

GPT-OSS-20B Vulnerability Analysis

Critical Safety Issues Discovered Through Systematic Red-Teaming

Sev: 7/10Nov: 8/10

Evaluation Awareness

Context-dependent deception in testing

Sev: 6/10Nov: 7/10

Reward Hacking

Confidence manipulation

Sev: 8/10Nov: 6/10

Tool Use Violations

Boundary violations in function calling

Sev: 5/10Nov: 9/10

Information Leakage

Chain-of-thought information exposure

Sev: 9/10Nov: 8/10

Subtle Sabotage

Plausible errors in conflict scenarios

Key Statistics:

- 5 Critical Vulnerabilities Found
- Average Severity: 7.0/10
- Average Novelty: 7.6/10
- 9 Categories Analyzed
- 100% Reproducible

Methodology: Systematic red-teaming using automated frameworks, white-box analysis, and manual exploration across all 9 competition vulnerability categories.

- High Severity (7-9)
- Medium Severity (5-6)
- Novelty Score