Risikobasiertes Framework für robuste LLM-Integration

Sandro Hartenstein, M.Sc. sandro.hartenstein@hwr-berlin.de

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

MOTIVATION

Semantische Fragilität, Inkonsistenzen und Halluzinationen führen häufig zu Fehlentscheidungen in LLM-basierten Systemen. Trotz vorhandener Leitlinien fehlt eine praxisnahe Operationalisierung zur systematischen Absicherung derartiger Integrationen.

METHODIK

Design Science Research

PROBLEM	SOLUTION	DESIGN	EVALUATION	COMMUNICATION
Gap-Analyse	Framework	Threat Modeling, Controls	Fallstudien, Tests	Publikation, Tools

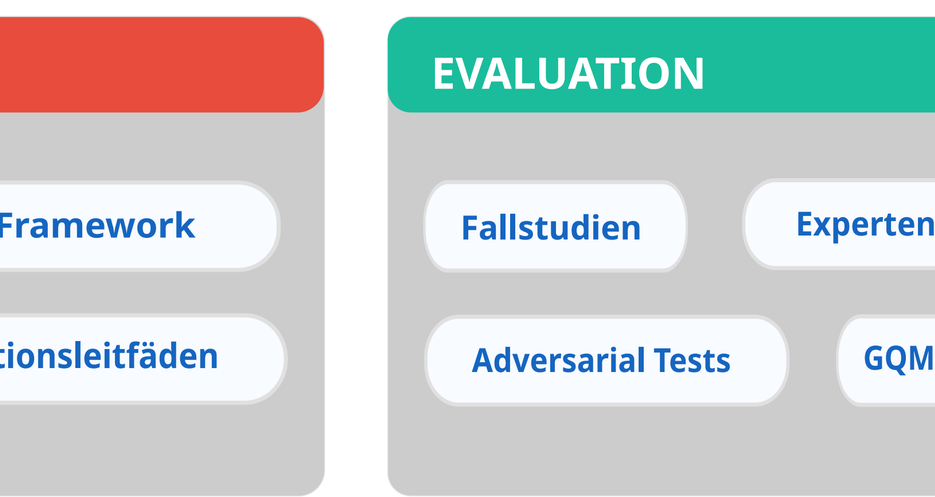
Empirische Erhebungen:
4289 OpenAPI Specs • 430 Adversarial Tests • 31 Online-Umfrage • 10 Experten-Diskussion

ERKENNTNISSE — Praxisprobleme & Robustheitslücken

Top-Praxisprobleme (Online-Umfrage, N=31)

Folgen:
- 41.9% Funktionale Fehler
- 41.9% Manuelle Korrekturen
- 29% Fehlentscheidungen

Provider-Robustheit (Adversarial Testing)



Kernerkenntniss:
- 21.6% Provider-Varianz (66–87.6%)

Rahmendaten:
- 430 Test-Prompts (promptfoo)
- Zeitpunkt: Juli 2024

SOLUTION & DESIGN

- Risk Assessment Matrix
- Monitoring Toolkit
- Test Framework
- Mitigation Catalog
- Robustness Checklist
- Integrationsleitfäden

EVALUATION

- Fallstudien
- Experten-Review
- Adversarial Tests
- GQM-Metriken

ZIEL

Ein risikobasiertes Framework für robuste LLM-Integration: Empirische Erkenntnisse werden mit bestehenden Standards verknüpft, um durch Threat Modeling, priorisierte Controls und praxisnahe Leitfäden die Verlässlichkeit von LLM-basierten Systemen zu erhöhen.

Details & Quellen

https://ha.rtenstein.com/esapi.html