**Project Proposal: RAG Application Development**

**Project Overview**

A Retrieval-Augmented Generation (RAG) application enhances the responses of Large Language Models (LLMs) by explicitly retrieving relevant information from a domain-specific knowledge base, combining retrieval with generation for accurate, context-rich answers.

**Project Phases & Timeline**

|  |  |  |
| --- | --- | --- |
| Phase | Description | Duration |
| Requirements Analysis | Define use cases, scope, requirements | 2 weeks |
| Data Preparation | Gather, clean, and chunk documents for retrieval | 4 weeks |
| Model Selection | Choose embedding, LLM, and vector database solutions | 2 weeks |
| System Design | Architect pipeline, integration plans | 2 weeks |
| Development | Build ingestion, retrieval, LLM inference modules | 8 weeks |
| Testing/QA | Evaluate accuracy, security, scalability | 4 weeks |
| Deployment | Launch to production, monitor, optimize | 2 weeks |
| Documentation/Training | Project docs, user/admin training | 2 weeks |

**Total Duration:** 26 weeks (~6 months).[[1]](#fn1)

**Cost Estimates**

* **Basic RAG App:** $40,000–$200,000 (simple pipeline, small team, basic interface)
* **Advanced RAG Solution:** $600,000–$1,000,000+ (enterprise-grade, sophisticated infrastructure, complex data, extensive team, custom features)

**Operational Costs:**

* Vector Database: ~$25–$70/month
* LLM API: $0.0003–$0.0046/query[[1]](#fn1)

**Team Composition**

|  |  |  |
| --- | --- | --- |
| Role | Number of Members | Responsibilities |
| Project Manager | 1 | Planning, coordination, reporting |
| AI/ML Engineers | 2–4 | Model development, evaluation |
| Data Engineers | 1–2 | ETL, data preparation, storage |
| Backend Developers | 1–2 | Integrating components, API dev |
| DevOps Engineer | 1 | Deployment, monitoring, scaling |
| QA/Testers | 1 | Testing accuracy, edge cases, security |

**Typical Team Size:** 6–10 members for production-grade deployment.[[1]](#fn1)