Retrieval-Augmented Context Extension

Knowledge Sources:

- 📄 PubMed Articles
- 📋 Clinical Guidelines
- 💊 Drug Databases
- 🧬 Patient Records

User Query
"Patient history of diabetes"

Benefits:

✓ Unlimited effective context
✓ Always up-to-date info
✓ Lower inference cost
✓ Selective information
✓ Modular updates

Hybrid RAG Architecture

similarity search

Recent History (Direct)

Historical Data (Retrieved)

Last 3 months
In context window

Vector Database
📊 10+ years of records
Dense embeddings

| 2019 | 2020 | 2022 |

Fast access

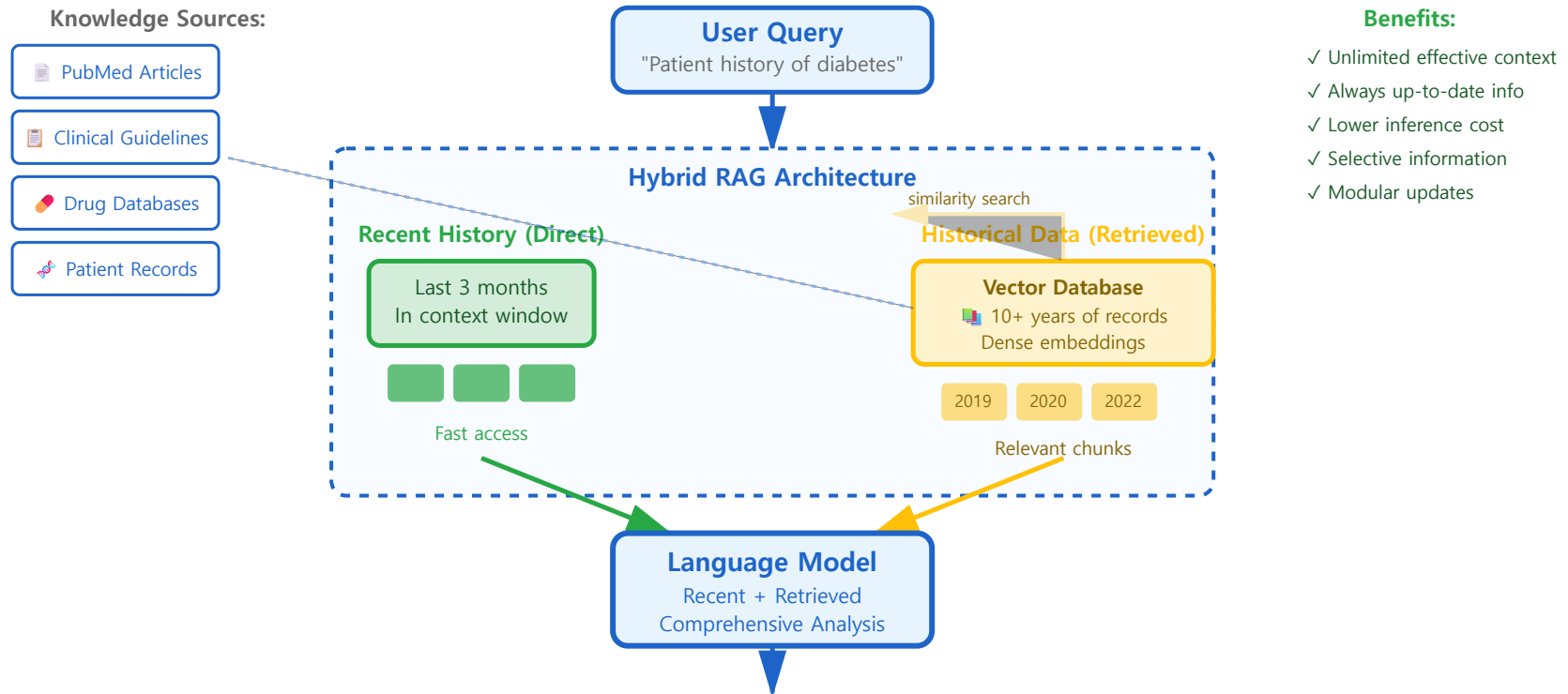Relevant chunks

Language Model
Recent + Retrieved
Comprehensive Analysis

🔍 Best of Both Worlds

RAG combines long-context models with retrieval systems, enabling access to vast medical knowledge while maintaining efficient inference and real-time updates. **Recent data stays in context, historical data retrieved on-demand.**