

Agent Evaluatie & Security

Maak je agenten betrouwbaar, veilig en meetbaar voordat ze live gaan

Niveau	Expert
Duur	5 weken · 4 - 6 uur/week
Lessen	9 lessen · ±271 minuten totaal
Categorie	Fundamentals

Beschrijving

Het verschil tussen een demo en een product zit hem in evaluatie en security. Leer eval-frameworks opzetten, prompt injection tegengaan en agent-gedrag monitoren in productie.

Voorkennis

- Ervaring met het bouwen van AI-agenten
- Kennis van Python of TypeScript
- Begrip van LLM API's en function calling

Wat je leert

- Zet een eval-framework op met LLM-as-judge
- Bouw testsets voor agent-gedrag en regressie
- Verdedig tegen prompt injection en jailbreaks
- Implementeer guardrails op in- en output
- Monitor agenten in productie met tracing
- Beheers kosten en latency onder belasting

Vaardigheden

Evaluatie · Prompt injection · Observability · Guardrails

Curriculum

Module 1: Evaluatie van agenten

Systematisch testen van agent-gedrag: van ad-hoc testing tot geautomatiseerde eval-pipelines.

Waarom agenten moeilijk te testen zijn 20 min [video]

Non-determinisme, lange output en subjectieve kwaliteit: waarom traditionele testing niet werkt voor agenten.

Eval-frameworks en LLM-as-judge 38 min [video]

Automatische evaluatie met LLM-as-judge, pairwise comparison en rubric-based scoring.

Testsets en regressietests bouwen 41 min [exercise]

Hands-on: bouw een testset met edge cases en implementeer automatische regressietests.

Module 2: Security en guardrails

Bescherm je agenten tegen misbruik, prompt injection en ongewenst gedrag.

Prompt injection: aanval en verdediging 33 min [video]

Hoe prompt injection werkt, waarom het gevaarlijk is en welke verdedigingsstrategieën effectief zijn.

Guardrails op input en output 30 min [video]

Implementeer guardrails: input filtering, output validatie en content moderation.

Sandboxing van tool-aanroepen 27 min [video]

Beperk de impact van gecompromitteerde agents: sandboxing, permission scoping en audit logging.

Module 3: Observability in productie

Monitor je agenten in productie: tracing, kosten en incident response.

Tracing met Langfuse en OpenTelemetry 35 min [video]

Implementeer distributed tracing voor je agent-pipelines: van request tot response.

Kosten- en latency-budgetten 25 min [video]

Houd kosten en performance onder controle: budgeting, alerting en optimalisatie.

Incident response voor agenten 22 min [video]

Wat te doen als het misgaat: runbooks, rollback-strategieën en post-mortems.
