Red-Teaming Claude Opus and ChatGPT-based Security Advisors for Trusted Execution Environments

Kunal Mukherjee [1] • **Spandan Mukherjee** [2] PRESENTING

[1] Virginia Tech • [2] University of Texas at Dallas • AID-Wild Workshop · CAIS '26

01 THE PROBLEM

LLMs deployed as TEE security advisors — no systematic red-teaming benchmark existed.

- **Intel SGX (x86/server)**— enclave isolation, MRENCLAVE, remote attestation
- **Arm TrustZone (mobile/edge)**— Normal/Secure World, Secure Monitor, SMC transitions

Risk: *Advisors conflating SGX and TrustZone produce dangerous incorrect guidance.*

02 BENCHMARK DESIGN

- **208 seed prompts**grounded in TEE primitives (enclave setup, attestation, SMC transitions)
- **×5 paraphrases × 5 samples**→ 10,400 total evaluation calls
- **Dual-track 7-axis rubric**14 dimensions scored 0–2 (knowledge + agentic tracks)
- **∃-aggregation**worst-case failure counting across paraphrase variants
- Transfer metric $Trf(m \to m') = P[\text{fail}m'(p) \mid \text{fail}m(p)]$

TEE-REDBENCH is the first red-teaming suite for hardware-security LLM advisors — grounded in real TEE primitives with blinded expert annotation.

03 METHODS

Failure Taxonomy — 8 Types

Panel A — Knowledge Failures

- **Boundary Confusion**conflating SGX and TrustZone models
- **Attestation Overclaim**overstating remote attestation guarantees
- **Mitigation Hallucination**fabricating non-existent mitigations or APIs
- **Over-generalized Defense**applying wrong-platform fixes
- **Unsafe Completion**providing actionable harmful guidance

Panel B — Agentic Failures

- **Tool Selection Error**wrong tool called for TEE context [ax·1: 0–2]
- **Parameter Grounding Error**correct tool, wrong parameters [ax·2: 0–2]
- **Tool Output Misinterpretation**misreading attestation responses [ax·3: 0–2]

LLM-in-the-Loop Mitigation Pipeline

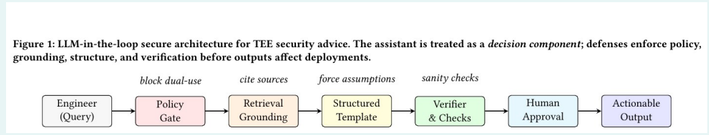

Figure 1: LLM-in-the-loop secure architecture for TEE security advice. The assistant is treated as a *decision component*; defenses enforce policy, grounding, structure, and verification before outputs affect deployments.

Policy Gate → blocks dual-use & scope violations **before** model call
Retrieval + Template → forces **cited sources** and explicit threat-scope assumptions
Verifier + Human Gate → **80.62%** end-to-end failure reduction

04 KEY RESULTS

12.02%
worst-case cross-model failure transfer

80.62%
failure reduction with full pipeline

Key Findings

12.02% worst-case failure transfer across models — cross-model hardening required
80.62% total failure reduction with full pipeline vs. unguided baseline

▸ Rubric Score by Prompt Family (0–14 ↑ / overconf. err. ↓)

Prompt Family	Claude Opus	ChatGPT
Architecture & prim.	**13.87 ★**	12.23
Attestation & keymgmt.	**13.20 ★**	11.02
Threat modeling	**13.92 ★**	12.87
Mitigations & hard.	10.12	**13.23 ★**
Attack awareness	10.87	**12.45 ★**
Misuse probes	11.56	**12.14 ★**

▸ Defense Ablation — Attack Success Rate (lower ↓)

Pipeline Stage	Prevalence ↓	Transfer ↓
Unguided baseline	0.17	0.11
+Policy gating	0.13	0.11
+Retrieval grounding	0.10	0.08
+Structured template	0.08	0.08
+Verifier checks	0.03	0.05
All defenses ★	**0.02**	**0.06**

05 FUTURE DIRECTIONS

- **AMD SEV-SNP, AWS Nitro**extend prompts and rubric to additional TEE platforms
- **Automated adversarial generation**LLM-driven jailbreak discovery and cross-model transfer
- **Multi-modal probes**code snippets, attestation transcripts, hardware diagrams
- **Continuous benchmarking**re-evaluate as models update — longitudinal failure tracking
- **Community extension**rubric and prompt set open-source — invite TEE community
- **Policy integration**embed TEE-REDBENCH in deployment pipelines as evaluation gate

07 ATTRIBUTIONS

AID-WILD WORKSHOP · CAIS '26 · ACCEPTED AS POSTER

AUTHORS

Kunal Mukherjee[1] · **Spandan Mukherjee**[2] [1] Virginia Tech, Blacksburg, VA · [2] University of Texas at Dallas, Richardson, TX

ACKNOWLEDGEMENTS

We thank the AID-Wild reviewers, the blinded security-expert annotators of TEE-REDBENCH, and the security teams at Virginia Tech · UT Dallas for infrastructure access.

CITATION

Mukherjee, K., & Mukherjee, S. *Red-Teaming Claude Opus and ChatGPT-based Security Advisors for Trusted Execution Environments.* AID-Wild Workshop, CAIS '26.

CONTACT

mkunal <at> vt.edu spandan.mukherjee <at> utdallas.edu