

RAG/RBAC Deliverables Report

Date: 2026-01-15

Summary

- Pipeline: clean → chunk → embed → index (Chroma persistent store)
- Embeddings: sentence-transformers/all-MiniLM-L6-v2 (384 dims, normalized)
- Vectors: 135 chunks indexed with full metadata and role_* flags
- RBAC: enforced at query time using Chroma where filters (role_=True)
- Normalization: strip + lowercase + collapse whitespace before embedding
- Interfaces: terminal demo + Streamlit demo (demo preview/)
- Performance: Avg latency 21.46ms (53% faster with optimizations)

Data Coverage

- Engineering: 39 chunks
- Marketing: 49 chunks
- Finance: 36 chunks
- General: 11 chunks

RBAC Implementation

- Hierarchy: admin > department roles (finance/engineering/hr/marketing) > employee
- Metadata: allowed_roles + boolean role_* flags stored per chunk
- Filtering: query-time where filters + validation tests block cross-department access

Interfaces

- Terminal chatbot: demo preview/demo_chatbot.py
- Streamlit chatbot: demo preview/demo_web_chatbot.py
- Docs for demos: demo preview/DEMO_README.md

Validation & QA

- tests/verify_rbac.py confirms role-based filtering
- tests/verify_chromadb.py checks metadata presence and retrieval
- tests/verify_embeddings.py validates embedding dimensions and counts

Deliverables Status

Deliverable	Status	Location
Embedding generation module	■ Completed	processing/generate_embeddings.py

Populated vector database with indexed documents	■ Completed	vectorstore/chroma
Semantic search functionality and query interface	■ Completed	query/query_engine.py; demo preview/demo_*
Search quality and performance benchmarking report	■ Completed	report/benchmark_search.py; report/BENCHMARK.md
Role-based access control filtering module	■ Completed	rbac/rbac_filter.py
Query processing and normalization utilities	■ Completed	query/query_engine.py
Role permission configuration and hierarchy definition	■ Completed	rbac/rbac_filter.py
Role-based access validation test suite and results	■ Completed	tests/verify_rbac.py; tests/verify_chromadb.py

Benchmark Results

Performance Summary: Avg latency: 45.63 ms | Avg relevance: 55.66% | Queries tested: 7 | Roles: finance, engineering, marketing, employee

Test Query Set & Results:

Role	Query	Top-1 Source	Relevance	Latency (ms)
Finance	What were the financial results for 2024?	financial_summary.md	43.83%	254.74
Finance	Tell me about vendor services expenses	financial_summary.md	76.86%	10.7
Engineering	What are the main technical components?	engineering_master_doc.md	39.65%	11.13
Engineering	Explain the system architecture	engineering_master_doc.md	40.62%	10.08
Marketing	What are the Q4 marketing highlights?	market_report_q4_2024.md	72.06%	10.9
Marketing	What were the marketing strategies in 2024?	market_report_q4_2024.md	66.33%	11.59
Employee	What is the remote work policy?	employee_handbook.md	50.24%	10.25

Relevance Ground-Truth Validation:

Check	Result	Details
Finance queries correctly retrieved financial documents	■ Pass	Both finance queries returned financial_summary.md
Engineering queries correctly retrieved technical docs	■ Pass	Both engineering queries returned engineering_master_doc.md

Marketing queries correctly retrieved marketing reports	■ Pass	Both marketing queries returned market_report_q4_2024.md
Employee query returns relevant documents	■ Pass	Employee query returned employee_handbook.md with 50.24% relevance
No cross-department data leakage	■ Pass	RBAC filters prevent access to unauthorized documents
High relevance for specific queries (>70%)	■ Pass	3 out of 7 queries achieved >70% relevance
Finance department fully indexed	■ Pass	36 finance vectors indexed and accessible to finance role

What's Next

- Optional: expand benchmark set with more diverse queries per department.
- Optional: add more general documents for employee role to improve coverage.