Building Resilient Gen-AI Systems: Fault Tolerance and Recovery Patterns

Gaurav Bansal

Uttar Pradesh Technical University, India

Abstract*: This comprehensive exploration of resilient generative AI systems delves into the critical architecture, methodologies, and strategies required to ensure continuous operation in mission-critical applications. The article examines fault tolerance mechanisms and recovery patterns that form the foundation of reliable Gen-AI systems, beginning with robust detection systems, including distributed monitoring, comprehensive health checks, and ML-based predictive failure detection. It then analyzes essential recovery patterns such as graceful degradation, backup model deployment, state replication, and automated rollback capabilities. The article demonstrates how these resilience patterns translate into tangible benefits through real-world applications across healthcare, enterprise, and financial sectors. The implementation challenges of balancing redundancy against cost, testing failure scenarios, managing state complexity, and handling external dependencies are addressed with evidence-based best practices. By synthesizing cutting-edge research and industry experience, this article provides system architects and organizations with a practical framework for building Gen-AI applications that maintain operational integrity despite inevitable failures, establishing new standards for AI system reliability.*

Keywords: Generative AI resilience, fault tolerance, state management, graceful degradation, chaos engineering

Copyright to IJARSCT
www.ijarsct.co.in

DOI: 10.48175/IJARSCT-23955

ISSN
2581-9429
IJARSCT

338