Sicherheitsrisiken und Schutzmechanismen von LLMs: Angriffsmöglichkeiten, Compliance und Qualitätsbewertung

Masterarbeit - Exposé 31. März 2025

Autor: Fabian Berger, B. Sc.
Betreuer: Prof. Dr.-Ing. Michael Tielemann
Prüfer: Prof. Dr.-Ing. Michael Tielemann

1 Motivation

Large Language Models (LLMs) haben in den letzten Jahren eine rasante Entwicklung erfahren. Die Modelle sind in der Lage, komplexe Sprachaufgaben zu lösen und menschenähnliche Texte zu generieren. Ihr Einsatz reicht von der Automatisierung von Kundeninteraktionen über die Unterstützung wissenschaftlicher Forschung bis hin zur Erkennung von Mustern in großen Datenmengen. Unternehmen und Forschungseinrichtungen nutzen LLMs, um ihre Produktivität zu steigern und neue Anwendungsfelder zu erschließen.

Mit der zunehmenden Verbreitung von LLMs wachsen jedoch auch die sicherheitsrelevanten Herausforderungen. Die Sprachmodelle können potenziell anfällig für verschiedene Angriffstechniken sein, sowohl während der Trainingsphase als auch bei der Inferenz. Dies stellt Forschungseinrichtungen und Unternehmen vor die Aufgabe, LLMs hinsichtlich ihrer Sicherheit zu bewerten und geeignete Schutzmaßnahmen zu implementieren.

Trotz dieser Herausforderungen können LLMs richtig eingesetzt Vorteile in sicherheitskritischen Anwendungen bieten. Einige vielversprechende Anwendungsfälle, in denen LLMs innovative Lösungen liefern können, sind *Intrusion Detection*, *Phishing-* und *Spam-Detection*, sowie *Software Vulnerability Detection* [2].

Diese Ansätze zeigen das Potenzial von LLMs nicht nur als Angriffsziel, sondern auch als Werkzeug zur Verbesserung der Cybersicherheit genutzt zu werden [7].

Die Relevanz dieser Arbeit liegt darin, einen fundierten Beitrag zur Sicherheitsbewertung von LLMs zu leisten. Durch die systematische Analyse von Schwachstellen, die Entwicklung geeigneter Schutzmaßnahmen und die Bewertung bestehender Modelle soll diese Forschung dazu beitragen, den sicheren und verantwortungsvollen Einsatz von LLMs zu gewährleisten.

2 Problemstellung

Mit den zunehmenden Fähigkeiten von LLMs wachsen auch die damit verbundenen Sicherheitsrisiken. Wie das Open Web Application Security Project (OWASP) in seinem Bericht

OWASP Top 10 for LLM Applications [3] feststellt, ergeben sich neue Angriffsvektoren, die gezielt ausgenutzt werden können. Dazu gehören Angriffsvektoren wie *Prompt Injection*, *Adversarial Attacks* und *Side-Channel Exploits*, die Datenschutzverletzungen oder den Missbrauch der Modelle begünstigen können. [2, 4]

Ein zentrales Problem ist die Möglichkeit, Sicherheitsmechanismen von LLMs zu umgehen. Insbesondere *Jailbreaking* ist eine weit verbreitete Methode, um mithilfe von bestimmten Eingaben die internen Sicherheitsvorkehrungen des Modells außer Kraft zu setzen [8]. Dadurch können Modelle dazu gebracht werden, ethisch bedenkliche oder sicherheitskritische Inhalte auszugeben, die unter normalen Umständen blockiert wären. Zusätzlich gibt es eine Vielzahl von LLMs, die von vornherein ohne Sicherheitsmaßnahmen zur Verfügung stehen. Diese Modelle können direkt für kriminelle oder unethische Zwecke missbraucht werden, etwa zur Generierung von Phishing-Inhalten. [3]

Neben direkten Angriffen auf Modelle existieren auch weitergehende Bedrohungen auf Systemebene. Side-Channel-Angriffe könnten eine Möglichkeit darstellen, um sensible Informationen aus LLMs zu extrahieren. Dabei werden durch die Analyse von Systemkomponenten private Informationen mit einer höheren Erfolgsrate extrahiert als durch direkte Modellabfragen [8]. Solche Angriffe nutzen beispielsweise Speicherzugriffe oder Laufzeitanalysen, um vertrauliche Daten zu gewinnen.

Ein weiterer wichtiger Aspekt ist die Einhaltung regulatorischer Vorgaben. Die Bayern KI-Richtlinie und weitere Datenschutzrichtlinien wie die DSGVO oder die KI-Verordnung der EU setzen grundsätzliche Anforderungen und Regelungen an den sicheren und verantwortungsvollen Umgang mit Künstlicher Intelligenz [1, 5, 6]. Die Entwicklung robuster Sicherheitsmechanismen ist daher nicht nur aus technischer, sondern auch aus regulatorischer Sicht von hoher Relevanz.

Um diesen Herausforderungen zu begegnen, ist es notwendig sowohl technische Sicherheitsmechanismen, als auch regulatorische Anforderungen zu analysieren und praktikable Abwehrstrategien zu entwickeln. Dazu gehören robuste Modell- und Trainingsarchitekturen, sowie integrierte Sicherheitsvorkehrungen. [8]

3 Konkrete Ziele

Diese Arbeit setzt sich zum Ziel, bestehende Schwachstellen systematisch zu identifizieren, geeignete Gegenmaßnahmen zu evaluieren und eine strukturierte Methodik zur Sicherheitsbewertung von LLMs zu entwickeln.

Dabei werden sowohl technische Angriffsmöglichkeiten als auch regulatorische Anforderungen betrachtet. Es soll eine Qualitätsbewertung verschiedener Modelle erfolgen, um diejenigen zu identifizieren, die gut gegen die identifizierten Bedrohungen gewappnet sind. Zudem wird eine mögliche Entwicklung eines Webinterfaces zur Durchführung von Sicherheitsexperimenten in Betracht gezogen.

Zur systematischen Bewertung der Sicherheitsaspekte wird ein Bewertungsansatz für die Sicherheitsprüfung von LLMs entwickelt. Dieser Ansatz soll als Grundlage für die Evaluierung unterschiedlicher Modelle dienen und mögliche Sicherheitslücken aufzeigen.

Literatur

[1] European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending various Regulations and Directives (Artificial Intelligence Act)*. 2024. URL: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32024R1689 (besucht am 03. 03. 2025).

[2] Mohamed Amine Ferrag u. a. *Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities*. 2025. arXiv: 2405.12750 [cs.CR]. URL: https://arxiv.org/abs/2405.12750.

[3] OWASP Foundation. *OWASP Top 10 for Large Language Model Applications – 2023 (Version 1.1)*. 2023. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_1.pdf (besucht am 03. 03. 2025).

[4] Mohammed Hassanin und Nour Moustafa. *A Comprehensive Overview of Large Language Models (LLMs) for Cyber Defences: Opportunities and Directions*. 2024. arXiv: 2405.14487 [cs.CR]. URL: https://arxiv.org/abs/2405.14487.

[5] Europäische Union. *Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates*. 2025-03-08. 2016. URL: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016R0679.

[6] Bayerisches Staatsministerium für Wissenschaft und Kunst. *Rechtsgrundlagen für den Einsatz von Künstlicher Intelligenz in der Verwaltung*. 2024. URL: https://web.archive.org/web/20241124035458/https://digitalverbund.bayern/wp-content/uploads/sites/12/2024/07/3.-Rechtsgrundlagen-fuer-den-Einsatz-von-KI.pdf.

[7] Jiacen Xu u. a. *AutoAttacker: A Large Language Model Guided System to Implement Automatic Cyber-attacks*. 2024. arXiv: 2403.01038 [cs.CR]. URL: https://arxiv.org/abs/2403.01038.

[8] Yifan Yao u. a. „A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly“. In: *High-Confidence Computing* 4.2 (2024), S. 100211. ISSN: 2667-2952. DOI: https://doi.org/10.1016/j.hcc.2024.100211. URL: https://www.sciencedirect.com/science/article/pii/S266729522400014X.