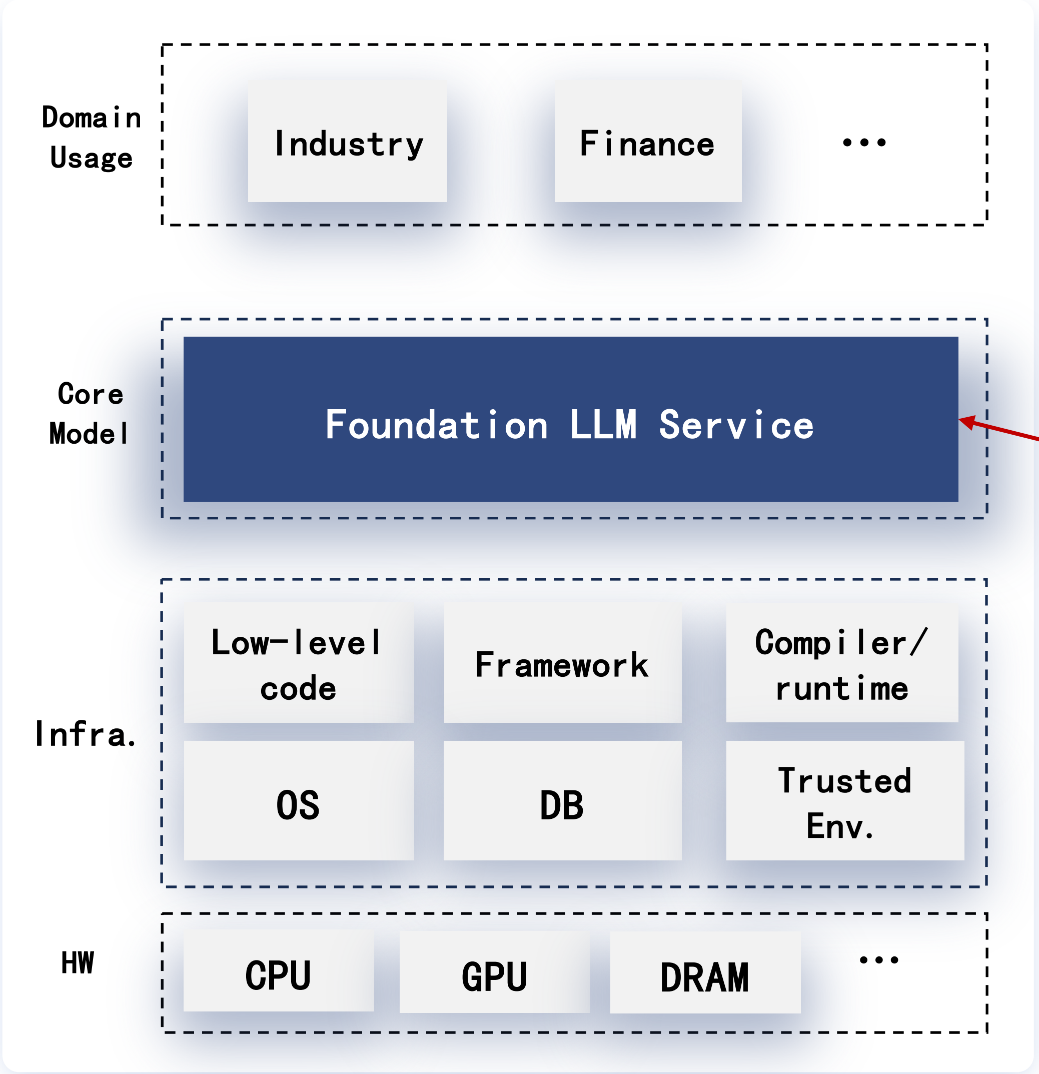


(a) Before BFA

(b) 3 Different types of outcomes after BFA

Using BFA (“bit-flip attack”), we can control model behavior.

Case Study – Red Teaming



Domain Usage

Industry

Finance

…

Core Model

Foundation LLM Service

Infra.

Low-level code

Framework

Compiler/ runtime

OS

DB

Trusted Env.

HW

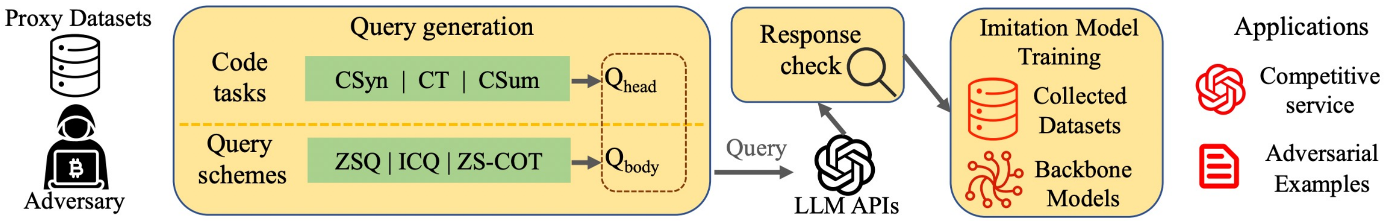
CPU

GPU

DRAM

…

In turn, attack your LLM easily!

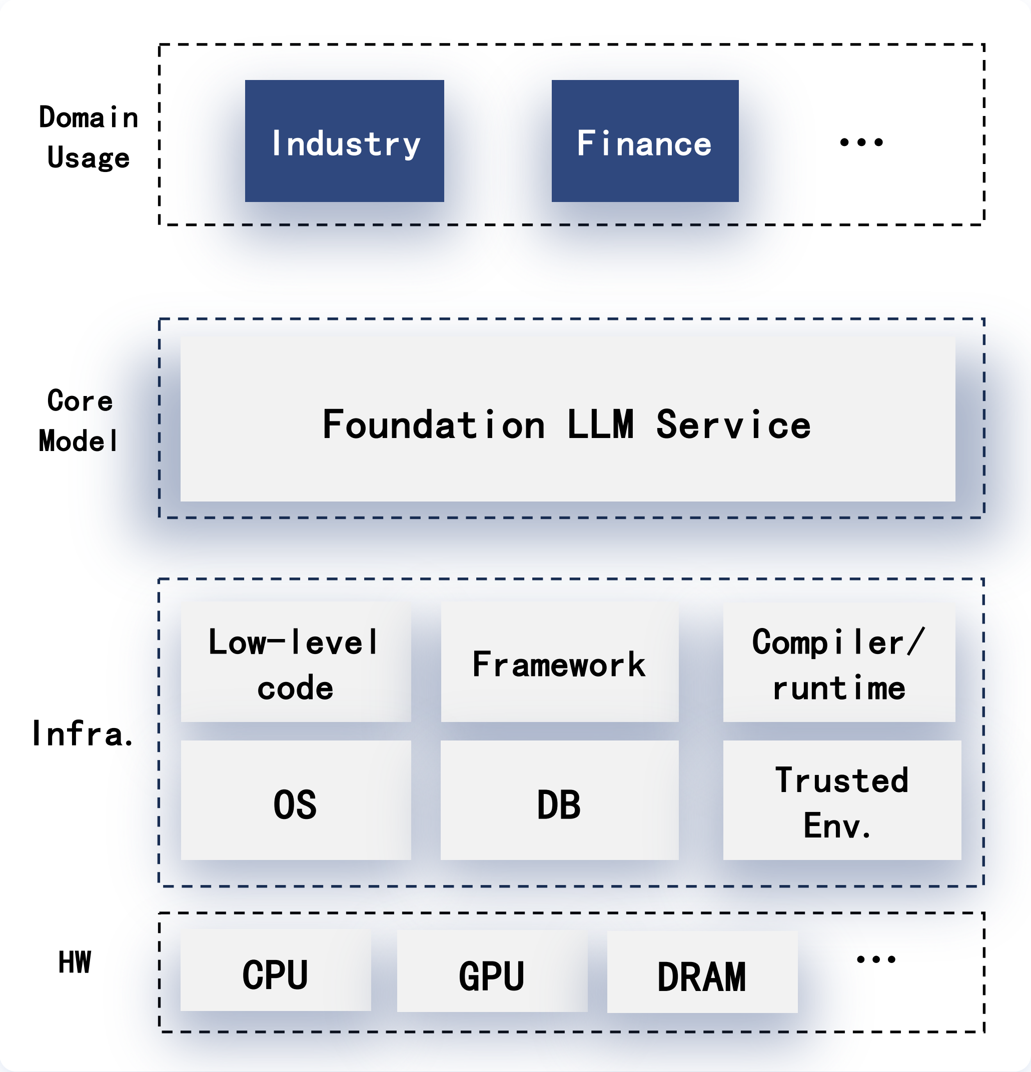


Imitation attack can ”extract” your LLM’s knowledge

Attacker’s local model can be easily trained with only a few thousand USD

Our advocation --- LLM usable, but not “stealable”

Case Study – Red Teaming

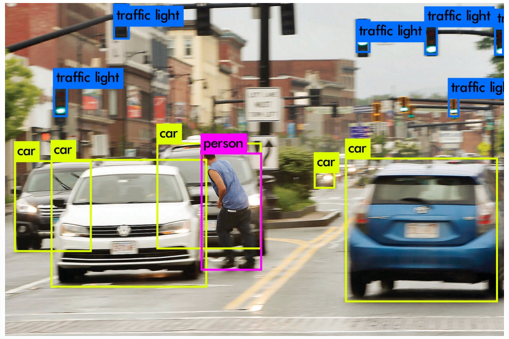


Domain Usage: Industry, Finance, …

Core Model: Foundation LLM Service

Infra.: Low-level code, Framework, Compiler/ runtime, OS, DB, Trusted Env.

HW: CPU, GPU, DRAM, …



GitHub Copilot

OpenAI

With well-designed methods, we find many defects in varying domains/applications.

CAII⁺ 中国工业互联网研究院 China Academy of Industrial Internet

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

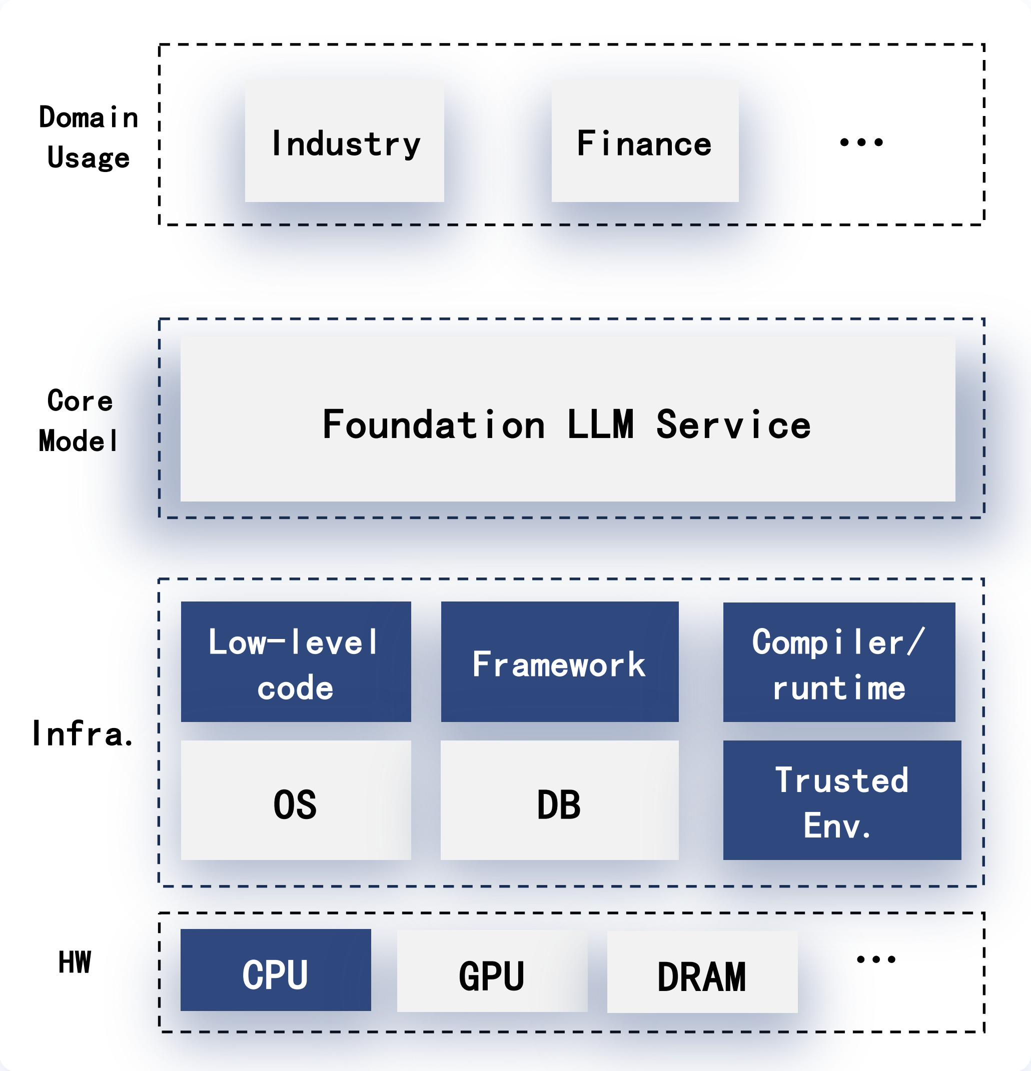
中国经济信息社 CHINA ECONOMIC INFORMATION SERVICE

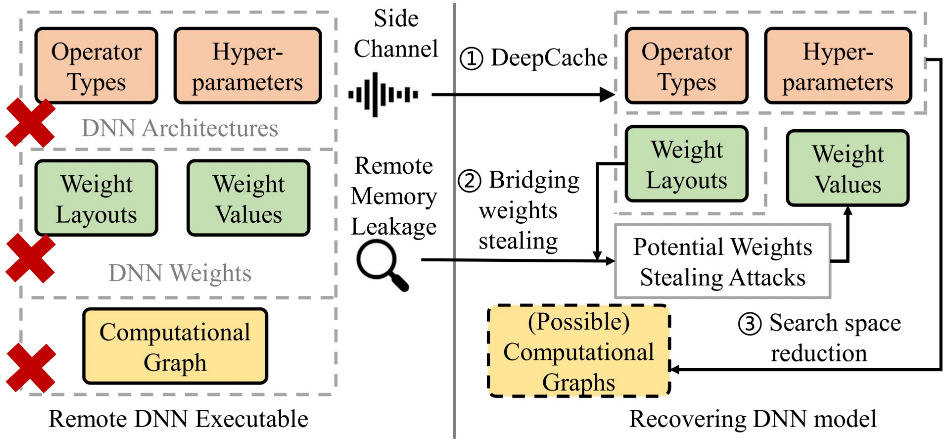
人工智能大模型
工业应用准确性测评

2024年3月版

LLM dependability in industrial usages

Case Study – Blue Teaming





Holistic attack pipeline

We harden model infrastructures using

- Obfuscations
- Sanitizations
- Oblivious RAM (ORAM)
- ...

Our solutions are applicable for various scenarios.

Building Trustworthy LLM

Dr. Shuai WANG
Assistant Professor @ CSE
shuaiw@cse.ust.hk
https://home.cse.ust.hk/~shuaiw/