CAREER TRANSFORMATION GUIDE

SK

6-MONTH ROADMAP

QA → GenAI QA Transformation

From test execution to AI system validation. A practical month-by-month guide for QA engineers stepping into the GenAI era.

M1 MINDSET
M2 EVALS
M3 RAG
M4 OBSERV.
M5 SAFETY
M6 AGENTS

6	30+	6	QA
MONTHS	TOOLS & SKILLS	DELIVERABLES	→ GENAI QA

THE FULL JOURNEY AT A GLANCE

M1

MONTH 1

Foundations & Mindset Shift

Understand why traditional testing breaks for GenAI. Learn LLMs, tokens, embeddings, temperature. Build your first prompt evaluator.

▸ Prompt-response evaluator + failure mode log

M2

MONTH 2

LLM Evaluation & Metrics

Learn how GenAI quality is measured. Build automated evaluation pipelines. Compare model versions objectively with RAGAS, DeepEval, Braintrust.

▸ LLM evaluation report with metrics & failure classification

M3

MONTH 3

RAG & Knowledge Reliability Testing

Validate AI systems backed by enterprise data. Test retrieval precision & recall. Inject corrupted documents. QA now validates data pipelines, not just app logic.

▸ RAG test suite with retrieval + faithfulness validation

M4

MONTH 4

Observability, Tracing & Production Readiness

Make GenAI debuggable in production. Trace prompt → tool → response chains. Detect latency spikes, token explosions, model drift with LangSmith & Arize.

▸ Production-ready GenAI observability dashboard

M5

MONTH 5

Safety, Guardrails & Risk-Based AI Testing

Prevent enterprise-level AI failures. Build red-team prompt suites. Validate refusal behavior, test boundary violations. Policy-as-code checks.

▸ Red-team suite + refusal behavior test report

M6

MONTH 6

Agentic AI Testing & Enterprise Reality

Test AI systems that plan and act. Validate multi-step agents, goal completion rate, unsafe action rate, recovery from failure, human-in-the-loop gates.

▸ Agent test framework + enterprise risk assessment

THE CORE MINDSET SHIFT

Traditional QA Thinks...

Pass / Fail · Deterministic outputs · Fixed test cases · Application logic · Binary correctness

→

GenAI QA Thinks...

Quality spectrum · Probabilistic outputs · Evaluation pipelines · Data + model + system · Usefulness + safety

MONTH 1 — Foundations & Mindset Shift

Objective: *Shift from test execution to system validation thinking. Understand why GenAI breaks traditional QA.*

📚 LEARN

- LLMs: tokens, embeddings, temperature
- Determinism vs variability in AI outputs
- Why traditional test cases fail for GenAI
- Core failure modes: hallucination, bias, unsafe output, prompt sensitivity, latency & cost instability

🛠️ HANDS-ON

- Build a simple LLM prompt-response evaluator
- Compare fixed vs variable outputs across temperature settings (0.0 → 1.0)
- Log prompts, responses, metadata in structured format
- Document 5 real hallucination examples you trigger

🔧 TOOLS

`OpenAI API` `Gemini API (free)` `Python` `Jupyter Notebook`

💡 PROJECT IDEA

Build a **Temperature Explorer** — run the same prompt 10x at temp 0, 0.5, 1.0. Chart the output variation. Your first GenAI QA insight.

MONTH 2 — LLM Evaluation & Metrics

Objective: *Learn how GenAI quality is measured — the core QA skill upgrade for AI systems.*

📚 LEARN

- Evaluation dimensions: correctness, faithfulness, relevance, context recall
- Ground truth vs reference-free evaluation
- Accuracy vs usefulness — they're not the same in GenAI
- Batch evaluation strategy and prompt variation testing

🛠️ HANDS-ON

- Build automated evaluation pipelines
- Run batch evaluations on prompt variations
- Compare model versions objectively with metrics
- Write unit-style LLM tests with DeepEval

🔧 TOOLS

`RAGAS` `DeepEval` `Braintrust`

💡 PROJECT IDEA

Build a **Model Comparison Dashboard** — evaluate GPT-4o vs Gemini on the same 50 prompts using DeepEval. Score each.

🏆 DELIVERABLE

LLM evaluation report with metrics & failure classification by type and severity.

MONTH 3 — RAG & Knowledge Reliability Testing

Objective: *Validate AI systems backed by enterprise data. QA now validates data pipelines, not just application logic.*

RAG PIPELINE — WHAT YOU'RE TESTING

User Query → Embedding Model → Vector DB Retrieval → Context Window → LLM Answer

- ⚠ Bad chunking → wrong context
- ⚠ Embedding mismatch → poor retrieval
- ⚠ Retrieval drift → hallucination

🛠️ HANDS-ON

- Test retrieval precision & recall on known documents
- Inject corrupted / outdated documents into the vector DB
- Validate answer faithfulness to source documents
- Test chunking strategies — size vs overlap impact

💡 PROJECT IDEA + TOOLS

Build a **RAG Corruption Test Suite** — inject wrong facts into docs, measure how often the LLM still hallucinates vs refuses.

`RAGAS` `LangChain` `ChromaDB` `DeepEval`

MONTH 4 — Observability, Tracing & Production Readiness

Objective: *Make GenAI debuggable in production. Logs ≠ traces for LLMs.*

📚 LEARN

- Why logs ≠ traces for LLM systems
- Prompt lineage & versioning strategy
- Model behavior drift detection over time
- Token usage monitoring & cost alerting

🛠️ HANDS-ON

- Trace prompt → tool → response chains end-to-end
- Detect latency spikes & token explosions
- Compare behavior across deployments (staging vs prod)
- Set up drift alerts on quality metric regression

🔧 TOOLS

`LangSmith` `Arize AI` `Prometheus`

💡 PROJECT IDEA

Build a **GenAI Health Dashboard** — track p95 latency, token cost, quality score, and hallucination rate across 100 daily requests.

🏆 DELIVERABLE

Production-ready GenAI observability dashboard with live metrics.

MONTH 5 — Safety, Guardrails & Risk-Based AI Testing

Objective: *Prevent enterprise-level AI failures before they reach users or regulators.*

🔒	⚠️	📋	💉
Data Leakage	**Unsafe Instructions**	**Compliance Violations**	**Prompt Injection**
PII in prompts & responses	Harmful or illegal guidance	GDPR, HIPAA, regulatory	Jailbreaks & overrides

🛠️ HANDS-ON

- Build red-team prompt suites (50+ adversarial prompts)
- Validate refusal behavior — does it refuse the right things?
- Test boundary violations — what slips through guardrails?
- Implement policy-as-code checks for content classification

💡 PROJECT IDEA + TOOLS

Build a **Red Team Test Library** — categorize 50 adversarial prompts by risk type, run against your LLM, score refusal rate and leakage rate.

`Guardrails AI` `NeMo Guardrails` `Custom policy checks`

MONTH 6 — Agentic AI Testing & Enterprise Reality

Objective: *Test AI systems that plan and act. Enterprise reality: legal, security, and QA intersect.*

AGENT ARCHITECTURE — WHAT YOU'RE TESTING

✓ Goal completion rate | ✓ Unsafe action rate | ✓ Recovery from failure

📚 LEARN

- Agent architectures: planner, executor, memory
- Non-deterministic multi-step workflows
- Why step-based test cases fail for agents
- Human-in-the-loop gates and when to trigger them

🛠️ HANDS-ON + PROJECT

- Test multi-step agents end-to-end
- Validate goal completion vs unsafe action rate
- Introduce intentional failures — test recovery
- **Project:** Build an **Agent QA Harness** — automate goal injection, capture full execution traces, score on 3 metrics above

COMPLETE GENAI QA TOOLS STACK

🤖 LLM APIS

- OpenAI API
- Gemini API
- Anthropic API
- Ollama (local)

📊 EVALUATION

- RAGAS
- DeepEval
- Braintrust
- PromptFoo

🗄️ RAG STACK

- LangChain
- ChromaDB
- Pinecone
- LlamaIndex

🔭 OBSERVABILITY

- LangSmith
- Arize AI
- Phoenix
- Prometheus

🛡️ SAFETY

- Guardrails AI
- NeMo Guardrails
- Custom policy

🤖 AGENTS

- LangGraph
- AutoGen
- CrewAI
- LangSmith

WHAT YOU'LL BE ABLE TO DO AFTER 6 MONTHS

🎯 Design LLM Evaluation Pipelines

Build automated eval suites that measure quality beyond pass/fail.

🗄️ Validate RAG Systems End-to-End

Test retrieval, chunking, faithfulness — not just the final output.

🔭 Monitor AI in Production

Detect drift, token explosions, latency spikes before they become incidents.

🛡️ Red-Team & Harden AI Systems

Build adversarial test suites. Validate guardrails. Own AI safety in your org.

🤖 Test Agentic AI Systems

Validate goal completion, unsafe action rate, recovery — beyond step-based tests.

💼 Speak the Language of AI Teams

Bridge QA, legal, security and engineering in AI product decisions.