

大模型安全 威胁框架

Ver1.0

ChatGPT GPT-4 Claude DeepSeek (Large Language
Models, LLMs) AI

“ ”

AI





CPU GPU

CPU

CPU

GPU

GPU

GPU

GPU

Rowhammer



TensorFlow PyTorch Numpy

●

API

API

Ollama

“

”

“

”

APP

DeepSeek APP

APP

/

CI () / CD ()

;

; RAG

()

(API)

Ollama 0.0.0.0 IP 11434

(

)

“

”

2.2

:

Q ÒË@
E



●

“ ”

2.3

AI

●

●

● /

2.4

AI

AI

AI

Web

(

XSS)

AI

(

AI

)

AI

AI

AI

AI



Web

Web

API

XSS

SQL

CSRF

API

- API

API (

●

AI

AI-R-IAM

AI

AI Identity

Non-Human Identity, NHI

RAG

IAM

RAG

API

API

(

)

IAM

- 1 AI 安全白皮书：AI-R-IAM . 2025 2 .
- 2 AI 安全白皮书：Demon AI Security Handbook 2025 5 .
- 3 DeepSeek 安全白皮书 . 2025 4 .
- 4 OWASP Top 10: LLM & Generative AI Security Risks <https://genai.owasp.org>.

LLMs

AI

1.

/

AI

AI

- NIST (AI RMF NIST AI Risk Management Framework)

AI RMF 1.0

NIST 2023

1

AI

AI

NIST AI RMF

- (EU AI ACT)

EU AI ACT

2024

5 8 2025 2

AI AI

AI AI

- ISO/IEC 42001 ()

ISO IEC AI

AIMS

AI AI

- IEEE AI (IEEE AI Ethics Guidelines)

IEEE AI

- /

“ ”

“ ”

2023 8

1.0 2024 9

4

2.

AI

● OWASP Top 10 for LLM Applications 2025

Web

OWASP

LLM

LLM

● MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)

ATLAS

MITRE

AI

AI

MITRE ATT&CK

AI

TTPs

● MIT AI Risk Repository

MIT AI

3.

● AI

AI

AI

AI

AI

AI

AI

AI

AI

AI

AI

AI

AI

● Google AI

SAIF Security AI Framework

Google 2023

AI

AI

AI

AI

Google DeepMind

FSF Frontier Safety Framework

DeepMind

AI

SAIF

FSF

● OpenAI AI

OpenAI LLM

AI

“ Preparedness Framework ”

AI