Module 2: Security en guardrails

Agent Evaluatie & Security

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

Lessen in deze module

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.

Leerdoelen

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