

AI system design and security review

Use for the Week 9 deployment and the capstone code review. Risks are grouped in the spirit of the OWASP Top 10 for LLM Applications; check the current list at owasp.org.

Architecture

- [] A diagram: clients, API, queue, model providers, vector store, database, observability
- [] Every model call goes through one module with timeouts, retries with backoff, and a fallback
- [] Long jobs run in a queue, not in the request
- [] Caching for repeated prompts and embeddings
- [] Clear data flow: what personal data goes where

Prompt injection and untrusted content

- [] User input and retrieved documents are marked as data (tags or separate messages)
- [] Rules that matter (refund limits, permissions) are enforced in code
- [] Tools that change things need confirmation; tool inputs are schema-validated
- [] Model output is never executed or rendered as raw HTML without sanitising

Data and privacy

- [] Personal data minimised or removed before model calls
- [] Provider retention and training settings checked and documented
- [] Vector store access scoped per tenant or user
- [] Compliance with the Nigeria Data Protection Act 2023, and GDPR for EU users
- [] Secrets in a secret manager or environment variables; rotated if exposed

Agents

- [] A maximum number of steps and tool calls
- [] Allow-list of tools per task; least privilege on credentials
- [] Human approval for irreversible actions
- [] Full trace of each run for review

Quality

- [] Eval set of 50+ cases, run in CI before prompt or model changes
- [] Regression threshold that blocks a release
- [] Online signals: thumbs, escalations, refusals
- [] Hallucination checks on a sample each week

Cost and reliability

- [] Cost per request and per user tracked; dashboard and alerts
- [] Rate limits per user and per IP; spend caps at the provider
- [] Smaller models for simple steps; prompt caching where supported
- [] Load tested at expected peak; graceful degradation when a provider is down

Deployment

- [] Docker image builds reproducibly; environment-specific config

- ☐ Health checks and structured logs without personal data
- ☐ Rollback plan; versioned prompts and models
- ☐ README with setup, architecture, evals and known limits