Meeting Transcript: RAG Slack Chatbot Sync
Date: April 16, 2025
Time: 10:00 AM – 10:35 AM

10:00 AM – Laura Martínez (LM): Good morning, everyone. Today we need to review our progress on the RAG Slack chatbot integration, walk through the API updates, and outline our deployment steps.

10:02 AM – Mennatuallah Hefny (MH): Morning! I've drafted the Slack event subscription handler. I hit an edge case with Slack retrying duplicate events, so I'm adding idempotency checks in the middleware. I plan to finish the signature verification and idempotency logic by end of day Friday.

10:04 AM – Sarah Lee (SL): That's important. On the API side, the FastAPI endpoints for embedding and retrieval are up and running. I need to parameterize the Qdrant collection name—currently hardcoded as "engineering-transcripts"—so it reads from `.env`. I'll refactor that tomorrow.

10:05 AM – MH: Perfect. Once you do, I'll adjust my tests to point at the dynamic collection name.

10:06 AM – Ahmed Khan (AK): I'll spin up the CI pipeline to deploy on pushes to the `develop` branch. The staging Kubernetes cluster on AWS is ready; once your container images are tagged, I can deploy the Slack bot as its own microservice.

10:07 AM – SL: Good catch on versioning. I'll tag Docker images with the git commit hash and date so we can track releases.

10:08 AM – MH: And I'll update the Helm chart to pull images by those tags.

10:09 AM – AK: Let's also move our secrets into AWS Secrets Manager instead of plaintext in `.env`. I can integrate that with Kubernetes so they mount as environment variables.

10:10 AM – SL: I'll update the README and internal docs to reflect the new secrets workflow and include instructions on key rotation.

10:12 AM – LM: Great. Mennatuallah, could you draft a simple user-flow script for a Loom demo? We want to showcase a Slack slash command triggering the RAG pipeline end-to-end.

10:13 AM – MH: Sure. I'll write the script today and record the Loom video by early next week, then share it in our channel for feedback.

10:15 AM – AK: Meanwhile, I'll instrument Prometheus metrics in the FastAPI service—tracking request latency, throughput, and error rates—and set up Grafana dashboards. I expect to have basic dashboards by Thursday.

10:17 AM – SL: Once I finish refactoring, I’ll run load tests against the retrieval endpoint. If I see any bottlenecks, I’ll tweak the embedding batch sizes and report back.

10:18 AM – MH: After your load tests, I’ll validate retrieval accuracy under stress, making sure our relevance thresholds still hold.

10:20 AM – LM: Excellent. Ahmed, will the dashboards cover both API and Slack bot metrics?

10:21 AM – AK: Yes—I'll include separate panels for bot event handling times and API call latencies.

10:23 AM – SL: I’ll coordinate with you to add relevant Prometheus exporters if needed.

10:25 AM – MH: I also noticed occasional 500 errors when the embedding service is under load. I’ll add retry logic in the client and surface a friendly error message in Slack.

10:27 AM – LM: Good idea. Let’s make the user experience smooth even when things hiccup.

10:28 AM – AK: I’ll expose error rates in Grafana so we can spot those spikes quickly.

10:30 AM – LM: Fantastic. Let’s meet again next Wednesday at 10:00 AM to review the Loom demo and the staging deployment. Thanks, everyone!

10:32 AM – MH, SL, AK (simultaneously): Thanks! See you then.