

## POLICYPULSE — FULL SYSTEM FLOWCHART (TEXT VERSION)

User -> Streamlit UI -> FastAPI -> RAG Engine -> Groq LLM -> Memory Manager -> Streamlit UI

Detailed steps:

1. User types a question in Streamlit UI.
2. Streamlit sends POST {user, question} to FastAPI (/ask).
3. FastAPI passes question to RAG Engine.
4. RAG Engine:
  - Normalizes input
  - Loads memory
  - Embeds query (MiniLM-L6-v2)
  - Searches Qdrant (HNSW, cosine)
  - Retrieves Top-k chunks
  - Builds context + system prompt
  - Sends prompt to Groq Llama-3.1
5. Groq LLM generates grounded answer.
6. RAG Engine post-processes answer + attaches sources.
7. Memory Manager saves interaction.
8. FastAPI returns final JSON.
9. Streamlit shows typing animation + full answer.