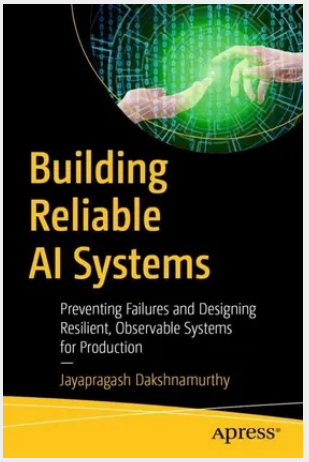
Dakshnamurthy

# Building Reliable AI Systems

Preventing Failures and Designing Resilient, Observable Systems for Production




**64,19 €**
64,19 € (zzgl. MwSt.)

*vorbestellbar, Erscheinungstermin ca. Januar 2027*

**Artikelnummer:** 9798868834097
**Medium:** Buch
**ISBN:** 979-8-8688-3409-7
**Verlag:** APRESS L.P.
**Erscheinungstermin:** 25.03.2027
**Sprache(n):** Englisch
**Auflage:** 1. Auflage 2027
**Produktform:** Kartoniert
**Seiten:** 491
**Format (B x H):** 155 x 235 mm

AI adoption has surged across industries, yet many production systems fail, not because of model accuracy, but due to weaknesses in the systems surrounding those models. While controlled environments often mask issues, real-world deployments expose complexities that emerge at scale: latency compounds, retries amplify failures, and dependencies trigger cascading breakdowns. Traditional observability dashboards may appear healthy, even as critical issues remain undetected. This book explores the often-overlooked gap between AI model success and production system reliability, focusing on the operational realities engineers face when AI moves from experimentation into mission-critical environments. Rather than emphasizing model development, the book treats AI systems as distributed systems and examines the unique failure patterns that arise at scale. It begins by analyzing why production systems fail and how latency amplification and retry storms can destabilize architectures. It then redefines observability beyond logs, metrics, and traces, introducing signal-based approaches and illustrating the importance of separating control and execution planes. Readers are guided through designing resilient AI architectures using microservices, circuit breakers, and fault isolation patterns, followed by deeper insights into event-driven systems, asynchronous processing, and idempotency. The book also addresses cascading failures, dependency risks, and techniques such as backpressure, rate limiting, and graceful degradation. Security is covered through practical patterns like token vaults and secure execution models. Real-world case studies reinforce these concepts, and a final hands-on chapter demonstrates implementation of reliability patterns including retry controls, circuit breakers, and observability instrumentation. By the end of this book, readers will understand how to build AI systems that remain reliable under stress, failure, and scale. They will gain practical strategies to design fault-tolerant, observable, and secure systems, supported by real-world lessons from enterprise-scale environments. This book equips engineers and architects with the tools needed to move beyond fragile deployments and create AI systems that not only perform but endure. What will you learn: - Design AI systems that remain stable under real-world production conditions - Identify and prevent hidden failure patterns such as retry storms and cascading failures - Implement production-grade observability beyond logs, metrics, and traces - Architect resilient microservices and event-driven systems for AI workloads - Apply production-grade reliability patterns used in large-scale distributed systems Who is it for: This book is intended for software engineers, backend developers, platform engineers, and system architects who are building or operating distributed systems and AI-enabled applications in production. It is especially valuable for engineers transitioning AI/ML workloads from experimentation to production environments and facing scalability and reliability challenges.

AI adoption has surged across industries, yet many production systems fail, not because of model accuracy, but due to weaknesses in the systems surrounding those models. While controlled environments often mask issues, real-world deployments expose complexities that emerge at scale: latency compounds, retries amplify failures, and dependencies trigger cascading breakdowns. Traditional observability dashboards may appear healthy, even as critical issues remain undetected. This book explores the often-overlooked gap between AI model success and production system reliability, focusing on the operational realities engineers face when AI moves from experimentation into mission-critical environments. Rather than emphasizing model development, the book treats AI systems as distributed systems and examines the unique failure patterns that arise at scale. It begins by analyzing why production systems fail and how latency amplification and retry storms can destabilize architectures. It then redefines observability beyond logs, metrics, and traces, introducing signal-based approaches and illustrating the importance of separating control and execution planes. Readers are guided through designing resilient AI architectures using microservices, circuit breakers, and fault