Building a RAG Chat: Full Pipeline GuideAuthor: Madhu DadiTags: AI, RAG, Architecture, FastAPI, ProductionSou

Step 1: The Embedding PipelineEvery published post is split into chunks and embedded. The chunks are stored in

- m = 16 â each node connects to 16 neighbors. Higher = better recall, slower build.16 is the sweet spot for datase

The user prompt includes the question and the chunk content:[code block]---Step 6: Source VerificationAfter the LL

---What’s NextIn the next post, I’ll cover the production RAG pipeline â streaming responses viaSSE, progressive