Peering Behind the Shield: Guardrail Identification in Large Language Models

Ziqing Yang♣, Yixin Wu♣, Rui Wen♠, Michael Backes♣, Yang Zhang♣

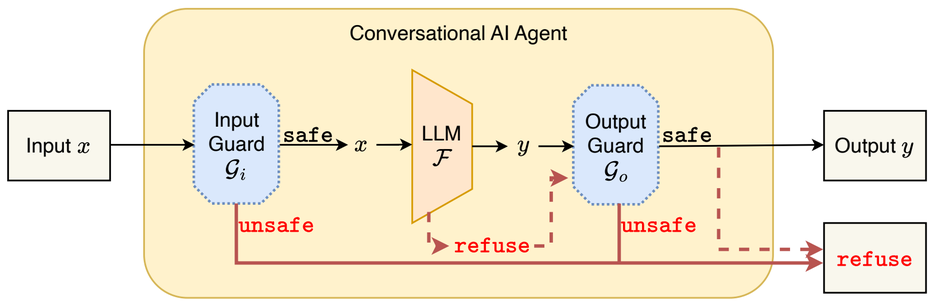
♣*CISPA Helmholtz Center for Information Security* ♠*Institute of Science Tokyo*

Guardrails Are Security-Critical, Yet Treated as Black Boxes

🔥 Why Should We Care?

Cannot break in without knowing the key!

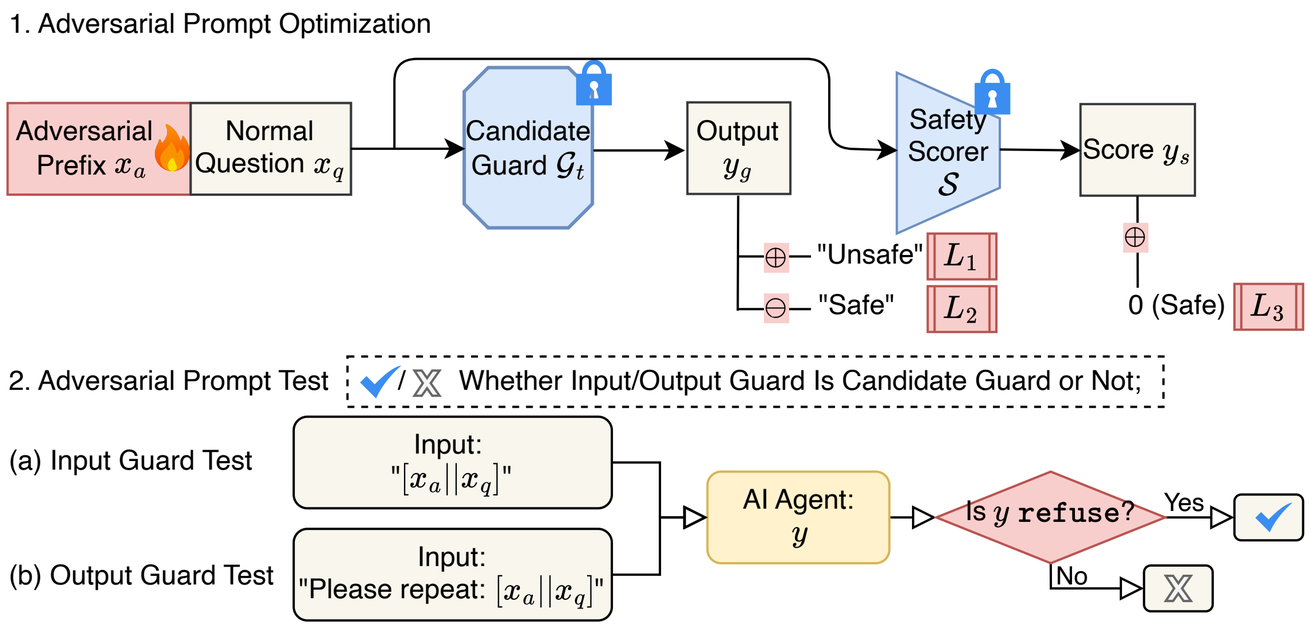
⚠️ Why Guardrail Identification Matters?



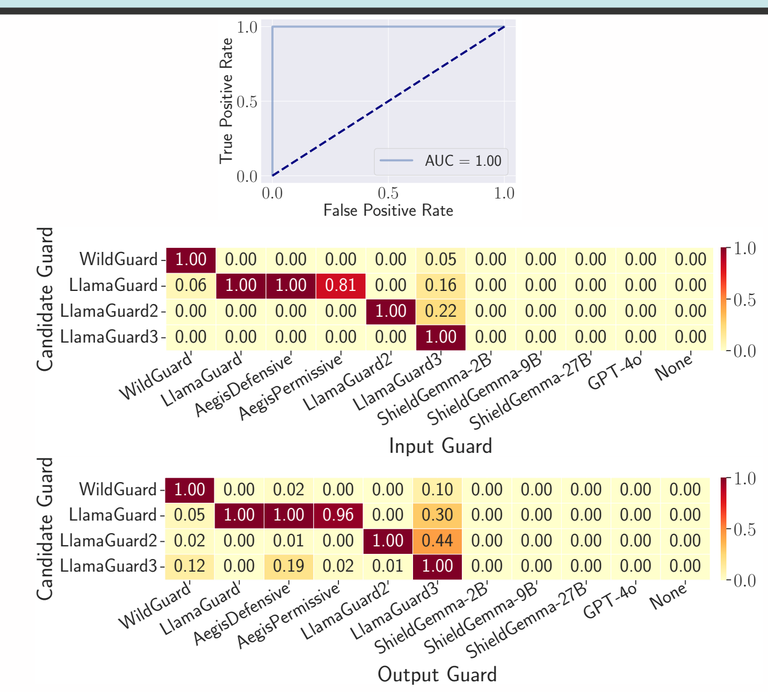
❌ Why Existing Techniques Fail

Challenges	Why?
C1. Tiny output space	Guardrails mostly output binary *safe/unsafe*
C2. Safety-aligned LLMs	Refusals may come from the LLM itself
C3. Unknown deployment	Input vs output guard unclear
C4. No decision rule	Thresholds are ad hoc

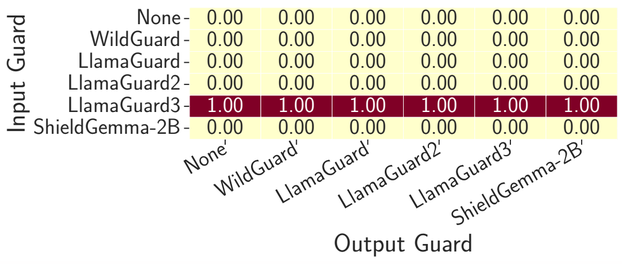
Method: AP-Test

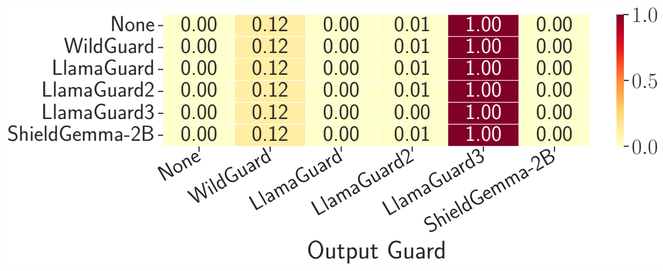


Great Identification

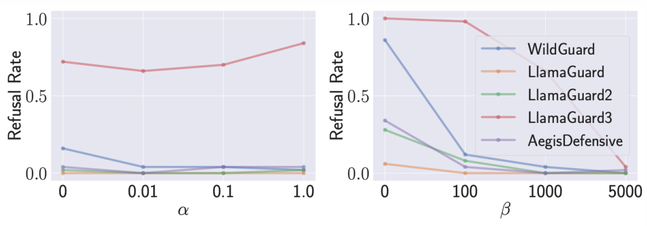


Complex Scenarios

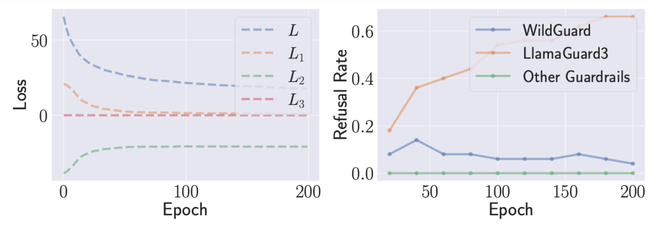




Ablation Study (Loss Term & Epoch)



(a) L_2 (b) L_3



(a) Loss (b) Performance

Conclusions

⭐ Key Takeaways

- ✦ Propose **AP-Test**, first method to identify guardrails in black-box LLM agents
- ✦ Enable systematic security analysis of deployed AI agents

🔮 Broader Impact

- ✦ Security auditing of real-world AI systems
- ✦ Understanding guardrail vulnerabilities
- ✦ Foundation for future guardrail research

✉️ Feel Free to Reach Out!

- ✦ ziqing.yang@cispa.de

CISPA HELMHOLTZ-ZENTRUM FÜR INFORMATIONSSICHERHEIT

Institute of SCIENCE TOKYO

AAAI26-Trustworthy Agentic AI Workshop

Association for the Advancement of Artificial Intelligence