How this could scale in a real product like Verba AI

In a real production system like Verba AI, this FAQ Retrieval Assistant could be scaled into a robust multilingual customer-support solution.

Currently, the system works on a small dataset (20 FAQs) and performs semantic search using a sentence-transformer model and cosine similarity. To scale this to thousands or millions of FAQs across multiple clients, languages, and domains, several architectural improvements are needed.

First, embedding management would move to a vector database, something like Pinecone or even Postgres with pgvector. This would allow much faster similarity search, real-time indexing, and efficient handling of large-scale data. Instead of computing similarity in memory, the system would perform optimized nearest-neighbor search.

Second, the system would need scalable infrastructure. A production setup would likely involve:

- a backend service (FastAPI / Flask / Node)

- a deployed embedding model (or API like OpenAI / Azure)

- caching layer for repeated queries

- load balancing to handle multiple users simultaneously

Multilingual capability is essential, so the system would benefit from stronger multilingual embedding models and possibly language detection to route queries correctly. If needed, queries could be normalized or translated to improve retrieval consistency.

Another important element in production would be reliability and control. The assistant should not just return the highest-matching answer blindly. A confidence threshold (already implemented) would be extended so that:

- High-confidence → answer directly

- Medium-confidence → maybe ask clarification

- Low-confidence → fallback to human support

Finally, analytics and monitoring are critical. Metrics like Hits@1, accuracy, latency, and customer success rate would be tracked continuously. This enables improvement over time and ensures the system stays aligned with user needs.

Thank you for taking the time to read this document, I really appreciate it.

Yours truly, Hristijan Bogdanoski