AI501 — Advanced Generative AI Systems (15 ECTS)

Redwood Digital University · Advanced course in production-grade GenAI engineering. **Formatted to match your syllabus template** (Program Overview → Learning Outcomes → Curriculum Map → Tracks/Facilities → Capstone → QA → References → Appendices).

Program Overview

- **Award:** Advanced Course Certificate (stackable toward M.Sc.)
- **Course Code: AI501**
- **Duration:** 1 Semester (14 teaching weeks + assessment)
- **Total Credits: 15 ECTS** (≈ 375–450 total hours)
- **Delivery:** Lectures (L), Tutorials (T), Hands-on Labs (P), Studio/Project (S)
- **Typical Load:** ~8–10 h/week contact; ~18–20 h/week independent
- **Prerequisites:** Programming (Python), basic ML/AI, Git/Linux shell
- **Co-requisites (recommended):** Containers/Kubernetes self-study

Pillars

Prompt Engineering • Production AI Systems • Retrieval-Augmented Generation • AI Security & Guardrails • Observability • Tool-Calling & Agents (MCP) • Multi-modal AI • Model Optimization (Quantization/Compression/Fine-tuning) • MaaS & Enterprise Deployment • Semantic Routing

Graduate Learning Outcomes

Graduates will be able to:

1. **Design** sophisticated GenAI applications with state-of-the-art models and techniques.
2. **Engineer prompts** and templates with evaluation/versioning for reliable behavior.
3. **Build** scalable, production-ready systems with CI/CD, IaC, and GitOps.
4. **Implement RAG** pipelines (ingestion, indexing, retrieval) with citations & provenance.
5. **Secure** LLM apps using guardrails, abuse monitoring, threat models, and GenAIBOM.
6. **Integrate multi-modal** (text-image-audio) models and evaluate cross-modal performance.
7. **Optimize models** via quantization, compression, and task-specific fine-tuning.
8. **Instrument & monitor** apps for quality/safety SLIs, SLOs, and incident response.
9. **Orchestrate agents** with tool-calling and Model Context Protocol (MCP).
10. **Operate MaaS** with APIs, routing, quotas, and SLA governance.

Curriculum Map (By Week)

L-T-P-S per week (typical): 2-1-3-1
Assessment: Continuous evaluation + Capstone (see below)

Week 1 — AI Orientation (Module 1): GenAI use-case taxonomy; risks/benefits; KPIs.
Week 2 — AI Linguistics I (Module 2): Advanced prompting; templates; eval harness.
Week 3 — Ready to Scale 101 (Module 3): Llama Stack; GitOps; prompt/config versioning.
Week 4 — Ready to Scale 201 (Module 4): Continuous evaluation; promotion gates; canary/shadow.
Week 5 — RAG Foundations (Module 5): Embeddings; chunking; ingestion pipelines.
Week 6 — Guardrails (Module 6): Safety taxonomies; filters; jailbreak defense; bias mitigation.
Week 7 — Observability (Module 7): Tracing, metrics, logs; SLI/SLO; on-call runbooks.
Week 8 — Tool-Calling & Agents (Module 8): Function/tool calling; MCP; planner/critic loops.
Week 9 — LLM Security (Module 9): Threat models; secrets; RBAC; GenAIBOM; attacks & defenses.
Week 10 — Small Models (Module 10): Efficient architectures; edge constraints; routing.
Week 11 — Multi-modal Models (Module 11): VLMs; ASR/TTS; OCR; evaluation pitfalls.
Week 12 — MaaS (Module 12): API design; multi-tenant scaling; quotas; SLAs.
Week 13 — Quantization & Compression (Module 13): PTQ/QAT; pruning; KV cache tricks.
Week 14 — Fine-tuning & Semantic Router (Modules 14–15): SFT/LoRA; domain adaptation; context-aware routing.

Assessment Week: Capstone demo & viva; portfolio hand-in.

Course Modules (Detail)

Foundation (1–3): Orientation; AI linguistics & prompt optimization; Ready to Scale 101 (Llama Stack, GitOps, templates/versioning).
Production (4–7): Ready to Scale 201; RAG implementation; Guardrails; Observability.
Advanced (8–11): Tool-calling & Agents (MCP); LLM Security (GenAIBOM); Small Models; Multi-modal.
Optimization (12–15): MaaS; Quantization & Compression; Fine-tuning; Semantic Router.

Practical Implementation Areas (Tracks)

- **Production AI Systems:** Llama Stack, GitOps, CI/CD.
- **Knowledge Grounding:** RAG design, vector DBs, doc pipelines.
- **AI Safety & Security:** Guardrails, red-teaming, observability.
- **Advanced Applications:** Agents/tool-calling, multi-modal, model optimization.

Technology Stack & Facilities

- **AI/ML Platforms:** Llama Stack, Hugging Face, OpenAI, Anthropic, Azure OpenAI, Vertex AI.
- **Dev Tools:** Python, PyTorch/TensorFlow, LangChain, LlamaIndex, Docker, Kubernetes, Git, MLflow.

- **Infrastructure:** GPU clusters; vector DBs (Pinecone, Weaviate); AWS/Azure/GCP.
- **Security & Monitoring:** Guardrails libs; observability platforms; security scanning; GenAIBOM frameworks.

Capstone Design Sequence (AI590/AI591 Template)

Phase 1 — Proposal & Architecture (AI590, within semester): Problem framing, requirements, risk & ethics, architecture, evaluation plan, deployment plan. **Gate A:** Design review & safety sign-off.
Phase 2 — Build, Test & Validate (AI591, end-semester): Implementation, verification/validation, monitoring dashboards, security/GenAIBOM dossier, demo & post-mortem. **Gate B:** Public demo & repository handover.

Capstone Examples

- **Enterprise RAG System** (secure deploy, citations, governance)
- **Multi-modal AI Assistant** (text+image+audio, tool-calling)
- **AI Security Framework** (guardrails, monitoring, GenAIBOM)
- **Optimized Edge Deployment** (quantized model + semantic router)

Assessment & Quality Assurance

- **A1 Prompting & Eval Harness (10%)** — prompts, datasets, metrics, baseline report.
- **A2 RAG Mini-System (15%)** — ETL→vector DB→retrieval→generation with citations.
- **A3 Guardrails & Red-Team (10%)** — policy design, tests, mitigations.
- **A4 Observability Pack (10%)** — tracing/metrics/logs, dashboards, runbook.
- **A5 Optimization Lab (10%)** — quantize/prune; benchmark vs. baseline.
- **A6 Agent with Tools (10%)** — MCP/tool-calling agent + reliability tests.
- **Capstone (30%)** — build + demo + viva + dossier.
- **Participation (5%)** — code reviews, discussions.

Rubrics: Correctness; reliability/safety; documentation; observability evidence; performance/cost; ethics/compliance.
Integrity: Original work; attribution for models/datasets/code; logs may be audited.
Accessibility: Captions, alt-text, contrast; inclusive design.

Suggested Texts & References

- Designing Data-Intensive Applications (Kleppmann)
- Vendor docs: OpenAI/Anthropic/Azure OpenAI/Vertex AI; Hugging Face
- Framework docs: LangChain, LlamaIndex, MLflow
- Security: OWASP LLM Top 10; supply-chain security guides
- Inference optimization guides (CUDA/ONNX/TensorRT, CPU acceleration)

Accreditation Mapping (Template)

- **Engineering/Systems:** ≥ 6 ECTS (production, observability, security)
- **Data/AI Methods:** ≥ 6 ECTS (prompting, RAG, optimization)
- **Design/Project:** ≥ 3 ECTS (capstone + labs)

Customization & Localization Notes

- Swap providers/stacks per institution; map security to local policy; enable on-prem GPU or sovereign cloud; add legal/ethics module for regional compliance.

Appendix A — Weekly Syllabi Snapshots (Examples)

Week 5 RAG Clinic: Chunking strategies; hybrid search; evaluation with grounded answers; failure analysis.
Week 7 Observability Drill: Instrumentation; SLI/SLO design; incident tabletop; on-call runbook.
Week 13 Optimization Lab: PTQ vs. QAT; accuracy/latency/cost trade-offs; rollback plan.

Appendix B — Example Capstone Briefs

- **Enterprise RAG:** Ingestion, indexing, retrieval, guardrails, governance dashboard.
- **Security Framework:** Threat model, guardrail pack, GenAIBOM, purple-team report.
- **Edge Assistant:** Small-model pipeline, quantization, semantic router, offline fallback.

Appendix C — Rubrics (Abbreviated)

- **Design Project:** Requirements (15), Architecture & trade-offs (20), Implementation (25), V&V evidence (20), Documentation (10), Security & ethics (10).
- **Lab Report:** Reproducibility (10), Method (20), Data & analysis (30), Discussion (20), Presentation (10), Safety/compliance (10).

This syllabus mirrors your provided format while preserving AI501's module content and outcomes.