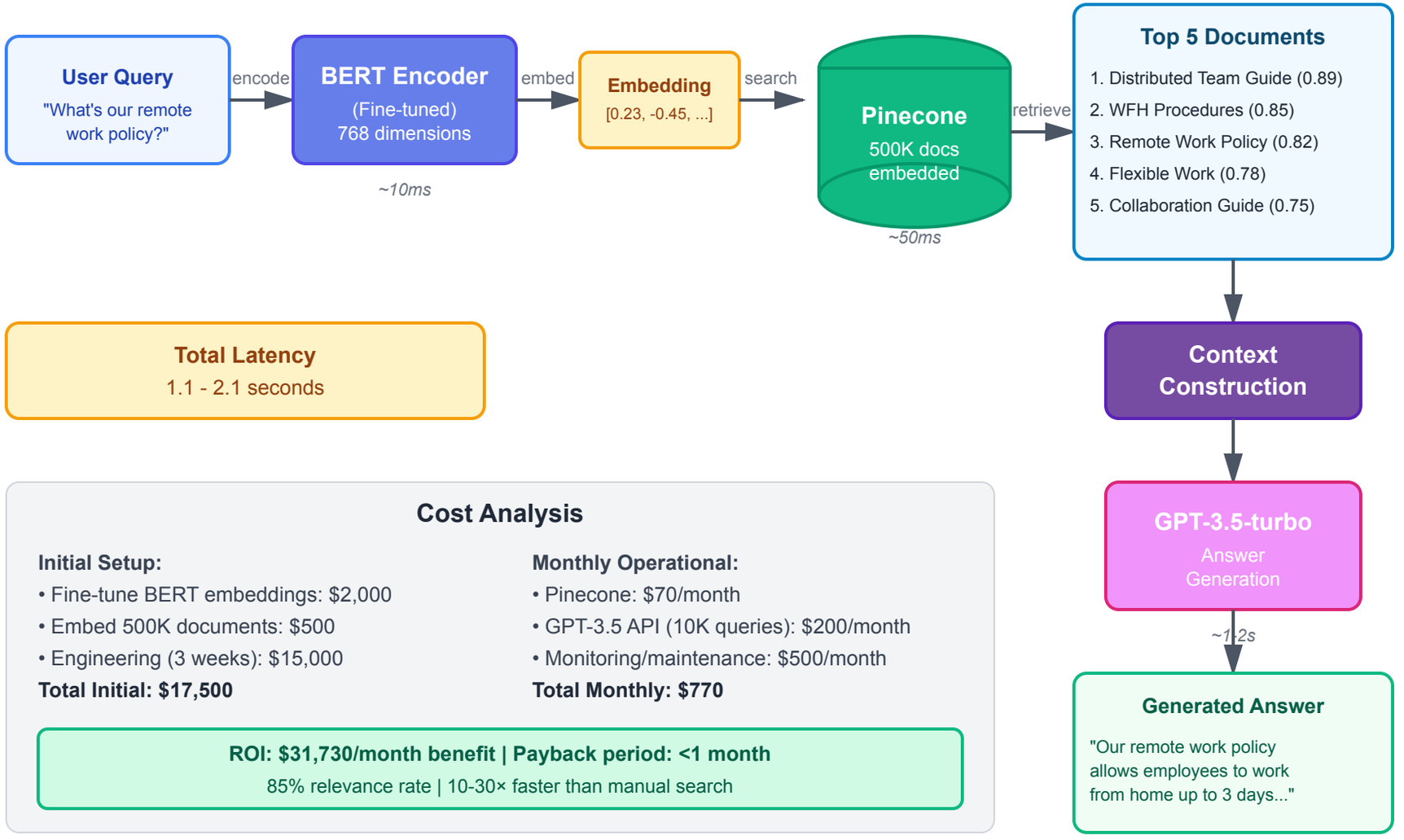
RAG Architecture for Semantic Search



Total Latency
1.1 - 2.1 seconds

Cost Analysis

Initial Setup:
- Fine-tune BERT embeddings: $2,000
- Embed 500K documents: $500
- Engineering (3 weeks): $15,000

Total Initial: $17,500

Monthly Operational:
- Pinecone: $70/month
- GPT-3.5 API (10K queries): $200/month
- Monitoring/maintenance: $500/month

Total Monthly: $770

ROI: $31,730/month benefit | Payback period: <1 month
85% relevance rate | 10-30× faster than manual search

Key Insight: RAG provides 80-95% cost savings vs fine-tuning while maintaining comparable accuracy.
Domain fine-tuning of embeddings ($2K) yields 17% accuracy improvement—worth the investment.