RAG Chatbot Quick Start Guide

Chunking Settings

- Chunk size: Start at 1100 characters. Adjust between 800–1200 depending on documents.
- Chunk overlap: Start at 220. Increase if answers cut off mid-sentence.

Retrieval Settings

- Top_k (chunks sent to LLM): Start at 6. Increase to 8–10 if answers are incomplete.
- Similarity threshold: Default 0.3. Raise (0.4–0.5) to avoid irrelevant chunks. Lower (0.2) if missing answers.

App Modes

- Chatbot = main Q&A; interface.
- Admin Dashboard = analysis + logs (feedback, hard questions, export tools).

Feedback Loop

- Use ■ / ■ after each answer.
- Logs include question, answer, retrieved chunks + similarity scores.
- Negative feedback → saved into hard_questions.jsonl.

Debug Mode

- Enable 'Show Retrieved Chunks (Debug)' in sidebar.
- Expander shows chunks + scores, sources, previews.
- Download retrieved chunks as JSON/TXT.

Admin Dashboard Tools

- Feedback vs Similarity scatterplot: green = helpful, red = not helpful.
- Export logs & plots for analysis.
- Use hard_questions.jsonl to identify unresolved queries.

When Accuracy Drops

- Rebuild index if outdated.
- Adjust chunk size/overlap.
- Increase multi-query expansion (currently 5 variants).
- Tune similarity threshold up/down.

Advanced Tips

- Use BM25-heavy weighting (0.4 / 0.6) for keyword-rich docs.
- Lower BM25 weight for semantic/conceptual queries.
- Check debug mode to confirm the text exists in your index.