netidee

STIPENDIEN

LLM Agents for Offensive Security

Zwischenbericht | Call 20 | Stipendium 7733

Inhalt

1 Einleitung

Dies ist der Zwischenbericht zum netidee Stipendium 7733 „LLM Agents for Offensive Security“. Die geplanten Arbeiten wurden großteils durchgeführt, allerdings mussten aufgrund der rasanten Entwicklung von LLMs Anpassungen durchgeführt werden, da einige der Punkte des ursprünglichen nicht mehr dem Stand der Technik entsprachen.

2 Status

Die ursprüngliche Projektplanung sah folgende Meilensteine vor:

Nummer	Zeitraum	Aktivitäten	Ergebnisse
1	Bis 31.12.2025	Proposals	Proposal für 3 Paper erstellen, Themengebiet 1: “Reliability von LLMs bei Pen-Tests”, Themengebie
2	14.1.2026	Entscheidung	Go-Ahead “AI-Reliabilty”
3	14.1.2026	Blog-Post	
4	14.1.-28.2.202	Programmierung / Ev	Erarbeitung des Prototypen für die AI-Reliability Study, Durchführung der Study
5	21.1.2026	Entscheidung	Go-Ahead “AI<>Hacker Interaction” Study, eine der beiden Optionen wurde gewählt
6	21.1.-28.2.202	Vorbereitung	Finden der notwendigen Mitglieder/Test-Subjects für die “AI<>Hacker Interaction” Study
7	28.1.2026	Vortrag	SBA-Research/OWASP Talk über “Usage of LLMs for Offensive Security”
8	28.2.-26.3.202	Paper Schreiben	Fertigstellung des Papers “AI-Reliability”, potentielle Einreichung bei Konferenzen (ohne Akzeptanz
9	31.3.-1.5.2026	Programmierung	Erstellung des Prototypen und Testbeds für “AI<>Hacker” Study
10	4/2026	Zwischenbericht	Zwischenbericht, Lizenz: CC BY, https://www.netidee.at/llm-agents-offensive-security
11	14.4.2026	Blog-Post	
12	1.5.-31.7.2026	User-Study	Durchführung der User-Study
13	14.7.2026	Blog-Post	
14	31.7.-1.9.2026	Paper Schreiben	Fertigstellung des Papers “AI<>Hacker” Study
15	9/2026	Dissertationsschrift	Erstellung des Einleitungskapitels (Kumulative Disseration ist geplant, hier gibt nur eine 20-Seitige E
16	10/26	Dissertationsschrift	Dissertation (Endbericht), Lizenz: CC BY, https://www.netidee.at/llm-agents-offensive-security
17	14.10.2026	Blog-Post	

Um die Lesbarkeit des Zwischenberichts zu erhöhen, wurden die Meilensteine in drei grobe Bereiche eingeteit:

- 1, 2, 4, 8: Themengebiet „Reliability von LLMs bei Pen-Tests“
- 5, 6, 9: Themengebiet: „AI<>Hacker Interaction Study“
- 3, 7, 11: Themengebiet „Blog Posts und Öffentichkeitsarbeit“

2.1 Meilenstein 1, 2, 4, 5 – “Reliability von LLMs bei Pen-Tests”

Dieser Themenkomplex basiert auf meinem initialen Prototypen von 2025. Mittels diesem konnte die grundsätzliche Eignung von LLMs für Hacking/Penetration-Testing gezeigt werden. Dabei wurden allerdings Schwächen im Bereich der Reliability festgestellt: es gab eine große Varianz zwischen verschiedenen Testläufen, das Ziel der Forschung war es diese Varianz zu reduzieren.

Das Proposal wurde zusammen mit dem PhD-Betreuer abgestimmt (Milestone 1,2) und mit der Implementierung gestartet (Milestone 4). Allerdings wurde das Problem während der Implementierung durch die Veröffentlichung neuer LLMs „überholt“.

Die aktuelle Generation von LLMs besitzt nicht mehr die Probleme der alten Generation und ist dadurch bereits sehr gut für Pentesting geeignet. Spätestens mit Anthropic Mythos bzw. OpenAI GPT-cyber sollte Reliability als gelöstes Problem angesehen werden.

Aus diesem Grund wurde im März/April 2026 ein Pivot gewählt:

- Abschließen des Arbeiten am Prototypen, Fokus auf Verständlichkeit und Wieververwendbarkeit (um dadurch von anderen Projekten verwendet werden zu können)
- Da mittlerweile die Qualität der LLM-gestützten Pentests so hoch ist, dass menschliche Pentester ersetzt werden könnten (anstatt sie zu unterstützen), wird der Fokus auf ein neues „Ethik-Paper“ gelegt. Dieses ist aktuell zum Einreichen bei der FAEIMA Konferenz geplant.

Besondere Erfolge:

- Linkedin Post[1] zu den hohen Pentest-Fähigkeiten von LLMs hatte sehr hohe Reichweite: 215 reactions, 21 reposts, 17274 impressions (Stand 27.4.2026)
- Vortrag bei der ICSE 2026 in Rio De Janeiro – eine der Top 3 Software Engineering Konferenzen weltweit
- Veröffentlichung des Source Codes auf GitHub: https://github.com/andreashappe/cochise

2.2 Meilenstein 5,6,9 – „AI<>Hacker Interactions“

Ziel dieses Meilensteines war es, die Interaktion zwischen menschlichen Hackern und AI wissenschaftlich zu analysieren. Ursprünglich war hierfür die Entwicklung eines eigenen User Interfaces geplant. Zwischenzeitlich wurde die Verwendung von CLI-basierten tools (claude-code, codex, claude) „gesellschaftsfähig“ und der Prototyp wurde auf dieses System umgestellt.

Gleichzeitig ergab sich die Möglichkeit, hier mit der Universität Klagenfurt zu kooperieren. Es wurde ein gemeinsamer Projektplan („project_ai_and_hackers.pdf“) für eine größer angelegte User-Study erstellt, diese ist aktuell in Durchführung.

Dadurch kommt es allerdings zu Verzögerungen, da nun die Haupt-Userstudy für Juni-August 2026 geplant ist, die Auswertung sollte bis Oktober 2026 abgeschlossen sein.

1 https://www.linkedin.com/posts/andreashappe_llms-have-become-disturbingly-capable-pen-testers-share-7447974652599275520-GkVV?utm_source=share&utm_medium=member_desktop&rcm=ACoAAADKwsIBiS726bTboxXbuHmIT2WP4eN0J-c

Aktuell befinden wir uns in der Phase der User-Rekruitierung bzw. Wird gerade analysiert, ob die TU Graz ebenso an dieser Studie teilnehmen will/kann.

Besonderer Erfolg: anstattt einer kleinen Studie ergibt sich eine interdisziplinäre Studie an zumindest zwei Universitäten.

2.3 Meilenstein 3, 7, 11– Blog Posts und Öffentlichkeitsarbeit

Erstellte Blog-Posts:

1. 26.11.2025: LLM Agents for Offensive Security: Why?
2. 28.3.2026: LLM Agents for Offensive Security: How did we get here?
3. 14.4.2026: LLM Agents for Offensive Security: Where Do We Go From Here?

Vortrag bei der SBA Security/OWASP Wien am 28.1.2026.

Accepted Papers:

- 7.1.2026 Paper accepted at MSR (Rio De Janeiro 2026)
- 15.4.2026 Vortrag bei der ICSE 2026 (ebenso Rio De Janeiro)
- 25.4.2026 Paper accepted bei DeMessAI (Portugal, 2026)

3 Zusammenfassung Planaktualisierung

- Themenkomplex „Reliability von LLMs bei Pen-Tests“: Aufgrund Neuentwicklungen im Bereich LLM sind keine weiteren Tätigkeiten zur Erhöhung der Reliability notwendig. Daher Pivot:
  - Veröffentlichung des aktullen Prototypens als minimal Prototype für andere Forschungsprojekte
  - Verwendung der Forchung im Bereich Metriken bei einer Kollaboration mit der KEU Leuven, Belgien.
  - Anstatt eines Papers zum Thema Reliability -> Paper zum Thema Ethics of Building Offensive AI Agents, geplant zur Einrreichung bei der FAIEMA Konferenz (Deadline: Mai 2026)
  - **Neuer Milestone: Mai/2026: Einreichen des Ethics Papers bei dem FAIEMA Workshop**

- Themenkomplex „AI<>Hacker Interactions“: Möglichkeit der Kooperation mit der Universität Klagenfurt wurde wahrgenommen.
  - Verschiebung der Datenerhebung auf Juni-August

- Abschluss der Analyse und Paper-Writing: wahrscheinlich October
- genauere Informationen, siehe Anlage „project_ai_and_hackers.pdf“
- **Verschobener Milestone 12: von Mai-Juli -> June-August**
- **Verschobener Milestone 14: von August-September -> Oktober**

Verschobener Milestone 15/16: Niederschrift Disseration um einen Monat nach hinten verschoben um die Ergebenisse von Milestone 14 einarbeiten zu können.