COURSE SYLLABUS

AI Safety & Guardrails Engineering for Government Systems (Advanced) (Self-Paced)

Engineer the safety controls around government AI - guardrails, prompt-injection defense, red-teaming, evaluation, and monitoring for systems in production.

Group classes (Live Online) and onsite training are available for this course. For more information, email onsite@graduateschool.edu or visit: https://www.graduateschool.edu/courses/ai-safety-guardrails-engineering-for-government-systems-advanced-self-paced



support@graduateschool.edu •
(888) 744-4723

Course Outline

Module 1: What can go wrong

- AI failure and abuse modes
- Prompt injection and jailbreaks
- Government-specific risks
- A safety mindset

Module 2: Building guardrails

- Input validation and filtering
- Output checks and constraints
- Limiting tool and data access
- Defense in depth

Module 3: Red-teaming and evaluation

- Red-teaming your own system
- Building an evaluation harness
- Measuring safety and quality
- Catching regressions

Module 4: Monitoring and response (capstone)

- Monitoring in production
- Incident response for AI
- Documenting safety
- A safety plan for your system