PROJECT REPORT: AI-POWERED RECOMMENDATION SYSTEM

Project Overview:
Built a scalable AI-powered recommendation system for TechCorp Inc. that serves over 2 million daily active users.

Technical Implementation:
- Backend: Python with FastAPI framework
- Database: PostgreSQL for user data, Redis for caching
- AI/ML: TensorFlow for deep learning models, OpenAI GPT-4 for natural language processing
- Infrastructure: AWS EC2, S3, Lambda, and RDS
- Vector Database: Pinecone for similarity search
- Message Queue: Apache Kafka for real-time data processing

Architecture:
1. Data Collection Layer: Collects user behavior data and preferences
2. Feature Engineering: Processes raw data into ML-ready features
3. Model Training: Uses collaborative filtering and deep learning
4. Inference Engine: Real-time recommendation generation
5. API Layer: RESTful APIs for frontend integration
6. Caching Layer: Redis for fast response times

Key Features Implemented:
- Real-time personalized recommendations
- Content-based filtering using NLP
- Collaborative filtering for user similarity
- A/B testing framework for model evaluation
- RAG (Retrieval-Augmented Generation) for explanations
- Multi-armed bandit for exploration vs exploitation

Performance Metrics:
- Latency: Average response time of 50ms
- Throughput: Handles 10,000 requests per second
- Accuracy: 35% improvement in user engagement
- Scalability: Successfully scaled from 10K to 2M+ users
- Cost Efficiency: Reduced infrastructure costs by 25%

Code Quality:
- 95% test coverage with unit and integration tests
- Comprehensive documentation and API specs
- CI/CD pipeline with automated testing and deployment
- Code review process with senior engineers
- Monitoring and alerting with Prometheus and Grafana

Challenges Overcome:
1. Cold Start Problem: Implemented hybrid approach combining content and collaborative filtering
2. Scalability: Designed microservices architecture with horizontal scaling
3. Real-time Processing: Used Apache Kafka for streaming data processing
4. Model Drift: Implemented automated retraining pipeline
5. Data Privacy: Ensured GDPR compliance with data anonymization

Results and Impact:
- 35% increase in user engagement metrics
- 40% reduction in system latency
- 92% user satisfaction score
- Successfully handling 2M+ daily active users
- Became core product differentiator for the company

Future Enhancements:
- Integration with more data sources

- Advanced deep learning models
- Real-time model updates
- Enhanced explainability features
- Mobile app optimization

Technologies Used:
Python, FastAPI, TensorFlow, PostgreSQL, Redis, AWS, Docker, Kubernetes, Apache Kafka, Pinecone, OpenAI GPT-4, Prometheus, Grafana, GitHub Actions

Project Duration: 8 months (2023-2024)
Team Size: 5 engineers (led the technical implementation)