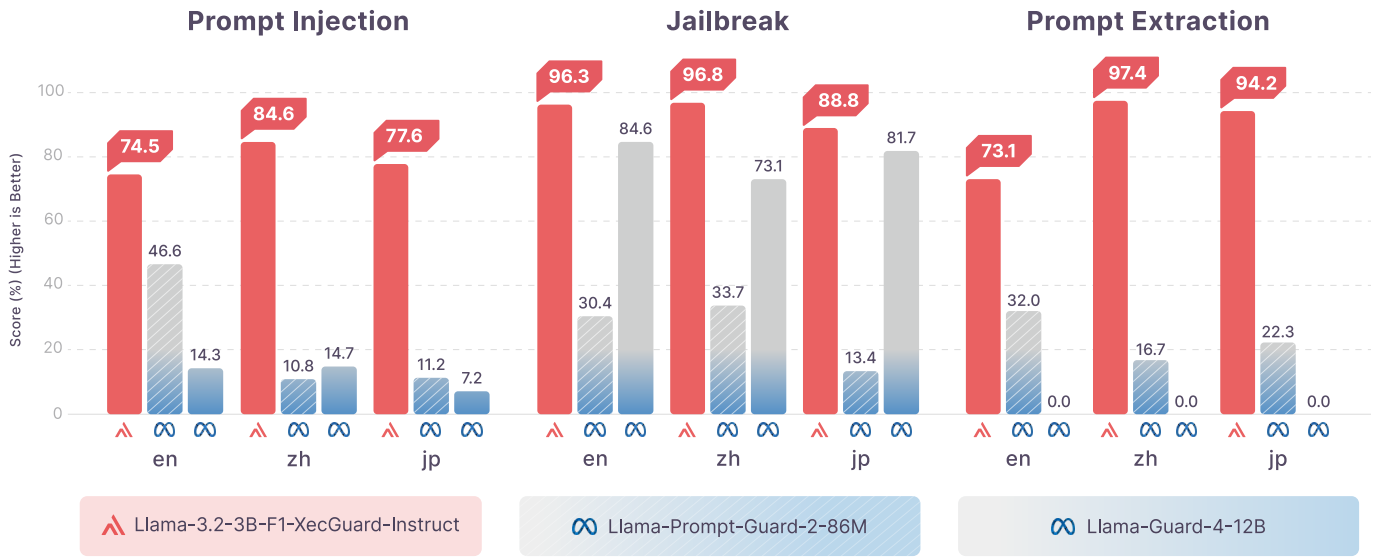
CYCRAFT

CyCraft Guards Trustworthy LLMs

XecGuard

AI Blue Teaming
Next-Generation AI Firewall

Outstanding Performance:
XecGuard's Industry-Leading Security Protection



Reinforce System Prompt Compliance in Normal Contexts
Reinforce Malicious User Prompt Detection

Prompt Injection Defense
Detect malicious instructions, output manipulation, instruction obfuscation, logical/emotional appeals, and role-playing attacks, preventing deviations from application-defined System Prompts to ensure compliance.

Prompt Extraction Defense
Prevent attackers from guiding inputs to exfiltrate application System Prompts, safeguarding internal logic and confidential information.

Jailbreak Attack Defense
Block safeguard-bypassing context manipulation, preventing outputs that violate ethical norms or security standards to preserve AI model's safety and reliability.

CYCRAFT

CyCraft Builds Trustworthy LLMs

XecART

AI Red Teaming Security Assessment

Chatbot Security Assessment

CyCraft AI RedTeam

API
Online Testing

Customized LLMs (Finetuned)
Chatbot System Prompt

Prompt Security Rapid Testing

CyCraft AI RedTeam

Offline Testing

Common LLMs
Chatbot System Prompt

AI Safety Resilience Report

AI Safety Compliance Mapping

- **Prompt Instruction Violation Testing**
  - Prompt Injection
  - Indirect Prompt Injection
  - Sensitive Data Leak
- **Prompt Leakage Testing**
  - Prompt Disclosure
- **Model Bias and Hallucination Testing**
  - Content Bias
  - Hallucinations
  - Input Leakage
- **Public Moral or Ethical Standard Violation Testing**
  - Unsafe Outputs
  - Toxic Outputs