THE AI RUNTIME · FIELD GUIDE

The Model Is the Smallest Part.

A guided reading path through The AI Runtime. Sixteen published pieces, four modules, one operating thesis for anyone shipping AI into production.

The model does the reasoning. The harness around it decides whether it ships, stays reliable, and stays defensible. That is where the work lives.

Curated for AI engineers, architects, and builders

theairuntime.com

START HERE

Why this guide exists

The AI Runtime covers one idea from many angles: in production, the model is the smallest part of the system. Reliability, value, and defensibility come from what wraps it. The harness around the model, the context it is given, the evaluations that gate it, and the identity it runs under.

This is a path through that idea. Sixteen published pieces, arranged into four modules. Read them in order to build the full mental model, or jump straight to the module that matches the problem on the desk this week.

How to read it

- Module 01 installs the operating thesis. Begin there.
- Modules 02 through 04 go deep: architecture teardowns, the operational layer, and the security frontier.
- Every entry names what it teaches and the framework it establishes. Follow the link to the full piece, with citations, on theairuntime.com.

Three pillars run through the path: **Model Reliability Engineering**, **Vertical Agents**, and **Lessons from the Trenches**.

MODULE 01

The Operating Thesis

Three pieces that install the mental model. Read these first.

01

MODEL RELIABILITY ENGINEERING INTRODUCES MRE

Model Reliability Engineering: Who Owns It When the AI Is Confidently Wrong?

Teams know their AI can be wrong. This installs the engineering discipline that makes it reliably right, and answers who owns the failure when it is not.

Read on theairuntime.com ↗

02

MODEL RELIABILITY ENGINEERING

Your AI Strategy Doesn't Need More Use Cases. It Needs a Production System.

Most enterprise AI strategies stall at the same point. Here are the five decisions that separate companies shipping AI products from companies running endless pilots.

Read on theairuntime.com ↗

03

VERTICAL AGENTS INTRODUCES VAA

The Anatomy of a Production Vertical Agent

Seven layers wrap every LLM that has shipped in healthcare, banking, and insurance. The model is the smallest of them. This names all seven.

Read on theairuntime.com ↗

MODULE 02

The Harness, Torn Down

Five production agents, opened up. What wraps the model is what actually ships.

01

VERTICAL AGENTS

Felix Is a Harness, Not a Model: How Rogo Built an Agent for High Finance

Rogo raised a $160M Series D led by Kleiner Perkins. Felix is the harness around the model, not the model itself, and the architecture is the lesson.

Read on theairuntime.com ↗

02

VERTICAL AGENTS

The Brain Isn't the LLM: How HockeyStack Built Revenue Agents

A $50M raise on a reasoning engine that is a custom ML pipeline, not a frontier model. Why that architectural choice matters for anyone building vertical agents.

Read on theairuntime.com ↗

03

VERTICAL AGENTS FEEDBACK ENGINEERING

How Vertical Agents Self-Improve in Production

Field notes on the harness loop at Harvey, Hippocratic, Anterior, and Azure SRE, where production failures compound into skill without retraining the model.

Read on theairuntime.com ↗

04

HARNESS ENGINEERING

Inside Mintlify's Agent Stack

A teardown of a two-harness architecture: async sandboxes for writes, virtual filesystems for reads. A concrete, copyable pattern for wrapping a model.

Read on theairuntime.com ↗

05 MODEL RELIABILITY ENGINEERING HARNESS TOPOLOGY

The Complete Field Guide to Browser Harnesses in 2026

Thirty plus harnesses, four topologies, one collapsing abstraction layer. The canonical map of how autonomous browser agents are actually built today.

Read on theairuntime.com ↗

MODULE 03

Context, Evals, and the Cost of Running It

The operational layer. What the model knows, how it is measured, what it costs to run.

---

01 MODEL RELIABILITY ENGINEERING CONTEXT ENGINEERING

Context Engineering for Code Agents: A Four-Level Spectrum

Deciding what the model knows about a codebase, its conventions, and the organization, laid out as a four level spectrum you can place your own system on.

Read on theairuntime.com ↗

---

02 MODEL RELIABILITY ENGINEERING

The Eval Lifecycle: What Actually Happens Between Proof of Concept and Production

Most AI projects die between it works on my laptop and it works in production. The eval lifecycle is the bridge across that gap, stage by stage.

Read on theairuntime.com ↗

---

03 MODEL RELIABILITY ENGINEERING SKILLOPS

PromptOps Is Dead, Long Live SkillOps

The shift from managing prompts to governing skills is the most consequential ops change in agentic AI. Most teams are already behind on it.

Read on theairuntime.com ↗

04

COST AND OPS

You're Paying 10x Too Much for LLM Inference

A practitioner guide to prompt caching across OpenAI, Anthropic, and Google, the single biggest lever for cutting cost and latency in production AI.

Read on theairuntime.com ↗

MODULE 04

The Security Frontier

Agent identity is the reliability surface nobody owns yet. Lessons from the Trenches.

01

LESSONS FROM THE TRENCHES AGENT IDENTITY

Shadow AI Agents

Your enterprise already has more agents than employees, and most have no identity, owner, or audit trail. Agent identity is the surface to claim before it claims you.

Read on theairuntime.com ↗

02

LESSONS FROM THE TRENCHES

The Vercel Breach RCA: Agent Identity Is the New Attack Surface

One OAuth grant, one compromised AI vendor, one platform breach. A root cause analysis to read against your own agent deployment, line by line.

Read on theairuntime.com ↗

03

LESSONS FROM THE TRENCHES

MCP Servers Are the Next Shadow Surface

Tool descriptions are now executable instructions, and the agent dependency graph runs through hundreds of unvetted servers. The next shadow surface, mapped.

Read on theairuntime.com ↗

BONUS TRACK

Prove You Can Build It

One more piece, for the reader who wants to be hired or promoted on this work.

★ CAREER INTRODUCES AIFOLIO

Your Portfolio Website Won't Get You Hired. Your AIfolio Will.

The to-do app is dead. The new portfolio playbook for developers entering the AI era, and the proof that you have made the leap.

Read on theairuntime.com ↗

KEEP READING

New teardowns, every week.

The AI Runtime publishes original architecture teardowns, named frameworks, and production postmortems for practitioners. Subscribe free to get each piece as it ships, across all three pillars.

The model does the easy part. The harness decides whether it survives production. That is the moat.

Model Reliability Engineering · Vertical Agents · Lessons from the Trenches

---

SUBSCRIBE FREE AT

theairuntime.com

The AI Runtime