

## Observability for Large Language Models

Site Reliability and Chaos Engineering for AI at Scale

This book is a comprehensive guide designed to equip engineers, data scientists, and AI practitioners with the principles, tools, and strategies needed to ensure reliability, performance, and accountability in Large Language Models (LLMs).

The book begins by laying the groundwork with the foundations of observability, introducing LLMs, their significance in modern AI, and the critical role observability plays in maintaining robust systems. It then explores SRE principles, service level objectives, and incident response, while distinguishing the unique observability challenges that arise in AI and ML systems. Building on this foundation, the book dives into measuring performance, from defining SLOs tailored for LLMs to monitoring computational and token-level metrics. Readers gain practical insights into structured logging, debugging, and distributed tracing methods that provide visibility into complex LLM workflows. Scaling challenges are addressed through strategies for cross-model observability, autoscaling, latency reduction, and fault-tolerant infrastructure design. The book further explores chaos engineering, guiding readers through resilience testing in LLMs and the automation of chaos experiments in CI/CD pipelines. Finally, it highlights monitoring, retraining, and ethical considerations in AI observability, including governance, privacy, and accountability.

In conclusion, this book provides a holistic roadmap to building reliable, transparent, and future-ready LLM systems.

What you will learn:

- How to design observability pipelines for LLMs, including token-level logging, prompt tracing, and

latency analysis.

- Techniques for applying chaos engineering principles to test LLM robustness under stress and

failure scenarios.

- Methods for building SLOs, SLAs, and dashboards tailored to inference quality and model

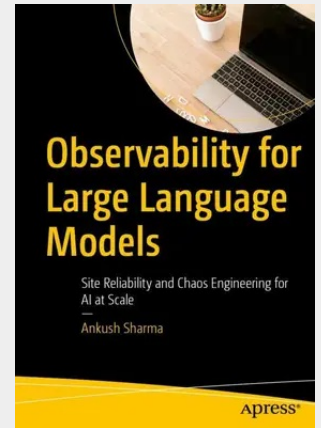
reliability.

- Strategies for monitoring hallucinations, drift, bias, and ethical failures in real-time.

Who this book is for:

This book is for AI infrastructure engineers, SREs, machine learning platform teams, and applied AI practitioners deploying or maintaining LLM-based applications.

This book is a comprehensive guide designed to equip engineers, data scientists, and AI practitioners with the principles, tools, and strategies needed to ensure reliability, performance, and accountability in Large Language Models (LLMs). The book begins by laying the groundwork with the foundations of observability, introducing LLMs, their significance in modern AI, and the critical role observability plays in maintaining robust systems. It then explores SRE principles, service level objectives, and incident response, while distinguishing the unique observability challenges that arise in AI and ML systems. Building on this foundation, the book dives into measuring performance, from defining SLOs tailored for LLMs to monitoring computational and token-level metrics. Readers gain practical insights into structured logging, debugging, and distributed tracing



**58,84 €**

54,99 € (zzgl. MwSt.)

*vorbestellbar, Erscheinungstermin ca. Juli 2026*

**Artikelnummer:** 9798868828263

**Medium:** Buch

**ISBN:** 979-8-8688-2826-3

**Verlag:** APRESS L.P.

**Erscheinungstermin:** 15.07.2026

**Sprache(n):** Englisch

**Auflage:** 1. Auflage 2026

**Produktform:** Kartoniert

**Seiten:** 236

**Format (B x H):** 178 x 254 mm

