THE LEARNER'S BRIEF · 5 RUNGS · 6 TOPICS · 3 CLOUD TRACKS

Q2 · 2026

L4

CURIOUS · BUILDER · PRACTITIONER · SPECIALIST · HERO

GenAI Professional

Learning. Zero to Hero.

One ladder. Six topics. Three cloud tracks — pick the rung you're on, not the one you wish you were.

THE BRIEF, IN THREE BEATS

01 The bottleneck isn't access to models. It's a missing rung.

02 Each rung has a build artifact — not a course completion.

03 Cloud track is the last call, not the first.

20 slides · 5 rungs · 6 topics · AWS/Azure/GCP tracks · 6-12 month roadmap

For engineers, scientists, architects, leads

SLIDE 02 · HOOK

"I've read 40 papers and watched

100 hours of lectures — and I still can't ship a RAG."

The most common GenAI learner sentence in 2026. The fix isn't more content — it's a different ladder.

Composite quote — field pattern observed across DeepLearning.AI cohort threads, r/LocalLLaMA, and HF agents-course discussion (2025-2026).

LEARNING CLAIM

GenAI mastery is not a content problem — it's a build-artifact problem. You don't level up by finishing a course. You level up by shipping the thing that course was setup for.

THREE COROLLARIES

01 The course is the *scaffold*. The repo is the artifact. No repo, no rung.

02 Cloud track is a *deploy decision*, not a learning prerequisite. Pick after L1.

03 Papers come *after* shipping — they explain what you already felt break.

THE HONEST TRADE

Two learning modes — pick deliberately.

Linear (textbook): Karpathy → CS25 → papers. Deepest, slowest, easy to drift.

Spiral (artifact-led): Ship a thing → break it → read the paper that names what broke → ship next.

> This deck commits to spiral. Most engineers reach L3 in 6-9 months on spiral; linear takes 18+ and many quit at L1.

Field pattern across DeepLearning.AI, HF agents-course, Karpathy "Neural Networks Zero to Hero" cohorts (2024-2026).

SLIDE 04 · THE CAPABILITY LADDER — FIVE RUNGS

Each rung is an artifact + a signal — not a course.

RUNG	NAME	TIME ON SPIRAL	SHIP ARTIFACT	SIGNAL YOU'RE THERE
L0	Curious	Week 0–2	A chat session log with 5 working prompts you actually used.	You stop asking "what is a token" and start measuring them.
L1	Builder	Month 1–3	A working RAG over a private doc set, deployed somewhere reachable.	You know why retrieval beats fine-tuning for "answer from my docs".
L2	Practitioner	Month 3–6	An agent with tool-use, an eval harness, and a per-call cost dashboard.	You can debug a wrong answer from logs alone — no re-running.
L3	Specialist	Month 6–12	A depth pick — fine-tune, agent platform, eval harness, or red-team rig.	Teams ping you when the playbook doesn't cover their case.
L4	Hero	Year 1+	An org-wide architecture doc others build against.	You're invited to the model-choice / cloud-choice meeting before it starts.

Ladder synthesis — based on AWS/Azure/GCP certification progressions, Karpathy NN Zero-to-Hero, DeepLearning.AI specialization sequence.

SLIDE 05 · HOW TO READ THE NEXT 15 SLIDES

READING MAP

Three sections. Read your row, then go shallow on the rest.

A · THE FIVE RUNGS

Slides 06–10

06 · **L0 Curious.** First chat, first prompt, first cost.

07 · **L1 Builder.** First RAG, hosted.

08 · **L2 Practitioner.** Agent + eval + cost.

09 · **L3 Specialist.** Pick a depth.

10 · **L4 Hero.** Architecture authority.

> Start here. Find your rung. The artifact tells you which.

B · THE SIX TOPICS

Slides 11–15

11 · **Foundations.** Tokens, transformers, attention.

12 · **Prompting & context.** The skill behind every rung.

13 · **RAG.** Retrieval, chunking, embeddings, grounding.

14 · **Agents.** Tools, memory, planning, multi-step.

15 · **Eval & safety.** The discipline that separates L2 from L3.

> Each topic surfaces at multiple rungs — you spiral, you don't graduate.

C · THE THREE CLOUD TRACKS

Slides 16–18

16 · **AWS track.** Bedrock + AIF-C01 + AgentCore.

17 · **Azure track.** Azure AI Foundry + OpenAI + AI-102 successor.

18 · **GCP track.** Vertex AI + Gemini + GenAI Leader cert.

> Pick after L1 — not before. The first RAG can run on any of the three; later choices are sticky.

Closing: slide 19 — five anti-patterns that hold learners at L1. Slide 20 — pick your next rung.

RUNG 1 OF 5 · CURIOUS — THE FIRST TWO WEEKS

WEEK 0–2 · ~20 HRS

L0 Stop being a user. Start measuring tokens.

SHIP THIS

A chat session log with 5 working prompts you used to do real work — not toy questions.

01 · Pick one real task you do weekly (PR review, ticket triage, log summary).

02 · Run it through Claude / ChatGPT / Gemini side-by-side. Same prompt.

03 · Note input + output token counts. (Every chat UI now shows them.)

04 · Keep the 5 prompts that worked. Discard the rest.

05 · Write a one-paragraph reflection: where did each model fail differently?

RESOURCES — PICK ONE

VIDEO · Karpathy — *Intro to LLMs* (1h, free, YouTube).

COURSE · DeepLearning.AI — *ChatGPT Prompt Engineering for Developers* (90 min, free).

READ · Anthropic / OpenAI prompt-engineering docs — whichever model you'll use most.

SIGNAL YOU'RE DONE

You stop asking "what's a token" — and start asking "how many of them did that cost".

When you catch yourself rephrasing a prompt to use fewer tokens, you've left L0.

Sources: karpathy.ai · deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers · docs.anthropic.com/prompt-engineering.

RUNG 2 OF 5 · BUILDER — YOUR FIRST RAG, HOSTED

MONTH 1–3 · ~80 HRS

L1

Ship a RAG over docs you actually own.

SHIP THIS

A working RAG over a private doc set (your team's wiki, a PDF corpus, your bookmarks) — deployed somewhere a colleague can hit.

Stack · Python + LangChain or LlamaIndex (pick one). Vector store: pgvector or Pinecone free tier.

Model · Whichever fits your cloud track — Claude on Bedrock, GPT-4 on Azure OpenAI, or Gemini on Vertex.

Deploy · A single endpoint — Streamlit + Lambda, Container Apps, or Cloud Run.

Must include · Citations on every answer + a daily token-cost log.

RESOURCES

COURSE · DeepLearning.AI — *LangChain Chat with Your Data* (free, 1h).

COURSE · Hugging Face — *LLM Course*, units 1–3 (free).

DOCS · LlamaIndex starter + your chosen cloud's RAG quickstart.

SIGNAL YOU'RE DONE

You can explain — from your repo — why retrieval beats fine-tuning for "answer from my docs."

And you know which 3 chunks in your store cause the most wrong answers.

Sources: deeplearning.ai/short-courses/langchain-chat-with-your-data · huggingface.co/learn/llm-course · docs.llamaindex.ai.

RUNG 3 OF 5 · PRACTITIONER — AGENTS, EVAL, COST

MONTH 3–6 · ~150 HRS

L2 Debug from logs. Trust your eval. Watch the bill.

SHIP THIS — THREE ARTIFACTS, NOT ONE

01 · AGENT
An agent with 3+ tools (search, calculator, internal API). Multi-step. Backed by your L1 RAG.

02 · EVAL HARNESS
30+ golden Q&A pairs. Auto-run on every prompt or model change. Pass/fail in CI.

03 · COST DASHBOARD
Per-call token+$ logged. p50 & p95 latency. Grouped by route, by user, by model.

RESOURCES

COURSE · Hugging Face *AI Agents Course* (free, certified, units 1–4).

COURSE · DeepLearning.AI — *Evaluating & Debugging GenAI*, *AI Agents in LangGraph*.

PAPER · ReAct (Yao et al., 2022) — the original tool-use loop.

SIGNAL YOU'RE DONE

A wrong answer arrives. You open the trace, find the bad tool call, fix the prompt, re-run eval — no guessing.

Mean-time-to-root-cause measured in minutes, not afternoons.

Sources: huggingface.co/agents-course · deeplearning.ai/short-courses · arxiv.org/abs/2210.03629 (ReAct).

RUNG 4 OF 5 · SPECIALIST — PICK A DEPTH

MONTH 6–12 · ~300 HRS

L3 Stop being a generalist. Own one corner of the stack.

DEPTH A

Fine-tuning & distillation

Ship: A distilled model that beats a frontier model on your domain at 1/10 the cost.

COURSE · HF LLM Course chapters 5–7. DeepLearning.AI *Finetuning LLMs*.

PAPER · LoRA (Hu et al., 2021). QLoRA (Dettmers, 2023).

CLOUD · Bedrock Distillation · Azure AI Studio fine-tune · Vertex Tuning.

DEPTH B

Agent platforms

Ship: Multi-agent system with shared memory, MCP gateway, and OBO identity.

COURSE · HF Agents Course units 4–5. LangGraph deep-dive.

PAPER · AutoGPT post-mortems. Anthropic's "Building Effective Agents" essay.

CLOUD · Bedrock AgentCore · Azure AI Agent Service · Vertex Agent Builder.

DEPTH C

Eval & observability

Ship: LLM-as-judge harness with calibration + drift detection in CI.

COURSE · DeepLearning.AI *Quality & Safety for LLMs*.

PAPER · Chatbot Arena. G-Eval. PoisonedRAG (USENIX '25).

CLOUD · Bedrock Eval · Azure AI Evaluation · Vertex Gen AI Eval.

DEPTH D

Security & red-team

Ship: Red-team CI rig + Guardrails policy that catches OWASP LLM Top-10.

READ · OWASP LLM Top 10 (2025). MITRE ATLAS.

PAPER · Spotlighting (arXiv:2403.14720). CaMeL.

CLOUD · Bedrock Guardrails · Azure Content Safety · Vertex Safety.

Pick ONE depth. Specialists earn the call; generalists at L3 just keep building L2 things.

RUNG 5 OF 5 · HERO — ARCHITECTURE AUTHORITY

YEAR 1+ · ONGOING

L4

Write the doc your org builds against.

SHIP THIS

An *org-wide GenAI architecture doc* — model choice, eval policy, security baseline, cost guardrails — that 5+ teams build against.

It defines · approved models, prompt review process, eval-as-CI, red-team cadence, IAM/KMS baseline.

It defers · use-case specifics. Heroes set the floor; teams ship above it.

It is reviewed · quarterly — the field moves; the doc moves with it.

RESOURCES AT THIS RUNG

CERT · AWS ML Engineer Associate or GenAI Developer Professional · Azure AI Engineer · GCP ML Engineer.

READ · NIST AI 600-1. OWASP LLM Top 10. Cloud-vendor well-architected GenAI lenses.

PEERS · Internal arch-review forum — your replacement input.

SIGNAL YOU'RE HERE

You're invited to the model-choice meeting *before* it starts — not asked to justify the decision after.

And junior engineers cite your doc, not Karpathy's video.

Sources: AWS/Azure/GCP well-architected GenAI guidance · NIST AI 600-1 (2024) · OWASP LLM Top 10 (2025).

SPIRALS AT L0 → L4

T1

You can't debug what you can't name.

THE MINIMUM VOCABULARY

TOKEN · The unit billed. Not a word, not a character. ~4 chars English, ~1 CJK.

CONTEXT WINDOW · Tokens in + tokens out, capped. 200K Claude, 1M Gemini.

ATTENTION · Every token reads every other token. $O(n^2)$. Why context is expensive.

EMBEDDING · Token → vector. The translator between text and math.

PRE-TRAIN · **FINE-TUNE** · **RLHF** · Three training stages. Most teams only touch the last two.

TEMPERATURE · **TOP-P** · Knobs that turn deterministic into creative. Default 0.7; set 0 for eval.

CANONICAL SOURCES

Karpathy · *Neural Networks: Zero to Hero* (YouTube, free, ~15h, build GPT-2 from scratch).

Karpathy · *Intro to LLMs* (1h talk).

Stanford CS25 · *Transformers United V6* (2026, free YouTube lectures).

THE TWO PAPERS — READ AFTER L2

Attention Is All You Need · Vaswani et al., 2017 — the transformer paper. 11 pages. Read it after you've felt context length bite.

The Annotated Transformer · Harvard NLP. Code + math side-by-side, the right second read.

Sources: karpathy.ai · web.stanford.edu/class/cs25 · arxiv.org/abs/1706.03762 · nlp.seas.harvard.edu/annotated-transformer.

EVERY RUNG USES IT

T2 It stopped being prompts. It's the whole context.

SIX TECHNIQUES WORTH YOUR TIME

01 SYSTEM PROMPT · Role, constraints, output format. Loaded once, reused.

02 FEW-SHOT · 2–5 input/output examples beats any adjective.

03 CHAIN-OF-THOUGHT · "Think step by step" — only when you need reasoning, not for lookup.

04 STRUCTURED OUTPUT · JSON schema / tool-use formatting. Parse, don't regex.

05 CACHE CONTROL · Mark stable prefixes — 6×+ cost reduction on long contexts.

06 CONTEXT BUDGET · Lead with the question, then the docs. Models forget the middle.

VENDOR PROMPT GUIDES

Anthropic · *Prompt engineering for Claude* + the Console's prompt library.

OpenAI · *GPT-4 best practices* + structured outputs guide.

Google · *Gemini API prompting strategies* + Vertex prompt gallery.

TWO PAPERS — THE ONES WORTH OPENING

Lost in the Middle · Liu et al., 2023 — models attend less to the middle of long context. Why context order matters.

Chain-of-Thought · Wei et al., 2022 — the technique that gave reasoning a name.

Sources: docs.anthropic.com · platform.openai.com/docs · ai.google.dev · arxiv.org/abs/2307.03172 · arxiv.org/abs/2201.11903.

THE L1 ARTIFACT

T3

The five-step pipeline. Each step has its own failure mode.

PIPELINE · FIVE STEPS, FIVE FAILURE MODES

01 INGEST · S3, SharePoint, Confluence. Fail: stale data, missing PII redaction.

02 CHUNK · 200–800 tokens, overlap 50–100. Fail: cuts mid-table, mid-list, mid-thought.

03 EMBED · Titan / OpenAI / Cohere / Gemini embeddings. Fail: model changed, index didn't.

04 RETRIEVE · Vector + BM25 hybrid + rerank. Fail: top-k all from one doc.

05 GROUND · Pass with citations, require quoting source. Fail: model invents around the chunk.

RESOURCES AT L1 → L3

COURSE · DeepLearning.AI — *LangChain Chat with Your Data*, *Building & Evaluating Advanced RAG*.

DOCS · LlamaIndex tutorials — chunking, hybrid retrieval, query rewriting.

CLOUD · Bedrock Knowledge Bases · Azure AI Search RAG · Vertex Vector Search.

PAPERS WORTH OPENING

RAG (Lewis et al., 2020) · The original. Read after L1, not before.

Self-RAG · HyDE · PoisonedRAG (USENIX '25) · The L3 reading list — retrieval becomes adversarial.

Sources: deeplearning.ai/short-courses · docs.llamaindex.ai · arxiv.org/abs/2005.11401 · usenix.org/conference/usenixsecurity25.

THE L2 ARTIFACT

T4

An LLM that calls tools in a loop. That's it.

FIVE PRIMITIVES — LEARN THEM BY NAME

01 TOOL · A function the model can call: search, calc, internal API. JSON-schema'd in/out.

02 LOOP · Think → tool-call → observe → think… until done or max-steps.

03 MEMORY · Short-term (this session) + long-term (across sessions). User-scoped, encrypted.

04 PLANNING · Break a goal into steps before executing. Optional — not always worth it.

05 MCP · Model Context Protocol — the emerging way tools advertise themselves to agents.

COURSES WORTH COMPLETING

Hugging Face · *AI Agents Course* (free, certified) — smolagents, LangGraph, LlamaIndex.

DeepLearning.AI · *AI Agents in LangGraph, Functions, Tools & Agents with LangChain.*

Anthropic · *Building Effective Agents* essay + cookbook recipes.

CLOUD-NATIVE AGENT PLATFORMS

AWS · Bedrock AgentCore (Runtime, Identity, Memory, Tools, Observability).

Azure · Azure AI Agent Service (Foundry).

GCP · Vertex AI Agent Builder + ADK (Agent Development Kit).

Sources: huggingface.co/agents-course · anthropic.com/research/building-effective-agents · modelcontextprotocol.io.

THE BORING DISCIPLINE

T5

"It works on the demo" is not eval.

FOUR EVAL MODES — PICK THE SMALLEST THAT WORKS

01 GOLDEN SET · 30–100 hand-curated Q&A. Exact-match / contains / regex. Cheap, fast.

02 LLM-AS-JUDGE · A second model scores answers on a rubric. Calibrate against humans on 50.

03 HUMAN A/B · Side-by-side comparison on real traffic. The truth, but slow + expensive.

04 RED-TEAM · Adversarial probes — prompt-injection, jailbreaks, PII exfil. Run in CI.

FRAMEWORKS WORTH KNOWING

OWASP LLM Top 10 (2025) · The vulnerability taxonomy every L3 should recite.

NIST AI 600-1 · The GenAI risk profile US regulators reference.

MITRE ATLAS · Adversarial ML threat matrix — the offensive playbook.

COURSES & TOOLS AT L2→L3

COURSE · DeepLearning.AI *Quality & Safety for LLMs*, *Red Teaming LLM Apps*.

TOOLS · promptfoo · LangSmith · Ragas · Bedrock/Azure/Vertex native eval.

Sources: genai.owasp.org · nist.gov/itl/ai-risk-management-framework · atlas.mitre.org · promptfoo.dev.

CLOUD TRACK 1 OF 3 · AWS — BEDROCK + AGENTCORE + CERTS

PATH: AIF-C01 → MLA → GENAI PRO

AWS If your org already runs on IAM, this is your shortest path.

RUNG	AWS SERVICE TO LEARN	CERTIFICATION	SHIP ARTIFACT ON AWS
L0	PartyRock playground · Amazon Q chat	—	5 prompts on Q + a token-aware chat log
L1	Bedrock InvokeModel · Knowledge Bases · S3	**AWS Certified AI Practitioner (AIF-C01)**	A Bedrock KB over S3 + Lambda endpoint with citations
L2	AgentCore Runtime · Guardrails · CloudWatch	ML Engineer Associate (MLA-C01)	Bedrock Agent w/ tools + Guardrail + cost dashboard
L3	Custom Model Import · Distillation · Trainium	**GenAI Developer Professional** (2026)	Distilled model + red-team CI rig on Bedrock
L4	Well-Architected GenAI Lens · Org SCPs	Solutions Architect Pro (top-up)	Org-wide Bedrock policy doc + landing zone pattern

Free study path: AWS Skill Builder — Standard Exam Prep Plan: AWS Certified AI Practitioner + AWS Educate generative AI courses. Hands-on: AWS Workshops · Bedrock samples.

Sources: aws.amazon.com/certification · skillbuilder.aws · aws.amazon.com/bedrock · AWS AI certifications 2026 catalogue.

AI-102 RETIRES JUN 2026

AZ

If your org runs on Entra and M365, this is the natural path.

RUNG	AZURE SERVICE TO LEARN	CERTIFICATION	SHIP ARTIFACT ON AZURE
L0	Copilot Chat · AI Foundry playground	**AI-900** Azure AI Fundamentals (until Jun 2026)	5 prompts on Copilot + token-aware chat log
L1	Azure OpenAI · AI Search RAG · Blob	**AI-102** Azure AI Engineer (or successor)	AI Search + Azure OpenAI on your blobs — Container Apps deploy
L2	AI Agent Service · Content Safety · App Insights	DP-100 Data Scientist Associate (adjacent)	Agent w/ Entra OBO + Content Safety + cost dashboard
L3	AI Foundry fine-tune · AI Evaluation · Phi family	AI-102 successor (Foundry-aligned, 2026)	Fine-tuned Phi + LLM-as-judge eval gating deploys
L4	Azure Well-Architected GenAI Workload	AZ-305 Solutions Architect Expert (top-up)	Org-wide Foundry policy + landing zone for AI

Note: Microsoft retires AI-102 + AI-900 on Jun 30, 2026 — check Microsoft Learn for the Foundry-aligned successor before booking your exam.

Sources: learn.microsoft.com/credentials/certifications · ai.azure.com · Microsoft AI-102 retirement notice (Jun 2026).

GCP If you want native multimodal + 1M-token context, this is the lane.

RUNG	GCP SERVICE TO LEARN	CERTIFICATION	SHIP ARTIFACT ON GCP
L0	Gemini app · AI Studio playground	**Generative AI Leader** (non-technical, 2026)	5 prompts across Gemini Pro/Flash + token log
L1	Vertex AI · Vector Search · Cloud Storage	**Cloud Digital Leader** + Vertex skill badges	Vertex RAG on a Cloud Storage corpus — Cloud Run deploy
L2	Vertex Agent Builder · ADK · Gen AI Eval	**Professional Cloud Developer**	ADK agent + Gen AI Eval Service + Cloud Logging
L3	Vertex Tuning · Model Garden · TPU v5p	**Professional ML Engineer**	Tuned Gemma + safety policy + canary rollout
L4	Vertex AI Architecture Center · Org policy	Professional Cloud Architect (top-up)	Org-wide Vertex policy + landing zone blueprint

Free study path: Google Cloud Skills Boost — Generative AI Leader Learning Path + Vertex AI quests. Hands-on: codelabs.developers.google.com.

Sources: cloud.google.com/learn/certification · cloud.google.com/learn/certification/generative-ai-leader · skills.google · cloud.google.com/vertex-ai.

The five traps. Pick one you're in. Then close this deck.

#	THE TRAP	WHY IT FEELS PRODUCTIVE	THE EXIT
01	**Tutorial purgatory.** Course after course, never ship.	Each course feels like learning. None forces a deploy.	Set a 2-week deadline for the L1 artifact. Cancel courses that don't fit it.
02	**Paper-first stalling.** Reading transformers math before touching an API.	Looks rigorous. Builds zero feel for what breaks in practice.	Ship one prompt + one API call first. Papers AFTER your first wrong answer.
03	**Premature cloud commitment.** Picking AWS/Azure/GCP before L1.	Feels strategic. You learn the cloud, not the model.	Ship L1 on any of the three. Pick cloud after the first real workload.
04	**Fine-tune fixation.** Trying to fine-tune at L1.	"Customization" sounds advanced. 80% of fine-tunes were RAG problems.	Prompt → RAG → agent. Fine-tune only at L3, with a specific eval delta.
05	**No eval, no progress.** "It works on the demo" as quality bar.	Each "fix" feels like progress. You're walking sideways.	30 golden Q&A pairs. Re-run on every change. Don't ship without them.

Field pattern across DeepLearning.AI cohort threads, HF agents-course discussion, r/LocalLLaMA (2025-2026).

2-WEEK ARTIFACT

Which artifact are you shipping in the next 14 days?

(a) · L0 → L1

The first RAG, hosted.

20 docs, vector store, one endpoint, citations + token log. Where most learners earn their first rung.

(b) · L1 → L2

Eval harness + cost dashboard.

30 golden Q&A, run in CI. My bet — this is where most teams quietly transition from "demo" to "system."

(c) · L2 → L3

Pick one depth. Ship the proof.

Fine-tune, agents, eval, or red-team. One of the four. The artifact that earns the call.

My bet is (b). Pick yours — but the rung you're on is the one whose artifact you cannot yet ship.