## **90-Day Generative AI System Design Plan – Phase 2 (Interview-Ready Pass)**

### **Approach**

* **End-to-End coverage of all 6 phases**
* **Target Depth**: Enough to explain confidently in interviews, draw diagrams, and cite relevant patterns.
* **Time Allocation**: ~3–4 days per module on average.
* **Output**:  
  1. 8–12 page theory notes per module
  2. HLD + LLD diagram for each
  3. 5–8 interview-ready scenario Q&As per module

### **Execution Sequence**

We’ll group the **25 modules** into weekly themes to maintain continuity.

#### **Week 1–2: Foundations (Phase 1)**

1. Evolution of System Design in AI
2. Core Principles of System Design (scalability, latency, trade-offs)
3. GenAI-Specific Constraints (latency, token limits, compliance)
4. HLD vs LLD in GenAI

#### **Week 3–4: HLD for GenAI (Phase 2)**

1. Component Layering in GenAI Systems
2. Core GenAI Architectural Patterns (RAG, agentic, multi-modal)
3. Integration Patterns (API gateway, microservices, serverless)
4. Security, Compliance, and Guardrails (at HLD level)

#### **Week 5–6: LLD for GenAI (Phase 3)**

1. Data Ingestion & Preprocessing (loaders, chunking)
2. Embedding & Vector Store Layer (Pinecone, Qdrant, hybrid search)
3. Prompt Orchestration Layer (templates, dynamic assembly)
4. Agent Patterns (ReAct, Plan-and-Execute, negotiation)
5. Validation & Structured Outputs (Pydantic, JSON schema basics)
6. Evaluation Layer (RAGAS, TruLens basics)
7. Observability Layer (LangSmith, cost tracking, logging)

#### **Week 7–8: Production Architectures (Module 4)**

1. FAANG-Grade RAG Pipelines (modular structure, domain tuning)
2. Enterprise Agent Systems (tool registry, HITL integration)
3. Scaling & Deployment (vLLM, serverless, AWS/GCP/Azure)
4. Caching & Performance Optimization

#### **Week 9: Specialized Designs (Module 5)**

1. Multi-Modal System Design (vision-language, speech)
2. Domain-Specific GenAI (legal, medical, financial, e-commerce)
3. Compliance-Aware Architectures (LexiGuard model)

#### **Week 10: Future-Readiness (Module 6)**

1. Self-Improving Systems (auto-eval + finetuning)
2. On-Device & Edge GenAI (Ollama, GGUF)
3. Hybrid Human+AI Systems (HITL patterns)
4. Anti-Patterns in GenAI System Design

### **Deliverables Per Module (Phase 2 Version)**

* **Notes**: 8–12 pages (clear, concise, interview-focused)
* **Diagram**: 1 HLD + 1 LLD diagram
* **Q & A**: 5–8 scenario questions
* **Examples**: 1–2 relevant architecture or project references

### **Post-Phase 2 (Deep-Dive Plan)**

When we revisit in deep-dive mode, each topic expands to:

* 20–30 pages of notes
* Hands-on reference implementation
* Real-world performance, cost, compliance metrics
* Multiple case studies & comparisons