

Leveraging Vector Databases to Enhance AI Applications in Pharmaceutical Research

Presenters:

Angu Sundaresh Krishnakumar

Ramanathan Srinivasan

19, March-2025, Phuse , Orlando





Introduction

Background

- LLMs have revolutionized unstructured data processing
- Vector databases enable efficient data organization and retrieval
- Critical for pharmaceutical research and clinical trials
- Focus on improving protocol design

Key Discussion Points

- Vector Databases Overview
- System Architecture and Evaluation Framework
- Methodology: Tools and Metrics
- Findings
- Conclusions
- Future Directions



Vector Databases Overview

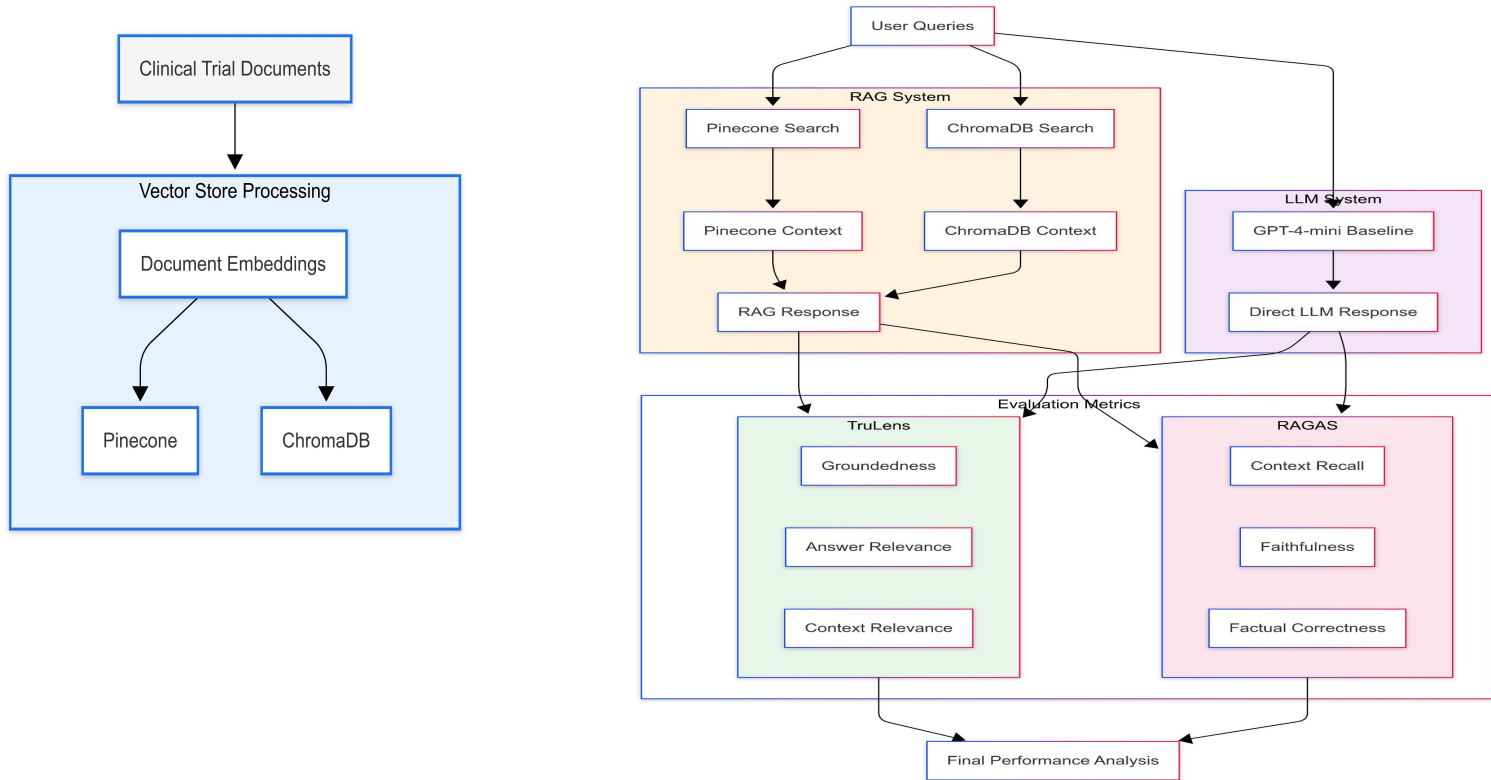
Key Characteristics

- **Optimized for Complex Embeddings:** Efficiently handles high-dimensional vectors
- **Real-time Similarity Search:** Advanced indexing for fast nearest neighbor lookups.
- **Deeper Insights from Any Data:** Find similar items across text, images, audio, etc.
- **Real-Time Results:** Deliver relevant recommendations and search.

Key Differences between Vector vs Traditional Databases

- **Flexible vs. Rigid Structures:** Handles complex data structures that traditional DBs can't.
- **Similarity vs. Exact Matching:** Finds semantically similar data, not just exact matches.
- **Meaning-Based Search:** Goes beyond keywords to understand context and intent.

System Architecture and Evaluation Framework





Methodology: Tools and Metrics

1. Tools

- Pinecone
- Chroma

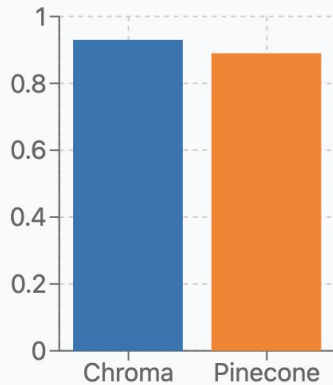
2. Metrics Overview

- TruLens Metrics
 - Groundedness: Uses facts from sources
 - Answer Relevance: Responds to actual question
 - Context Relevance: Found helpful information first
- RAGAs (Retrieval Augmented Generation Assessment) Metrics
 - Context Recall: Got all needed info
 - Faithfulness: Doesn't make things up
 - Factual Correctness: Facts are really true



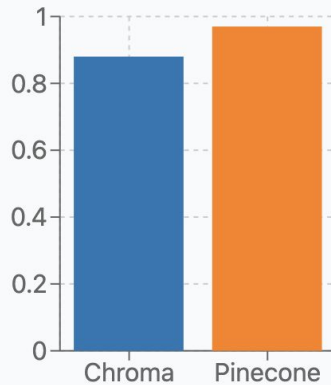
Findings - TruLens - Component-wise Analysis

Groundedness Comparison



Facts backed by sources

Answer Relevance Comparison



Responds to actual question

Performance Metrics

- Groundedness: Chroma (0.92) vs Pinecone (0.89)
- Answer Relevance: Pinecone (1.00) vs Chroma (0.88)



Findings - TruLens - Component-wise Analysis

Context Relevance Comparison



Found helpful information first

Average Latency Comparison
(seconds)



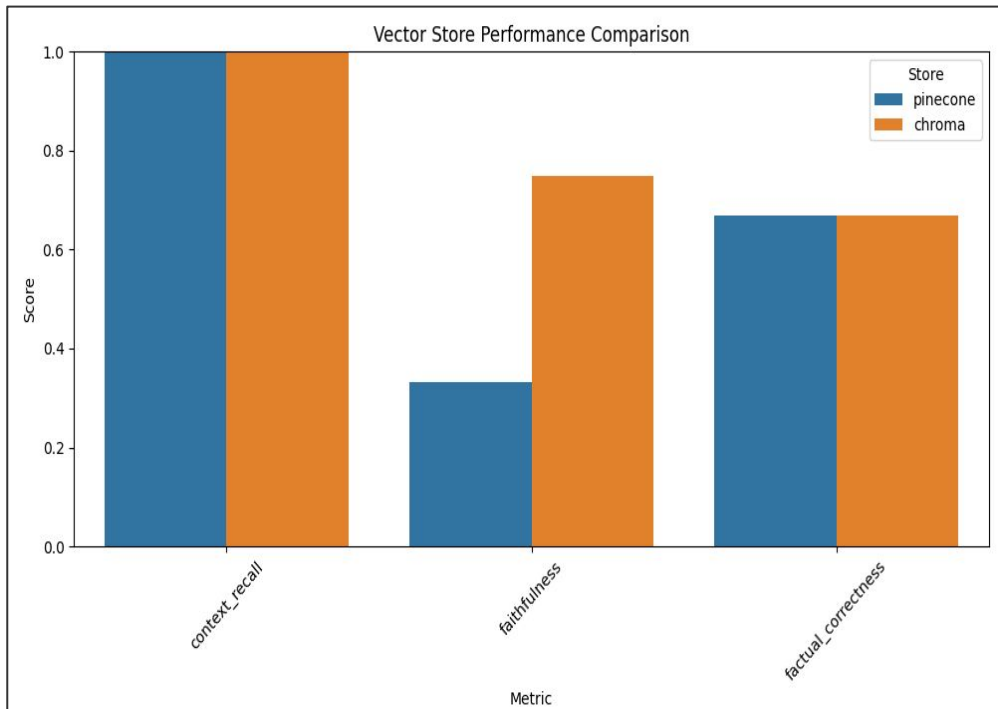
Lower is better

Performance Metrics

- Context Relevance: Pinecone (0.70) vs Chroma (0.60)
- Latency: Pinecone (0.45s) vs Chroma (0.60s)



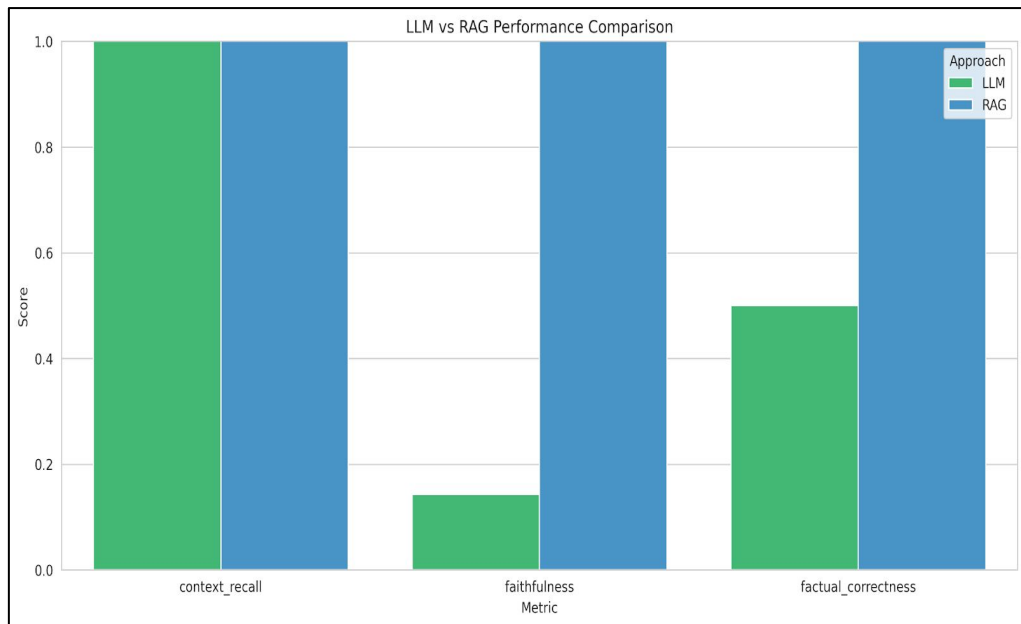
Findings - RAGAs - Vector Store Performance



Key Findings

- Perfect context recall (1.0) for both systems
- Chroma: Superior faithfulness (0.75)
- Pinecone: Lower faithfulness (0.33)
- Both systems: Similar factual correctness (0.67)

Findings - RAGAs- LLM vs RAG Comparison



Key Findings

- Perfect context recall in both systems
- RAG superior in faithfulness (~1.0 vs 0.15)
- RAG leads in factual correctness (~1.0 vs 0.5)



Findings - Response Latency And Tradeoff

- **RAG systems:** Better with clinical trial information but slower responses
 - Accesses specific trial data and research accurately
 - Takes extra time to search through trial documents
 - Great when using exact patient outcomes and drug details
- **Direct LLMs:** Faster answers with better flow but misses trial details
 - Gives quick responses without searching documents
 - Doesn't know your specific clinical trial information
 - Better when general medical knowledge is enough
- Pick what fits your medical needs best - **trial accuracy or speed**
 - Treatment decisions might need specific trial results
 - General medical questions might just need speed



Findings Summary

Category	Key Findings
Vector Store Performance	- Both Pinecone and Chroma demonstrate production-ready capabilities - Pinecone excels in speed and precision but shows lower faithfulness - Chroma offers better semantic consistency at the cost of higher latency
Overall System Efficacy	- Perfect context recall across implementations indicates robust information retrieval - High factual correctness in RAG implementation suggests viability for clinical applications - Response latency meets interactive use requirements
RAG Implementation	- RAG architecture substantially improves factual accuracy over pure LLM approaches - The trade-off between semantic faithfulness and factual correctness is evident.



Conclusions

Key Takeaways

- RAG implementation shows superior compare to direct LLM
- Both vector stores achieve production-grade performance
- Pinecone excels in speed and precision
- Chroma offers better semantic consistency



Future Directions

GraphRAG Implementation

- Enhanced relationship understanding
- Extension of traditional RAG capabilities
- Graph-based knowledge representations
- Improved clinical trial information retrieval

Thank You!

To learn more, email us at
angu.krishnakumar@saama.com
ramanathan.srinivasan@saama.com

