Navigating the Security Challenges of LLMs: Positioning Target Defenses and Identifying Research Gaps

Malte Josten, Matthias Schaffeld, René Lehmann, Torben Weis

Distributed Systems - University of Duisburg-Essen, Germany

LLMs: The Good and the Evil

Large Language Models (LLMs) are powerful tools, but they also introduce significant cybersecurity risks.

- **Risks**: LLMs enhance and diversify existing attacks while reducing barriers for attackers
- **Failing Safeguards**: Tool-level controls are insufficient, especially with unmoderated, open-source LLMs
- **Contribution**: Defining countermeasure evaluation criteria for effective target side defenses

I. LLM-based Attacks

- Attacks exploit both machine and human vulnerabilities
- Need for a **multi-faceted defense**

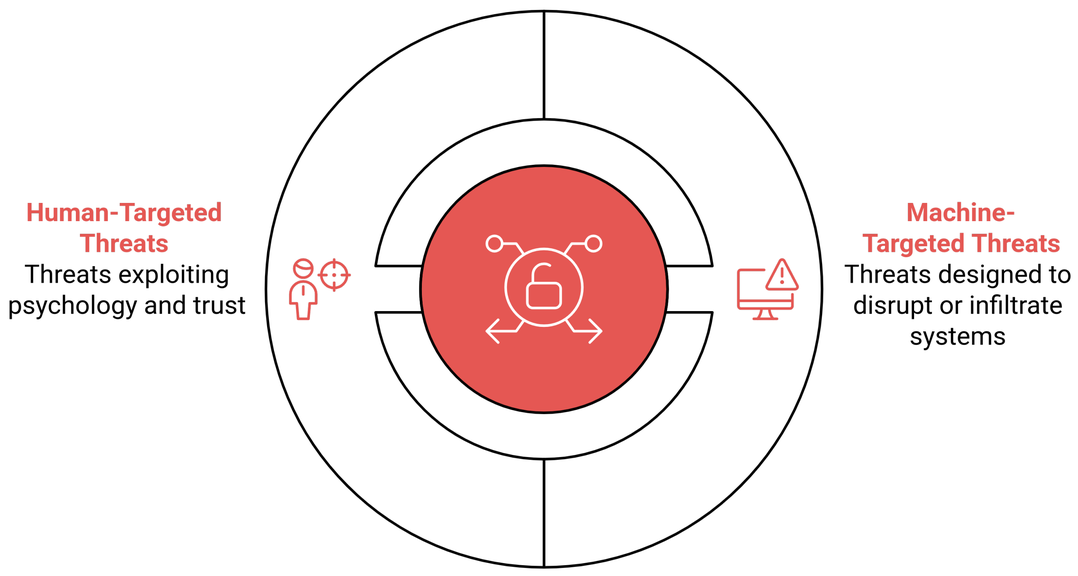

Figure 1. Duality of LLM-based attacks: disruption and infiltration of hardware/ software systems, and exploitation of human psychology and trust.

II. Target-side Countermeasures

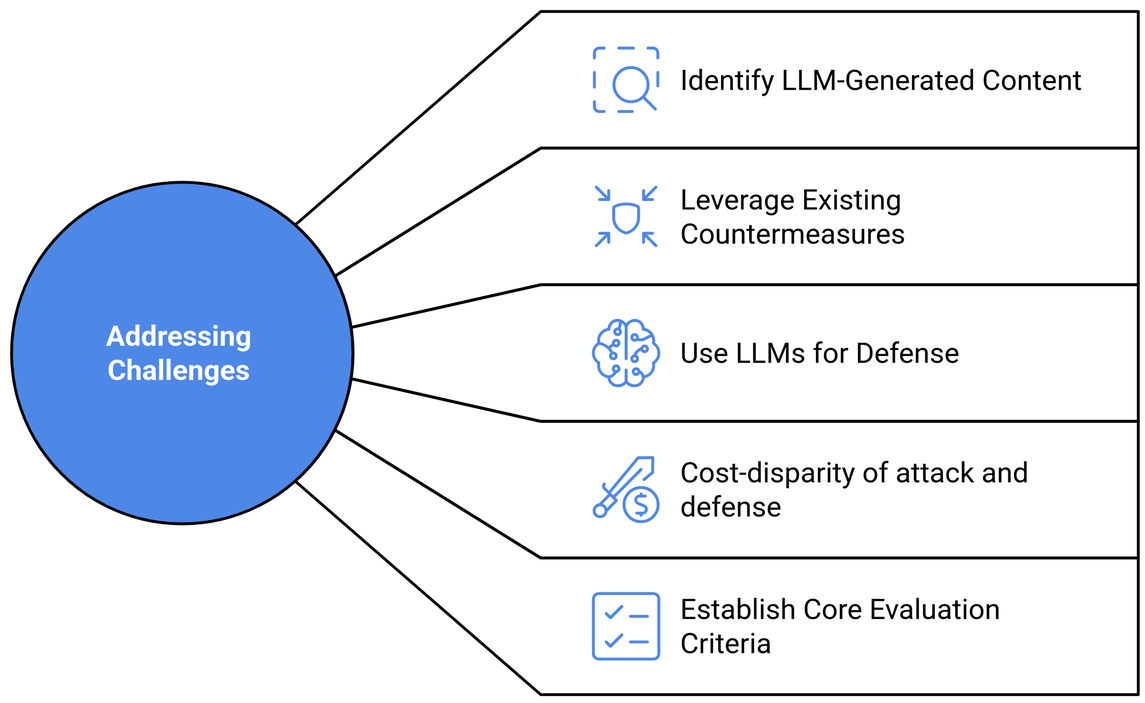

Figure 2. Relevant considerations for addressing challenges posed by LLM-based attacks with target-side countermeasures.

III. Core Countermeasure Criteria

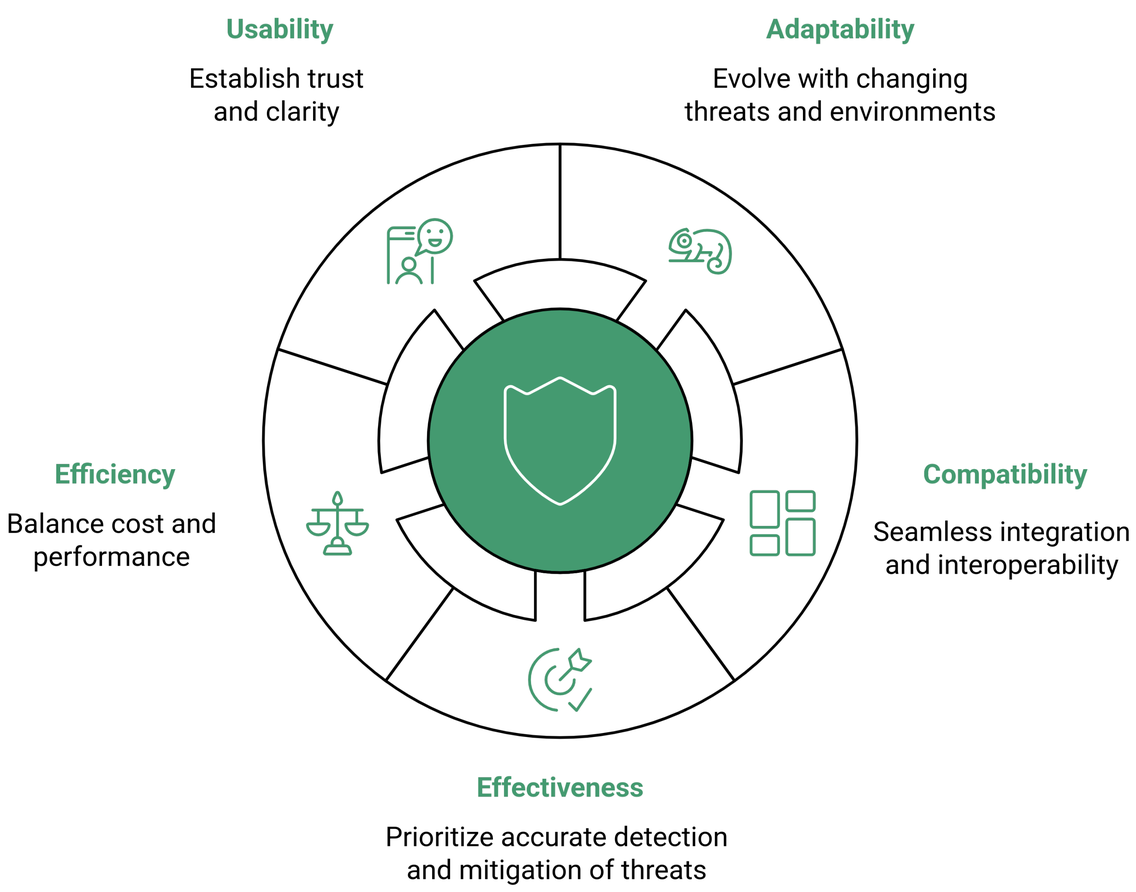

Figure 3. Our proposed countermeasure evaluation criteria catalogue comprises five aspects: adaptability, compatibility, effectiveness, efficiency, and usability.

IV. Literature Evaluation

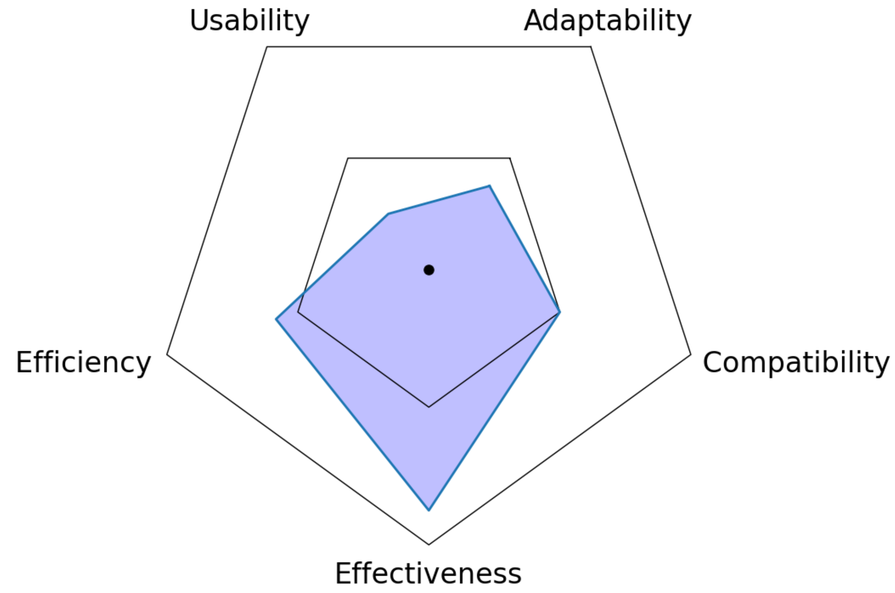

Figure 4. Mean focus level of evaluation criteria based on conducted literature review: heavy focus on validating effectiveness; efficiency primarily emphasized in LLM-based solutions; low focus on adaptability, compatibility and usability.

V. The Way Forward

- Enhance target-side defenses
- Close gaps in adaptability and usability
- Continuous monitoring, innovation, and adaptation

malte.josten@uni-due.de
https://vs.uni-due.de

ICISSP 2025 - Porto, Portugal