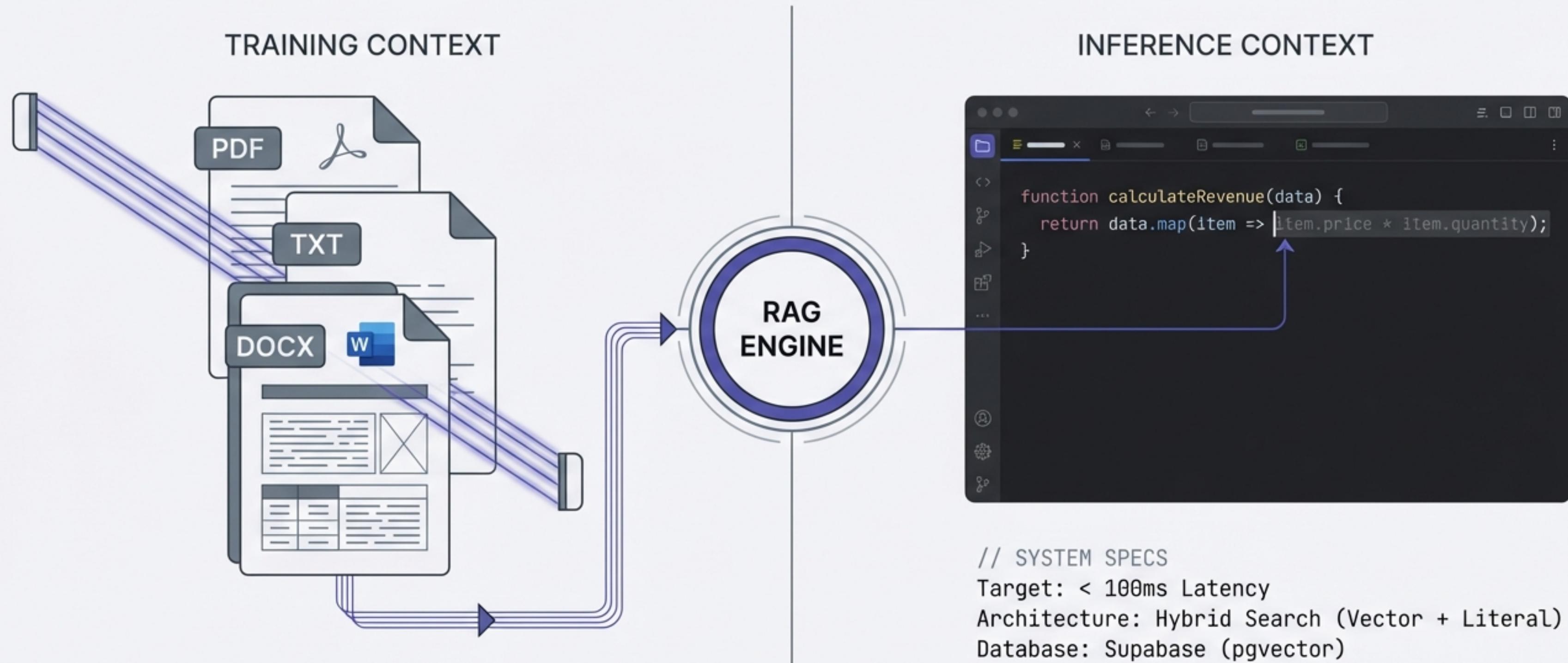


# TypeFlow AI: Intelligent Autocomplete Architecture

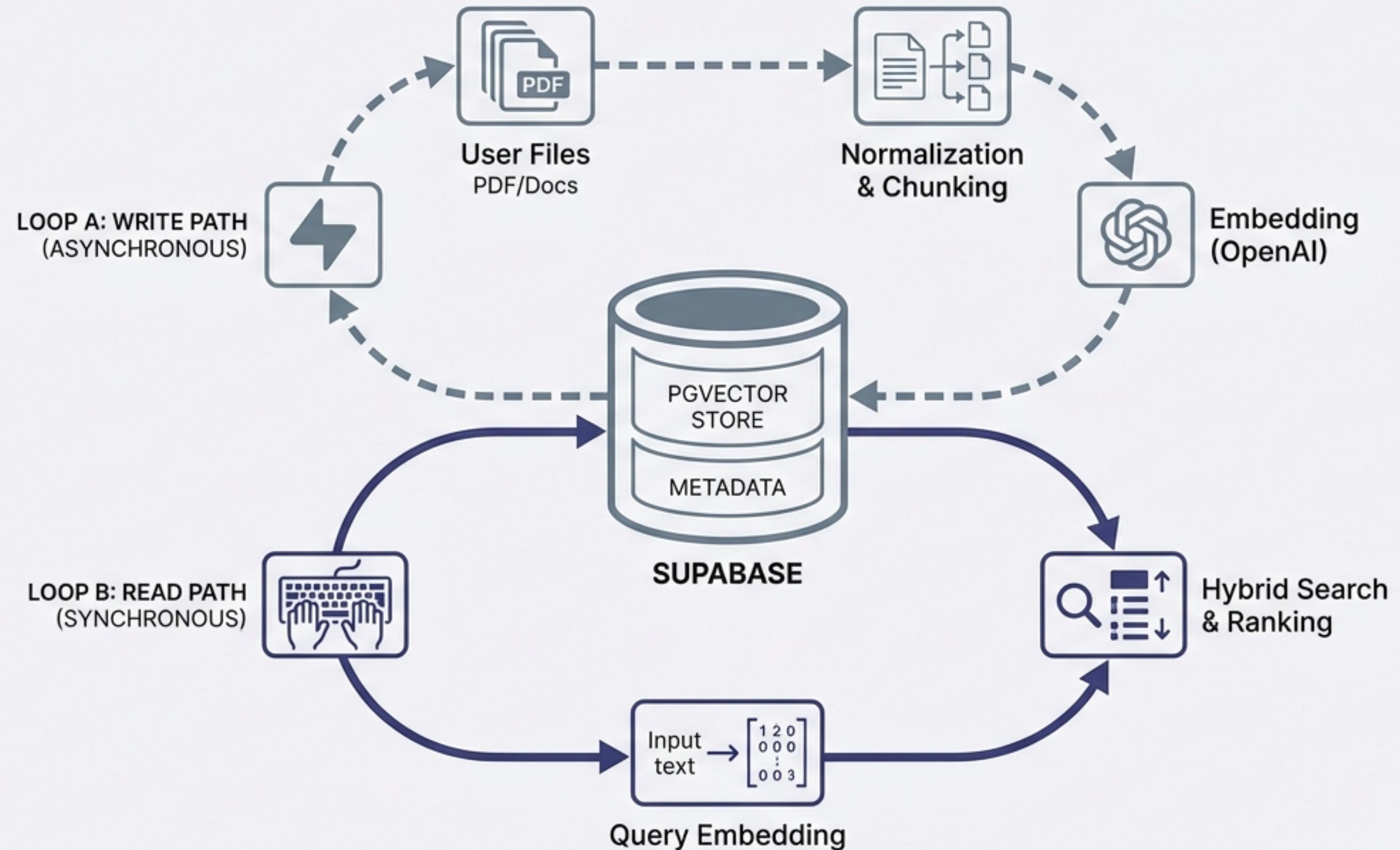
Engineering context-aware suggestions via Retrieval Augmented Generation.



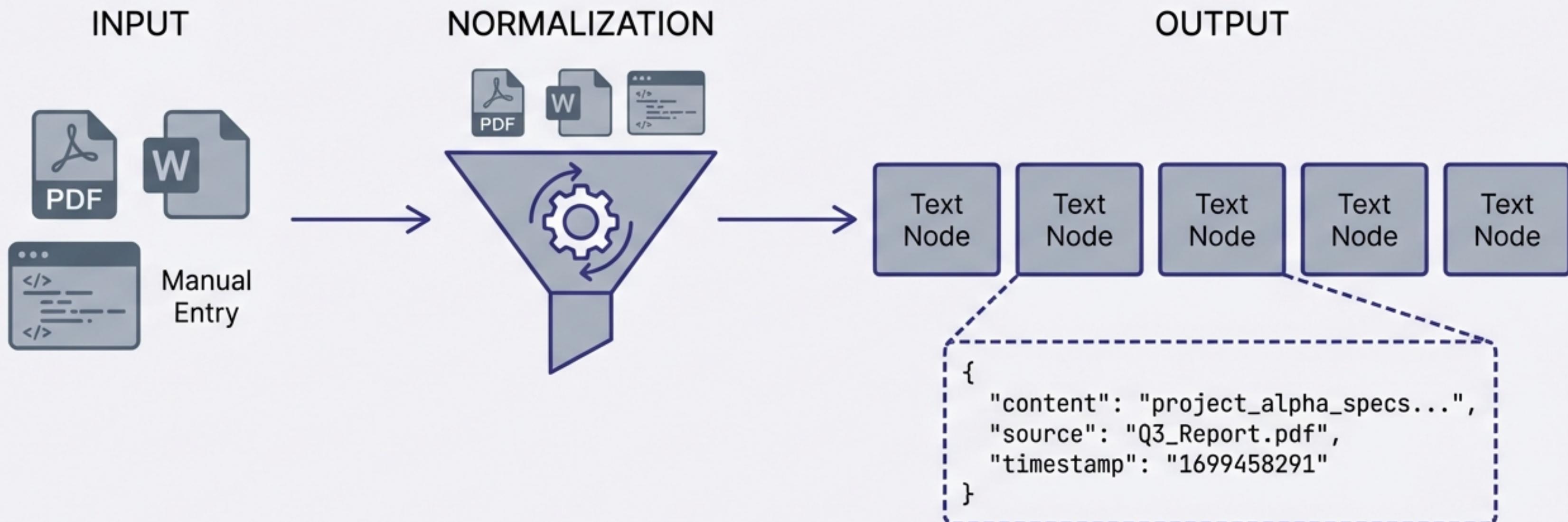
# The Dual-Pipeline Architecture

The system creates a "Data Journey" separated into two distinct timelines.

The Asynchronous Training Pipeline builds the knowledge base, while the Synchronous Suggestion Pipeline queries it in real-time.



# Ingestion: Transforming Files into Raw Data

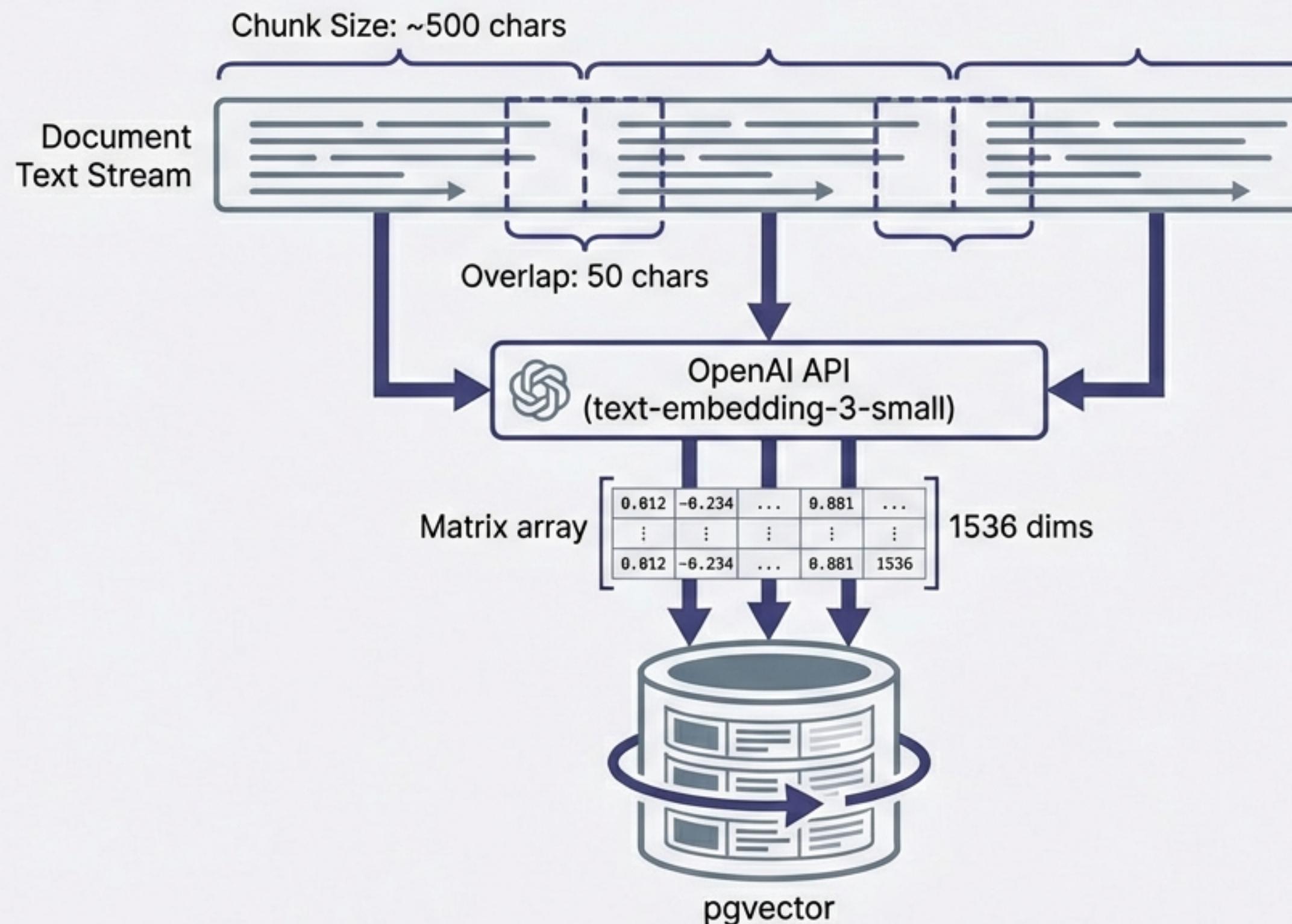


- **Input Agnostic:** Handles structured code and unstructured prose.

- **Sanitization:** Strips binary assets and non-text formatting.

- **Lineage Preserved:** Metadata tags allow for citation.

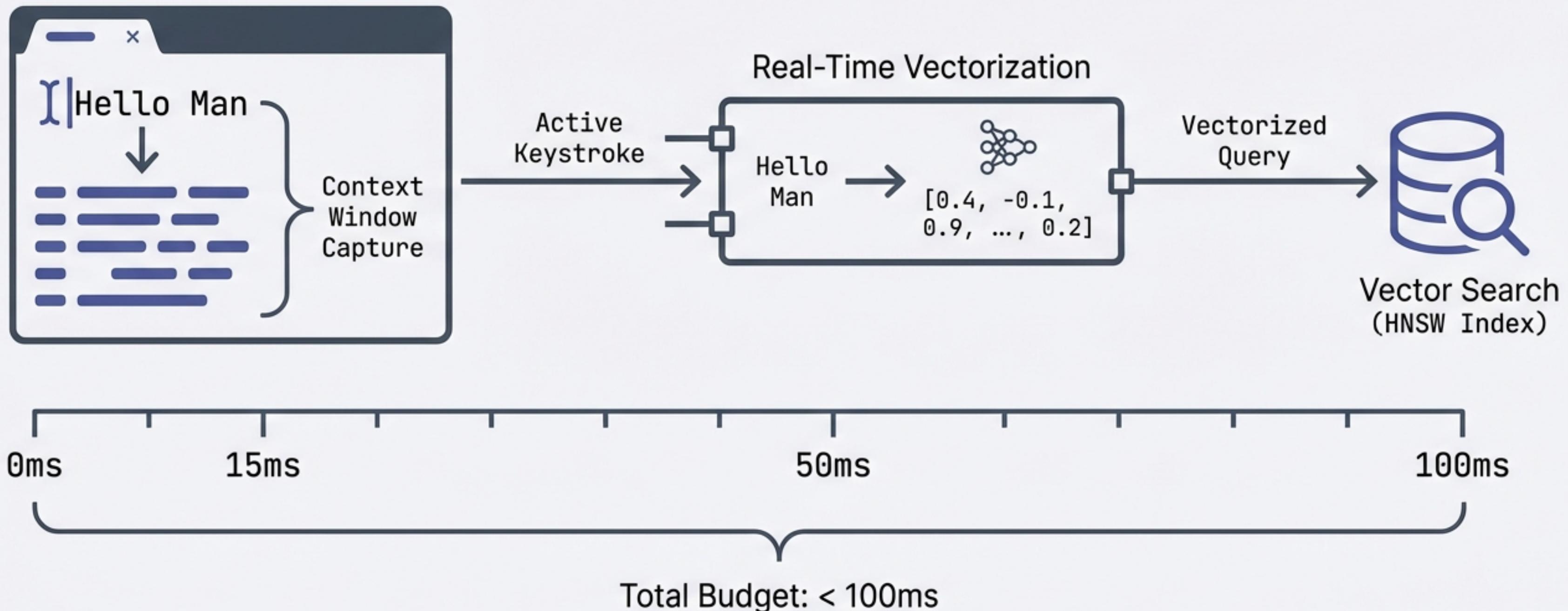
# Vectorization: Creating the Semantic Index



## TECHNICAL SPECS

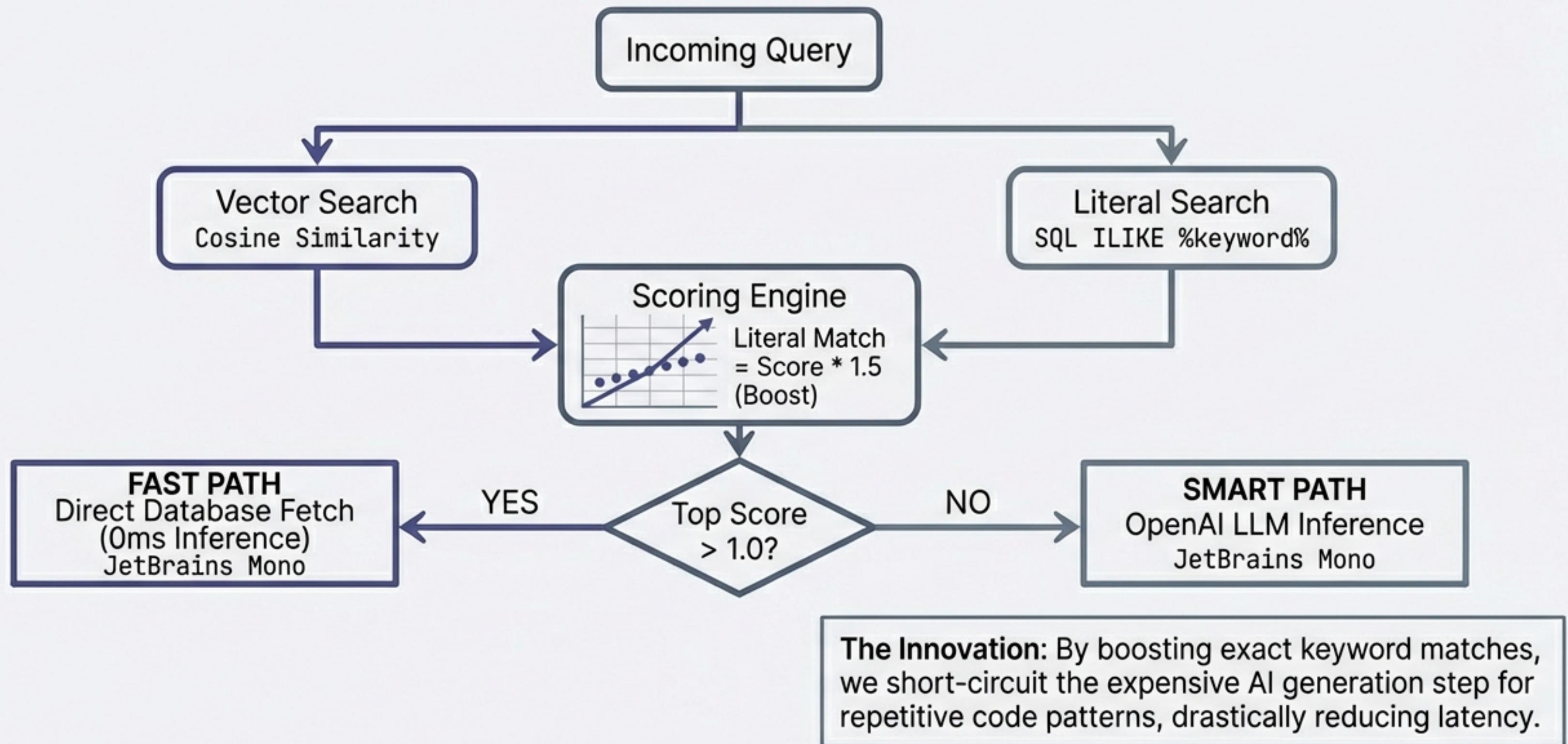
Parameter	Value
Model	text-embedding-3-small
Dimensions	1536
Chunk Size	500 tokens
Storage	pgvector (HNSW Index)

# RAG Usage: The Query Lifecycle

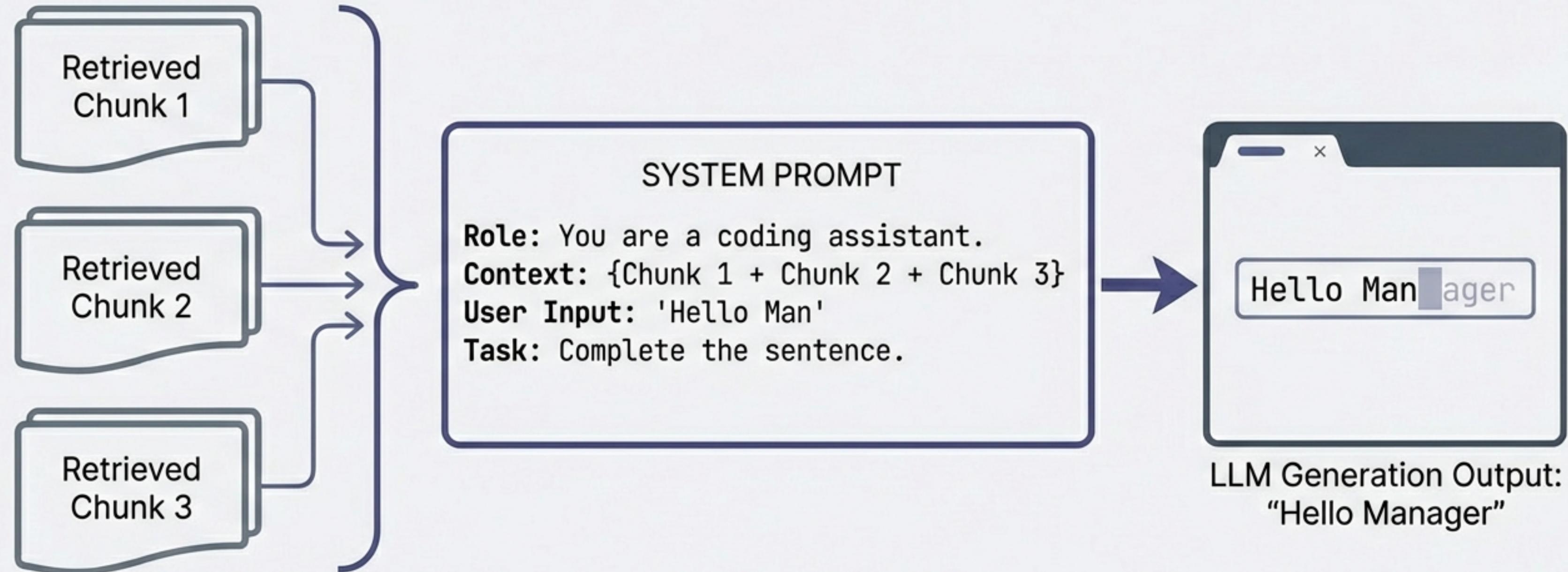


Unlike the training pipeline, the suggestion pipeline is on the critical path. We capture the user's cursor position and preceding lines to form a query, embedding it instantly to match the vector space of our index.

# Hybrid Search Strategy: Precision vs. Speed



# Generation: Delivering the Suggestion



The Smart Path constructs a contextual prompt only when necessary, ensuring the LLM has specific, retrieved knowledge to generate accurate completions.