Runbook: Chat API End-to-End Troubleshooting (UI → API → RAG → LLM)

Last updated: 2026-01-23

Purpose

Diagnose issues from the web UI through API Gateway/Lambda to vector retrieval and LLM completion.

Common Symptoms

- Spinner never stops / request hangs
- UI receives 502/504
- Answer returned but has no citations or irrelevant content
- CORS failures in browser console

Step 1: Browser & Network

- Check DevTools Network: status code, response time, response body.
- Verify API_BASE_URL points to correct environment (dev/prod).
- Verify request payload includes selectedAgent and question text.

Step 2: API Gateway & Lambda

- Check CloudWatch logs for requestId and stack traces.
- Confirm Lambda timeout is sufficient (start at 15–30s).
- Validate env vars: S3_BUCKET, S3_PREFIX, embedding endpoint config.

Step 3: Vector Retrieval

- Log retrieval query, filters, top_k, and number of results returned.
- If results are empty, validate index exists and was built for this env.
- If results are irrelevant, adjust filters or chunking.

Step 4: LLM Completion

- Use system prompt requiring citations and explicit grounding.
- Enable retries for transient 429/5xx from LLM endpoint.
- Cap max_tokens to avoid timeouts.