RAG Retrieval Platform Demo (Internal)

Date: 2026-03-01

Overview
- Internal grounding over runbooks, incident timelines, and config references
- Hybrid retrieval (dense embeddings + BM25 blending) with reranking hooks
- Citation grounding for auditability and operator trust
- ChatGPT-style UI with streaming and source panel

Data Sources Ingested (demo set)
- runbooks/: pod crash loop, DB connection pool exhaustion
- incidents/: API latency incident, OOM cascade incident
- configs/: kube-apiserver, prometheus scrape, nginx ingress values

Validation Results
- Full test suite: 190 passed
- API smoke test: /query returned status 200 with citations
- Chat backend smoke test: /api/chat returned status 200 with citations
- Local fallback model enabled when external API quota unavailable

Key Technical Enhancements
- Metadata-aware ingestion with source type classification (runbook/incident/config)
- Expanded ingest formats: md/txt/pdf/yaml/yml/json/ini/conf
- WebSocket done frame now returns citations + model metadata for UI rendering
- UI fixes for conversation creation and robust API error handling

End of Demo Report