pronix.ai
An AI-first business of Pronix Inc.

PLAYBOOK · PNX·PBK·809

Building your first production-grade agentic workflow

A field-tested blueprint for shipping your first agentic AI workflow into production — the same one we use with Fortune 500 clients. Covers intent boundaries, tool design, memory, guardrails, human-in-the-loop patterns and evaluation harnesses so your first agent survives real users, real data and real audits.

PREPARED FOR
CIO, Head of AI, VP Engineering, Head of CX Platforms

PREPARED BY
pronix.ai Strategy Practice, Enterprise AI & CX advisory

EDITION
Updated Q1 2026 · 18 min read · 34-page PDF

STRATEGY · IMPLEMENTATION · TALENT pronix.ai/resources

EXECUTIVE SUMMARY

Building your first production-grade agentic workflow

A field-tested blueprint for shipping your first agentic AI workflow into production — the same one we use with Fortune 500 clients. Covers intent boundaries, tool design, memory, guardrails, human-in-the-loop patterns and evaluation harnesses so your first agent survives real users, real data and real audits.

6-layer	12	4×
Reference architecture	Example tool schemas	Faster time-to-production vs ad-hoc builds

What you'll take away

- How to scope an agentic workflow so it delivers business outcomes, not demos
- The 6-layer reference architecture: intent, tools, memory, orchestration, guardrails, evaluation
- How to design tools and function contracts LLMs can actually call reliably
- Human-in-the-loop patterns for regulated and high-stakes decisions
- An evaluation harness you can run in CI — with pass/fail thresholds
- The org model — product, platform, and safety — that keeps agents in production

Audience: CIO, Head of AI, VP Engineering, Head of CX Platforms

CONTENTS

What's inside

SECTION 01

Why 90% of agentic pilots don’t ship

Pilots die at the seam between the model and the enterprise: brittle tools, weak evaluation, unclear ownership. The playbook opens by mapping the seven failure modes we see across 400+ enterprise programs and the architectural moves that neutralize them.

In our engagements, the practical work in this section usually breaks down into three deliverables: how to scope an agentic workflow so it delivers business outcomes, not demos.

Enterprise outcomes here come from production discipline, not slideware. Named owners, shipped guardrails, evaluation harnesses in CI and instrumented unit economics turn a promising pilot into a portfolio the board is willing to fund again next year.

SECTION 02

The 6-layer agentic reference architecture

A reference stack — intent boundary, tool layer, memory, orchestration, guardrails, evaluation — with concrete choices for each layer across OpenAI, Anthropic, Bedrock, Azure AI Foundry and Kore.ai Agent Platform. Each layer includes the anti-patterns to avoid and the SLOs to measure.

In our engagements, the practical work in this section usually breaks down into three deliverables: the 6-layer reference architecture: intent, tools, memory, orchestration, guardrails, evaluation; how to scope an agentic workflow so it delivers business outcomes, not demos.

Treat this as a reference, not a recipe. The layer boundaries matter more than the specific products you pick — a clean seam between orchestration, tools, memory and evaluation is what lets you swap providers as pricing, model quality and regulatory posture shift over the next 18 months.

WHAT GOOD LOOKS LIKE

The 6-layer reference architecture: intent, tools, memory, orchestration, guardrails, evaluation

SECTION 03

Designing tools LLMs can actually use

Function contracts, argument shapes, idempotency, retries, side-effect boundaries, and the ’1-verb / 1-noun’ rule that turns a flaky agent into a reliable one. Includes 12 example tool schemas from live production workflows.

In our engagements, the practical work in this section usually breaks down into three deliverables: how to design tools and function contracts llms can actually call reliably; the 6-layer reference architecture: intent, tools, memory, orchestration, guardrails, evaluation.

Treat this as a reference, not a recipe. The layer boundaries matter more than the specific products you pick — a clean seam between orchestration, tools, memory and evaluation is what lets you swap providers as pricing, model quality and regulatory posture shift over the next 18 months.

SECTION 04

Memory that scales without leaking

Session, task, entity and long-term memory — where each belongs, how to bound it, and how to keep PII out of the vector store. Includes retrieval evaluation recipes.

In our engagements, the practical work in this section usually breaks down into three deliverables: the 6-layer reference architecture: intent, tools, memory, orchestration, guardrails, evaluation; the org model — product, platform, and safety — that keeps agents in production.

Enterprise outcomes here come from production discipline, not slideware. Named owners, shipped guardrails, evaluation harnesses in CI and instrumented unit economics turn a promising pilot into a portfolio the board is willing to fund again next year.

WHAT GOOD LOOKS LIKE

Human-in-the-loop patterns for regulated and high-stakes decisions

SECTION 05

Guardrails, HITL and safety

Policy layers (input, tool, output), red-teaming, HITL routing, escalation UX and audit trails. Regulator-ready patterns for BFSI, healthcare and public sector.

In our engagements, the practical work in this section usually breaks down into three deliverables: the 6-layer reference architecture: intent, tools, memory, orchestration, guardrails, evaluation; the org model — product, platform, and safety — that keeps agents in production.

Governance that ships is architectural, not a policy PDF. Every enforcement point should be a callable service with test coverage, and every high-risk decision should have a named human owner, a documented review SLA, and an audit trail that a regulator or internal auditor can pull without a ticket.

SECTION 06

Evaluation harness in CI

Golden sets, LLM-as-judge, task success rate, cost per resolution, hallucination bounds. YAML config you can drop into GitHub Actions.

In our engagements, the practical work in this section usually breaks down into three deliverables: an evaluation harness you can run in ci — with pass/fail thresholds; the 6-layer reference architecture: intent, tools, memory, orchestration, guardrails, evaluation.

Evaluation is engineering, not vibes. Curate golden sets that reflect real traffic distributions, run them in CI on every model or prompt change, and hold the line on task success rate, hallucination bounds and cost per successful resolution — not just aggregate accuracy.

WHAT GOOD LOOKS LIKE

The org model — product, platform, and safety — that keeps agents in production

SECTION 07

The operating model

Product owner, platform team, safety review board — the RACI that keeps agents shipping weekly without breaking production.

In our engagements, the practical work in this section usually breaks down into three deliverables: the org model — product, platform, and safety — that keeps agents in production.

Structure follows outcome. Product teams own the business result, a single platform team owns shared tools and infrastructure, and a small safety and evaluation function reports independently. This is the org shape we see in every top-quartile program regardless of industry or platform choice.

FREQUENTLY ASKED

Questions from enterprise buyers

Is this framework tied to a specific LLM provider?

No — it’s provider-agnostic. Concrete examples span OpenAI, Anthropic, AWS Bedrock, Azure AI Foundry, Google Gemini and Kore.ai Agent Platform. The reference architecture assumes you’ll route across at least two providers for resilience and cost control.

How long does a first production agent typically take?

With this blueprint, our clients ship the first production agent in 8–14 weeks. Without it, 6–9 months is more typical because teams re-discover the same failure modes.

Does it cover regulated industries?

Yes — HITL, audit, red-teaming and evaluation patterns are written for BFSI, healthcare and public-sector use. Governance sections were reviewed with clients under applicable regulatory obligations.

TURN THIS INTO A DECISION

Book a 30-minute working session.

Bring your platform, data, seats and delivery constraints. A pronix.ai strategy or CX engineering lead will walk this document through with your team and adapt it — no slideware, no generic playbook.

ENTERPRISE SALES	hello@pronix.ai	pronix.ai/contact
PARTNERSHIPS & BPO	partners@pronix.ai	pronix.ai/company/partners
TALENT SOLUTIONS	talent@pronix.ai	pronix.ai/talent