READER RESOURCE · 02 OF 02 · ONE PAGE · CC BY-NC-ND 4.0

SERIES CHEAT SHEET

Production is not deployment. It is the architecture of trust under load.

AUTHOR: NABEEL A. KHAN · THREE BOOKS, THREE REFERENCE ARCHITECTURES, NINE RESEARCH ANCHORS · COMING SEPTEMBER 2026 · NABEELKHAN.COM/SERIES-RESOURCES

Three books, one company, one discipline observed from three altitudes. Set inside Nebula Financial, a regulated fintech running three systems, each owning a layer of one stack. A single fraud signal traces through all three and proves why each layer must exist: NexusCore routes it, AgentMesh investigates it, ThinkFlow ships the fix.

THE THREE BOOKS AT A GLANCE

	BOOK	SYSTEM	LAYER	PRIMARY READER	WHAT IT TEACHES
Book 1	*LLM Systems in Production*	NexusCore	Infrastructure / SRE	SREs, cloud architects, platform engineers entering AI	How a request finds the right model under a latency, cost, and risk budget, and is logged as evidence.
Book 2	*Prompt Systems & Agent Orchestration*	AgentMesh	Application / Agents	AI engineers, application developers, agent builders	How a pile of prompts becomes governed agents that plan, act, verify, and submit to human review.
Book 3	*DevOps for AI-Native Platforms*	ThinkFlow	Operations / Platform	DevOps and platform teams, internal-platform owners	How a platform builds, ships, and governs the models, agents, and code the other two layers depend on.

THE THREE REFERENCE ARCHITECTURES IN BRIEF

NEXUSCORE · BOOK 1

Routing and observability gateway

- **Data.** Gateway to Routing Brain to Latency Controller to the governed model pool (on-prem, cloud, frontier).
- **Control.** The Routing Brain chooses the smallest model a request can trust; the Latency Controller earns its SLO with speculative decoding.
- **Governance.** Routing and speculation policies as signed, versioned artifacts promoted through GitOps with a conformance gate.
- **Evidence.** Every routing decision, decode strategy, model version, and cost recorded as a queryable record.

The gateway is where an institution decides what it is allowed to think, and proves it thought it lawfully.

AGENTMESH · BOOK 2

Catalog of record and orchestration

- **Taxonomy and contracts.** Agents classified by domain, capability, and autonomy, each declaring a capability contract.
- **Four-module loop.** Planner decomposes, executor calls tools under contract, verifier checks output, generator responds.
- **Graph orchestration.** Workflows as activity-on-vertex graphs that can be inspected, parallelized, and refined.
- **Human review as runtime.** Tiered review (auto, lightweight, full supervisory) mapped to risk; feedback is first-class data.

An agent is not a clever prompt. It is a governed actor with a contract, a boundary, and a record.

THINKFLOW · BOOK 3

AI-native developer platform

- **Catalog and golden paths.** Services, models, and agents as first-class catalog citizens, with paved roads for AI workloads.
- **Policy-bounded delivery.** Agentic gates inside the pipeline, constrained by Delivery Guardrails and a trust-tier model; every decision logged.
- **RL-learned testing.** An adaptive agent chooses which tests to run, skip, or parallelize, bounded by per-service risk profiles.
- **PARA operations.** DevOps agents as perception, action, reasoning, reflection, each evaluated against benchmarks of real failures.

A platform is the paved road that decides what an organization can build without losing control.

THE NINE RESEARCH ANCHORS

#	BOOK	ANCHOR	CH.
1	NexusCore	Universal, workload-aware LLM routing	B1, 4
2	NexusCore	SLO-aware speculative and pipelined decoding	B1, 6
3	NexusCore	Secure and auditable router lifecycle	B1, 9
4	AgentMesh	Trainable planner-executor-verifier-generator loops	B2, 3
5	AgentMesh	Graph-based agent workflows with dynamic refinement	B2, 5
6	AgentMesh	Tiered human-in-the-loop orchestration	B2, 8
7	ThinkFlow	Policy-bounded AI-augmented CI/CD with trust tiers	B3, 4
8	ThinkFlow	Reinforcement-learned adaptive test selection	B3, 6
9	ThinkFlow	Perception-action-reasoning-reflection DevOps agents	B3, 7&8

THE SPINE IN FIVE CHAPTERS

1. **Book 1, Ch. 4.** The router selects a model worthy of the risk and logs the decision.
2. **Book 1, Ch. 12.** The routed request is handed forward as a workflow.
3. **Book 2, Ch. 3.** The workflow is decomposed across planner, executor, verifier, generator.
4. **Book 2, Ch. 8 & 11.** Tiered human review engages and the workflow produces an audit record.
5. **Book 3, Ch. 11.** The gap the investigation exposed becomes a new service the platform ships.

THE GOVERNANCE SPINE

TOGAF for architecture governance. DMBOK for data governance. Routing policies become governed artifacts with provenance and promotion. Agent catalogs become managed assets with lineage. Delivery decisions become audited events. Every book's Appendix E carries the TOGAF, DMBOK, and EU AI Act / NIST AI RMF alignment checklists.

> *Governance is not compliance. It is coherence made visible.*

FROM THE PAGE TO A SYSTEM

The template package.

Turn the series artifacts into governance-grounded implementation starting points, adapted to your stack and compliance obligations. **Scan the code or visit nabeelkhan.com/engagements** to start an engagement. The full guides and print-ready downloads live at nabeelkhan.com/series-resources.