

The Agent Engineer's Field Guide

10 production agent patterns — when to use each, what breaks first, and how to know it's working.

Imran Ahmad, PhD · 30 Agents Every AI Engineer Must Build (Packt, 2026) · Built with **Google ADK** · Runnable code for every pattern — no-API-key simulation + local Ollama: github.com/PacktPublishing/30-Agents-Every-AI-Engineer-Must-Build

FIVE RULES THAT SHOW UP IN EVERY CHAPTER

1. **An agent is a cognitive loop** — perception → cognition → action. LLM + tools + a loop. Learn the pattern first; then let a framework carry the plumbing (the workshop builds every agent with Google ADK).
2. **Pick a confidence threshold and gate autonomy on it.** 0.85 auto-approves an insurance claim; 0.79 routes it to a human.
3. **The prompt is a constitution, not an instruction.** Write it with PTCF: Persona, Task, Context, Format.
4. **Log everything into an immutable audit trail.** You cannot fix what you cannot see.
5. **Safety constrains the action space before planning; it never emerges from planning.** One unsatisfied constraint vetoes the whole plan.

1 Tool-Using Agent

CH 7

The pattern everything else is built on: the LLM reasons, selects from a tool registry (schemas are contracts), executes in a function-calling loop.

USE WHEN the task decomposes into calls on external functions/APIs with clear input-output contracts.

BREAKS FIRST semantic mismatch — the call *succeeds*, but the result misses the intent (chart sorted alphabetically, not by spend).

EVAL wrap every tool in timeouts + retry/backoff; a "failure memory" circuit breaker for dead tools; log every invocation.

COST unbounded retries against a dead tool burn tokens and latency. Cap attempts; escalate to a human path.

2 Knowledge Retrieval (RAG)

CH 6

Query understanding → chunk & embed → vector retrieval → grounded synthesis with provenance tracking.

USE WHEN answers must be grounded in current, verifiable documents — not the model's parametric memory.

BREAKS FIRST chunking misconfiguration — the most common source of retrieval-quality degradation in production. Defaults: 1000-char chunks, 200 overlap, k=3.

EVAL inspect returned source passages. Uniformly low similarity → re-ingest, add metadata filters, or go hybrid BM25+vector.

COST multi-stage retrieval buys recall with latency. Cache frequent queries; keep single-stage on the hot path.

3 Chain-of-Agents Orchestrator

CH 7

A supervisor routes specialist agents over a state machine — typed delegation, shared memory, formal arbitration. (Insurance claims: OCR → validation → risk → payout.)

USE WHEN the workflow exceeds one agent — heterogeneous sources, specialized skills, long-running processes.

BREAKS FIRST conflicting specialist outputs with no arbitration layer. Flag when the conflict score exceeds 0.5.

EVAL confidence-gated autonomy — 0.91 auto-approves, 0.79 goes to *Pending Human Review*. Every state transition logged: that log *is* your audit trail.

COST a human-in-the-loop gate on medium/high risk feels slow — and is far cheaper than one automated error.

4 Data Analysis Agent

CH 8

Plain-English question → generated Pandas/SQL → sandboxed execution → chart recommendation + statistical readout.

USE WHEN conversational analytics — moving a team from data literacy to insight literacy.

BREAKS FIRST keyword-rule intent classifiers snap on the first ambiguous query ("show me the data"). Production needs LLM intent classification with confidence scoring.

EVAL let statistics check the LLM — hypothesis tests, z-score anomaly detection ($|z| > 3$), confidence intervals in plain language. Keep the LLM out of the arithmetic.

COST generated code executes against real data. Sandbox it, or your agent becomes your biggest security hole.

5 Verification / Fact-Checking

CH 8

Extract atomic claims → retrieve evidence → NLI verdict (entail / refute / neutral) + recompute the arithmetic symbolically.

USE WHEN downstream of any generative or analytical agent whose errors propagate into decisions.

BREAKS FIRST confident hallucination — the summary cites a revenue figure hallucinated from a misread table, and nobody notices.

EVAL tolerance-based verdicts vs. a trusted store — Confirmed / Mostly True / Contradicted. 0.5 pp on rates, \$500K on monetary claims. Low confidence → human.

COST verification is a second LLM pass — run it as accept/revise/reject post-processing, not inline in the hot path.

6 Financial Advisory Agent

CH 14

A supervisor routes five specialists: onboarding, market intelligence, portfolio construction, risk monitoring, compliance.

USE WHEN regulated, suitability-bound advice. The goal is not to automate judgment — it is to structure it.

BREAKS FIRST speed without safeguards. Knight Capital lost ~\$440M in 45 minutes — the canonical warning.

EVAL compliance-by-architecture: enforcement in the supervisor, not the agents. Composite risk 0–10 (volatility, drawdown, 95% VaR), 25% concentration cap, revise→re-validate loop, immutable audit log.

PROOF response time 4 hours → 30 seconds; compliance accuracy 95% → 99.99%.

7 Healthcare Intelligence

CH 13

Strictly layered: ingestion / clinical knowledge / reasoning / explanation — Bayesian belief updating, calibrated confidence, safety monitor, signed audit trail.

USE WHEN clinical decision support where every action must be defensible to auditors, clinicians, and patients.

BREAKS FIRST false negatives — a missed MI or sepsis costs ~an order of magnitude more than a false alarm. Escalate at 0.15 probability; critical conditions escalate regardless of confidence.

EVAL calibration, not raw accuracy — 80% confidence should mean ~80% correct (Brier score). Append-only, cryptographically signed audit records, ≥7-year retention.

LATENCY tiered — sub-second reactive tier for drug interactions; deliberative tier returns top-2 differentials on timeout.

8 Education Intelligence

CH 15

Bayesian Knowledge Tracing (the belief state) + zone-of-proximal-development curriculum planner + SM-2 spaced repetition.

USE WHEN personalized tutoring at scale. It augments teachers; it does not replace them.

BREAKS FIRST rubber-stamp praise plus the correct answer — or generic advice that never connects to the student's actual misconception.

EVAL mastery threshold 0.85 before advancing; misconception-rule pipeline (~180 patterns, <50 ms, catches ~70%) with an LLM second stage below 0.7 confidence.

PROOF feedback latency 8 s → 2.5 s at >85% diagnostic accuracy.

9 Vision-Language Agent

CH 11

Visual encoder → projection into the LLM's token space → LLM cognitive core (adapters, cross-attention, or early fusion).

USE WHEN the model must directly perceive images — visual QA, grounding — not reason over textual summaries of them.

BREAKS FIRST confident visual hallucination. Cross-check VLM output against deterministic CV — an object detector verifies "three people."

EVAL greedy decoding for reproducibility, plus an accuracy-validation layer that catches confident errors before they propagate downstream.

COST visual tokens dominate the bill — 1024→336 px resolution cuts them ~90%. 7B+ models need float16 and ≥16 GB VRAM.

10 Embodied Intelligence — drone missions

CH 16

Asymmetric control loop: a slow LLM planner over a fast deterministic controller (task 0.1–1 Hz ... servo 1–10 kHz).

USE WHEN physical-world action where failure means material loss — and there is no transactional rollback.

BREAKS FIRST collapsing reasoning and control into one loop. A servo loop at 1 kHz cannot pause to query a weather API.

EVAL a Unified Constraint Envelope — go/no-go across five domains, conservative fusion: one unsatisfied domain vetoes the mission. The system can decline to arm, with an explanation.

PROOF –10 °C / 25 km/h limits; battery ≥30% departure, ≥20% per waypoint, 15% reserve; human in the perimeter → halt <100 ms.

Why these ten?

They sequence the whole discipline: how a single agent thinks (1–2), how agents work together (3–5), how specialists survive regulated domains (6–8), and how agents perceive and act in the physical world (9–10). Master these and the other twenty patterns in the book are variations, not mysteries.

DEEPER each card maps to a full chapter: case study, design variations, integration notes, common pitfalls.

RUNNABLE every pattern ships as a pre-executed notebook — read the outputs on GitHub without running anything, or run locally with Ollama, no API key.

Build all ten, live, in one day — 12 September, 11 AM ET

10 Essential AI Agents Every Engineer Must Build — a 6-hour, build-along Packt workshop. Every attendee builds all 10 agents with **Google ADK** and ships working code: OpenAI, Claude, Gemini, local Ollama, or the no-key simulation mode. Bring a laptop with Python 3.10+ and Git; leave with 10 working agents, the ebook *30 Agents Every AI Engineer Must Build*, and a certificate of completion. **40% off — code IMRAN40**

Register: eventbrite.co.uk/e/10-essential-ai-agents-every-engineer-must-build-tickets-1992470369511?aff=imran