Cheatsheet: Agent Evaluatie & Security

Snelreferentie voor alle kernconcepten per module.

Kernvaardigheden: Evaluatie · Prompt injection · Observability · Guardrails

Module 1: Evaluatie van agenten

- **Waarom agenten moeilijk te testen zijn**
  Non-determinisme, lange output en subjectieve kwaliteit: waarom traditionele testing niet werkt voor agenten.
- **Eval-frameworks en LLM-as-judge**
  Automatische evaluatie met LLM-as-judge, pairwise comparison en rubric-based scoring.
- **Testsets en regressietests bouwen**
  Hands-on: bouw een testset met edge cases en implementeer automatische regressietests.

Module 2: Security en guardrails

- **Prompt injection: aanval en verdediging**
  Hoe prompt injection werkt, waarom het gevaarlijk is en welke verdedigingsstrategieën effectief zijn.
- **Guardrails op input en output**
  Implementeer guardrails: input filtering, output validatie en content moderation.
- **Sandboxing van tool-aanroepen**
  Beperk de impact van gecompromitteerde agents: sandboxing, permission scoping en audit logging.

Module 3: Observability in productie

- **Tracing met Langfuse en OpenTelemetry**
  Implementeer distributed tracing voor je agent-pipelines: van request tot response.
- **Kosten- en latency-budgetten**
  Houd kosten en performance onder controle: budgeting, alerting en optimalisatie.
- **Incident response voor agenten**
  Wat te doen als het misgaat: runbooks, rollback-strategieën en post-mortems.