Retrieval Augmented Generation

RAG is a mechanism that enhances the quality of Large Language Models (LLMs) by connecting them to external knowledge sources in real-time.

Before generating a response, the system first retrieves relevant, up-to-date information and provides it to the LLM as context, which then augments the model's ability to generate a more accurate, detailed, and trustworthy answer.

Retrieval Augmented Generation

A vector database is a database designed to store and search through data as numerical representations called "**vector embeddings**"

Key Aspects of Vector Databases:

- Semantic Search
- Built for AI-Native Workloads
- Scalability and Speed

Popular Vector Databases:

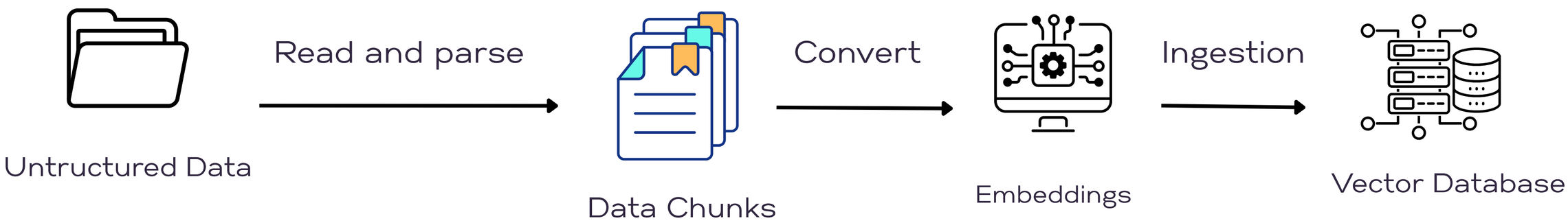
Pinecone

Chroma

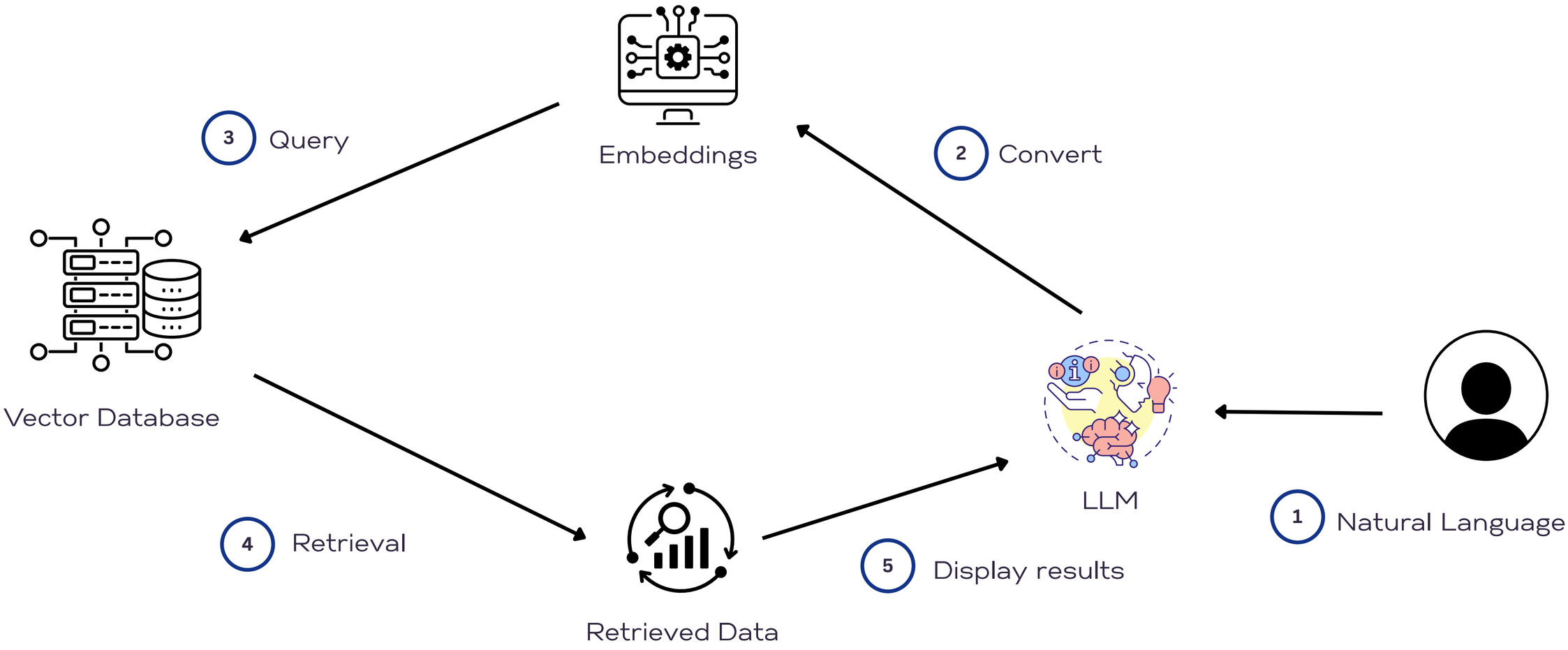
redis

elasticsearch

1 - Data Ingestion into Vector Stores



2 - Querying Vector Stores & Retrieval



Retrieval Augmented Generation

Why is RAG essential?

- Combats Hallucinations and Improves Factual Accuracy
- Access to Current and Real-Time Information:
- Use of Proprietary and Domain-Specific Data
- Increases Transparency and Trustworthiness

Retrieval Augmented Generation

RAG is Everywhere: The Core Pattern

- Classic RAG
- Function Calling & Tools
- AI Agents
- Multi-Agent Systems