K2 Chat

DOCUMENT RAG / ML INFRASTRUCTURE PROJECT BRIEF

A document-grounded workspace that separates interactive retrieval from asynchronous ingestion, with explicit tenant, version, and evidence boundaries.

Architecture

Interactive: Next.js client → Hono API on Cloudflare Workers → Cognito authentication, DynamoDB authority, Pinecone retrieval, and hosted Gemini generation.

Ingestion: S3 → SQS → Lambda for standard files; authorized files above 10 MiB use Step Functions and Glue/Spark. Uploads are capped at 100 MiB. AWS resources are defined in CDK.

Core engineering decisions

Authority before reuse. Versioned passages and cached private citations are checked against the current tenant-owned document record.

Job-bound state. Conditional writes bind processing to the original source, version, job, and workflow execution. Intermediate chunks are isolated by execution.

One embedding space. Documents and queries use 768-dimensional Gemini embeddings. Malformed vectors are rejected; another embedding model is not used as an automatic fallback.

Reproducible evaluation

The v2 synthetic corpus contains 11 documents, 30 source paragraphs, and 38 questions, including 8 unanswerable cases. SHA-256 seals source and label files. Evidence scoring covers partial recall, fixed-K precision, MRR, nDCG, citation validity, and abstention.

DEVELOPMENT LEXICAL REFERENCE	RESULT
Recall@5	91.7%
Precision@5	23.3%
MRR@5	0.697
Unanswerable abstention	0.0%

The reference uses 14 development questions and no model calls. It does not correctly abstain on either unanswerable development question. These figures describe a reference retriever, not generated-answer accuracy or deployed application performance.

Release scope

Model execution uses hosted APIs. Live ingestion requires configured AWS and provider access; older corpora created with a mixed embedding fallback require reindexing. The public website includes a clearly labeled scripted walkthrough and an engineering journal.