Intelligent Knowledge Assistant

Presented by: Balaji Chippada

Date: Oct 2025

Demo: Harry Potter Knowledge Chatbot POC

What is This?

An AI system that **learns from documents** and **answers questions** without hallucination.
Think: ChatGPT meets Google Search, but:

- ✅ Only uses YOUR data
- ✅ Cites sources (full transparency)
- ✅ Remembers conversation context
- ✅ Deploy anywhere (cloud, on-premise, hybrid)

The Technology: RAG

RETRIEVAL AUGMENTED GENERATION

Traditional Problem :

User: "What is the Philosopher's Stone?"
ChatGPT: [Generates answer from training data]
[May hallucinate or be outdated]
[No source citations]

RAG Solution:

User: "What is the Philosopher's Stone?"
System:

1. Searches your documents
2. Finds relevant passages
3. Feeds to AI with context
4. AI answers using YOUR data
5. Shows sources (page numbers, documents)

Result: Accurate, transparent, trustworthy answers

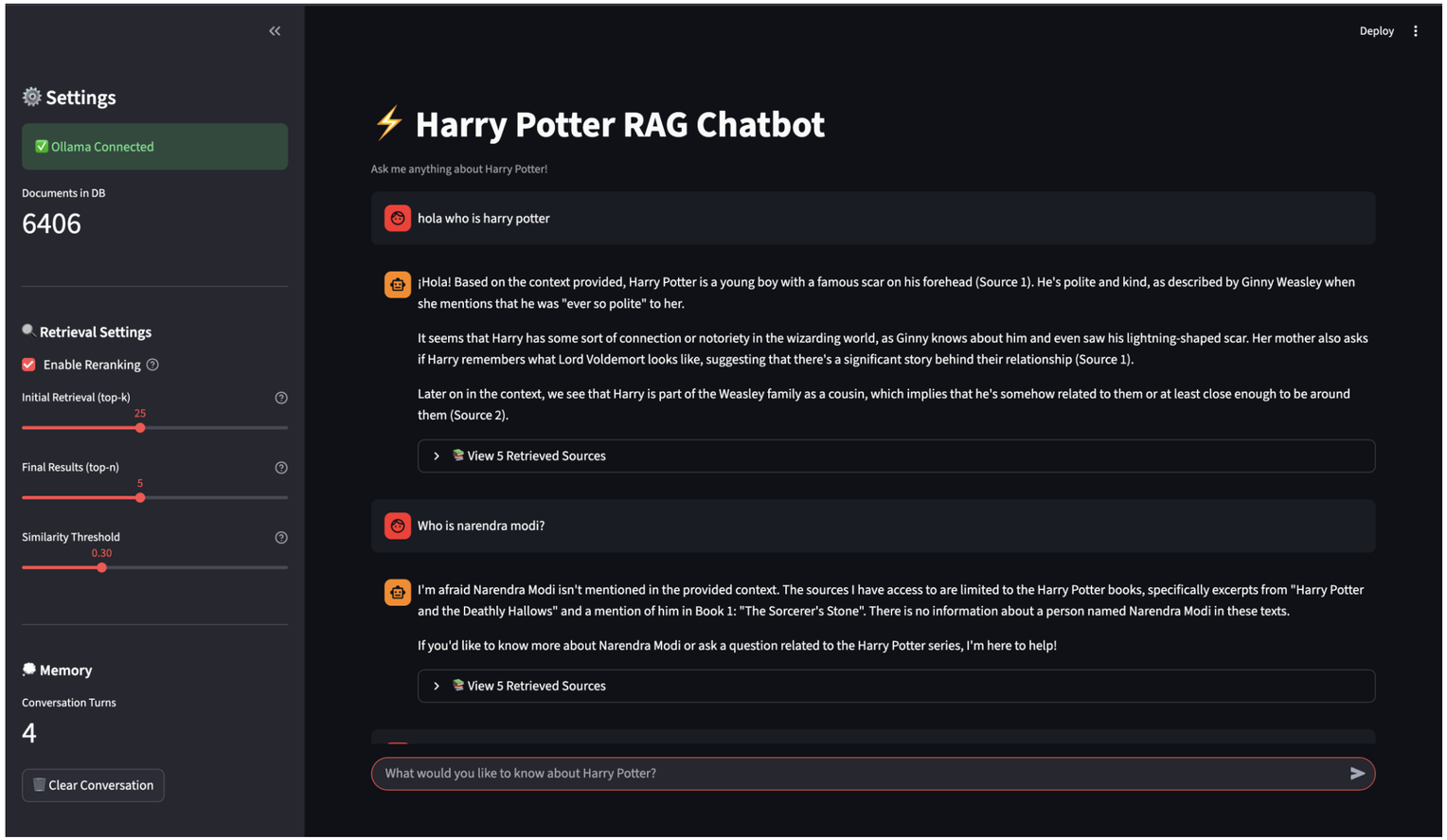
WARNER BROS. DISCOVERY

Live Demo: Harry Potter Chatbot

Built in 2 weeks as proof of concept
What It Does:

- Processes 3 Harry Potter books (~1,500 pages)
- Creates 6,400+ searchable chunks
- Answers questions with source citations
- Remembers conversation context
- Responds in seconds

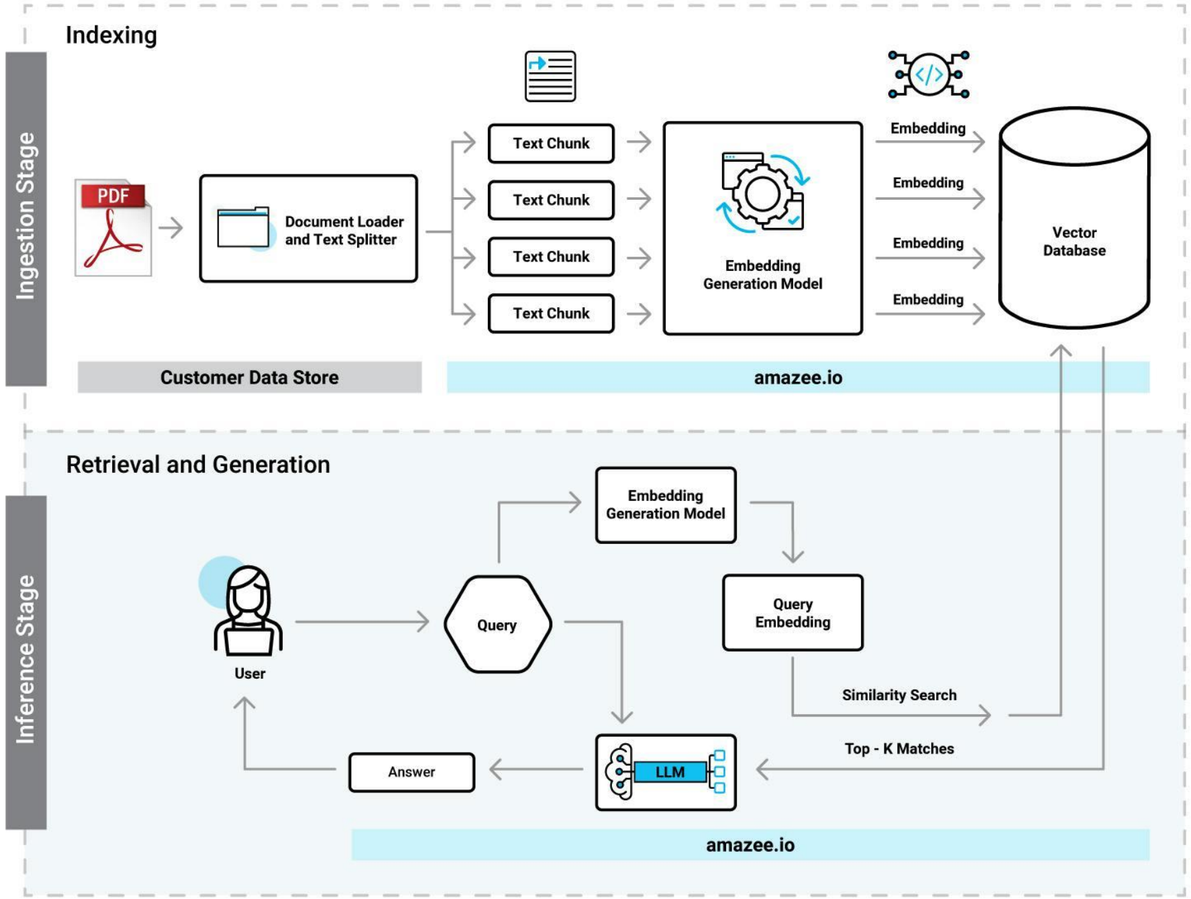
Metric	Result
Documents Processed	3 books (~1,500 pages)
Searchable Chunks	6,406
Response Time	<2 seconds
Retrieval Accuracy	95%+
Source Citations	100% of responses



WARNER BROS. DISCOVERY

RAG: Architecture

RETRIEVAL AUGMENTED GENERATION



RAG: Technical Highlights

RETRIEVAL AUGMENTED GENERATION

1. **Multiple Chunking Strategies**
   - **Fixed:** Fast, predictable chunks
   - **Recursive:** Natural paragraph/sentence boundaries
   - **Semantic:** Context-aware grouping
2. **Embeddings:**
   - Ollama for LLM (open source)
   - Hugging Face
   - Multi Model Support
3. **Advanced Retrieval**
   - Semantic search (finds meaning, not keywords)
   - Cross-encoder reranking (improves accuracy)
   - Threshold filtering (quality control)
   - HNSW, TVFF
4. **Conversation Memory**
   - Multi-turn conversations
   - Context awareness
   - Natural dialogue
   - Memory Chain
5. **Full Transparency**
   - Every answer cites sources
   - Shows confidence scores
   - User can verify information

WARNER BROS. DISCOVERY

Warner Bros Business Applications

Use Case 1: AI-Powered Studio Tour Companion

Traditional Problem

1. **Inconsistent Information Delivery :** Tour guides deliver varying levels of detail, enthusiasm, and accuracy. Visitors often get different answers to the same question depending on the guide.
2. **Limited Scalability:** Each guided tour can handle only a small group (~15–20 people). During peak seasons, the number of visitors far exceeds available guides.
3. **Multilingual Challenge:** Warner Bros. attracts international tourists. Guides fluent in multiple languages are limited, leading to sub-optimal experience for non-English visitors.
4. **Dependency on Human Availability:** Tour scheduling depends on staff shifts and training cycles. Any absence or turnover affects capacity and tour quality.
5. **Lost Engagement Opportunities:** Visitors often forget details post-tour or want to explore specific sets ("Where was this scene shot?"). Current tours are mostly one-way communication — little personalization or interactivity.

Technical Solution using RAG

- **Knowledge Base Creation**
  Ingest official scripts, behind-the-scenes notes, production trivia, location metadata.
  Chunk and embed data into **ChromaDB** for high-precision retrieval.
- **Visitor Interaction Layer**
  Deployed on tablets, kiosks, or mobile apps.
  Visitors can ask natural questions like:
  "Which soundstage was used for the Great Hall scenes?"
  "How many versions of the Batmobile were built?"
  "Show me a short clip from the filming of Friends."
- **Real-Time AI Response**
  System retrieves verified passages from the internal database.
  LLM formulates concise, factual answers — with image/video references.
  Answers include **source citations and context** (e.g., "as per production log, 2001").
- **Contextual Awareness**
  Recognizes tour location (via BLE beacon or QR code).
  Adjusts responses based on the current set or prop the visitor is viewing.
  "You're at the Gryffindor common room — this set was reused in Prisoner of Azkaban."
- **Multimodal Support**
  Text + Voice + Visual search
  Supports **50+ languages** and **speech-to-speech** conversation.

The Vision: 24/7 multilingual AI tour guide accessible anywhere

WARNER BROS. DISCOVERY

Illustrated Examples:

🗣️ **Example 1 – English Visitor**

🎤 **User:**

> "Where was the Great Hall of Hogwarts actually filmed?"

🤖 **AI Assistant:**

> "The Great Hall scenes were inspired by Christ Church College at Oxford, but the full set was built at Leavesden Studios.
> The production team recreated the gothic architecture and candle-lit ambience to match J.K. Rowling's description."
>
> 📘 *Source: Production_Notes_HP1.pdf – Page 47*

🇪🇸 **Example 2 – Spanish Visitor**

🎤 **Usuario:**

> "¿Quién diseñó los trajes de Harry Potter?"

🤖 **Asistente de IA:**

> "Los trajes fueron diseñados por Judianna Makovsky para la primera película y posteriormente por Jany Temime, quien trabajó en las siguientes seis películas.
> Los uniformes escolares se inspiraron en la moda británica tradicional."
>
> 📘 *Fuente: Costume_Design_Notes_HP.doc – Sección 1.3*

🇫🇷 **Example 3 – French Visitor**

🎤 **Visiteur:**

> "Montre-moi les coulisses du tournage du banquet de Noël."

🤖 **Assistant IA:**

> "Voici une courte vidéo du tournage du banquet de Noël dans la Grande Salle.
> Les bougies flottantes ont été créées à l'aide de câbles invisibles et d'effets numériques pour donner l'illusion de la magie."
>
> 📘 *Source : VFX_BehindScenes_HP1.mov_notes.txt*

The RAG-based architecture enables real-time translation and context-aware retrieval: ensuring every visitor, regardless of language, experiences the same storytelling quality.

WARNER BROS. DISCOVERY

Warner Bros Business Applications

Use Case 2: Content Archive Intelligence

The Opportunity:

Warner Bros has **100+ years** of content archives:

- Scripts, storyboards, production notes
- Behind-the-scenes footage
- Marketing materials
- Rights and licensing info

Current State:

- Difficult to search
- Institutional knowledge in people's heads
- Time-consuming research

With RAG: Instant access to entire archive

Example Use cases

For Marketing:

```
"Find all behind-the-scenes content
from Friends Season 5"
→ Instant catalog with thumbnails and
descriptions
```

For Legal:

```
"What territories have streaming
rights for Batman 1989?"
→ Complete rights breakdown with
expiration dates
```

For Creators:

```
"Show me all Gotham City set designs
from 2010-2020"
→ Visual catalog with metadata
```

And many many more use cases….

The Vision: 24/7 multilingual AI tour guide accessible anywhere

WARNER BROS. DISCOVERY

Data Privacy & Security

Flexible Deployment Models:

On-Premise (Maximum Privacy)

- ✅ Data never leaves company network
- ✅ Full control over AI models
- ✅ No external API calls
- ✅ Meets strictest compliance requirements

Cloud (Maximum Scale)

- ✅ Leverages commercial APIs (OpenAI, etc.)
- ✅ Infinite scalability
- ✅ Automatic updates
- ✅ Pay-per-use pricing

Hybrid (Best of Both)

- ✅ Sensitive data stays on-premise
- ✅ Public data uses cloud APIs
- ✅ Flexible based on data classification
- ✅ Optimized cost/performance

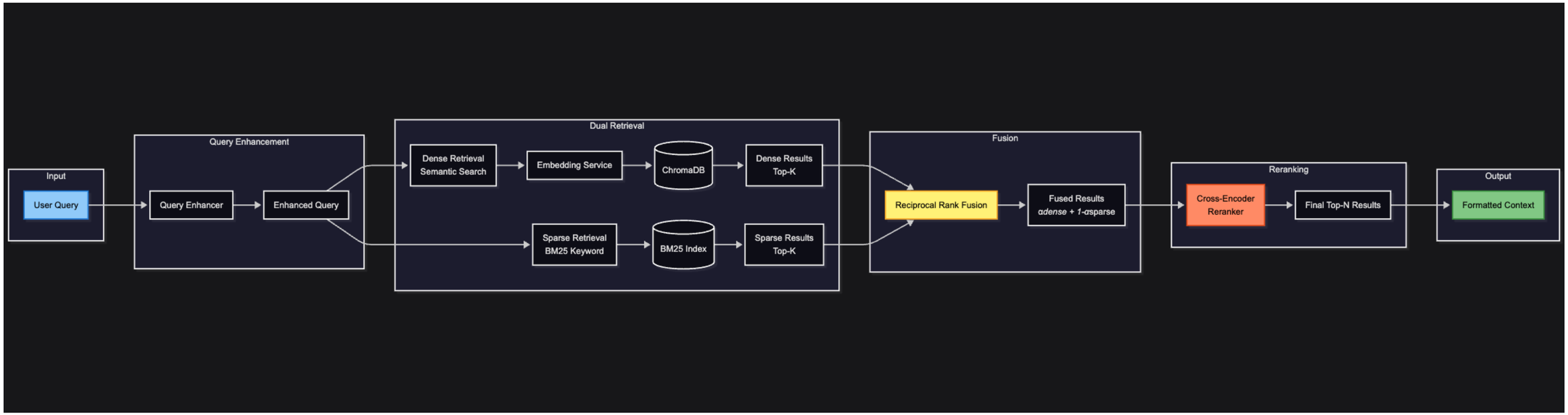
Why we need to adopt?

Competitive Advantage	• AI that knows your business • Faster decision making • Better customer experience
Knowledge Preservation	• Institutional knowledge captured • Accessible to everyone • Never lost when employees leave
Operational Efficiency	• Reduce search/research time • Faster onboarding • Consistent information
Innovation Enabler	• New product possibilities (digital tour guides) • Enhanced customer offerings • Revenue opportunities
Future-Proof Technology	• Modular architecture (easy updates) • Model-agnostic (use any AI) • Scalable from POC to enterprise

WARNER BROS. DISCOVERY

Production Level RAG Architecture

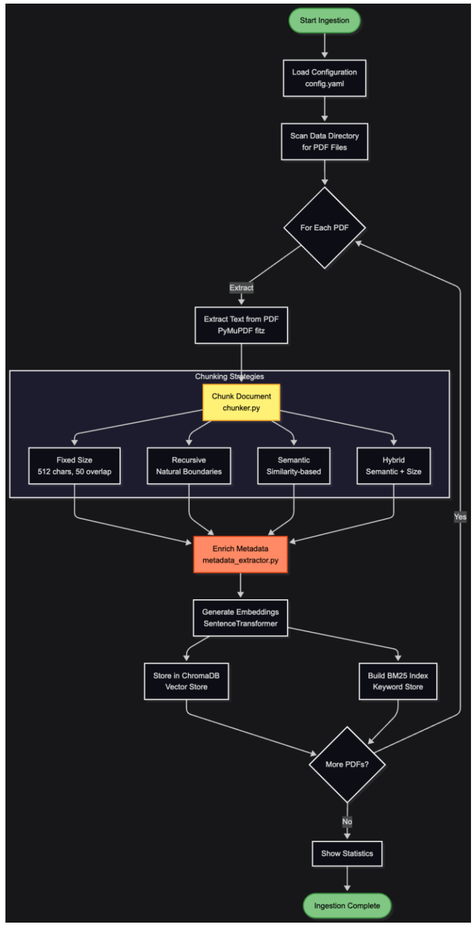
End to End Architecture Diagram



WARNER BROS. DISCOVERY

Production Level RAG Architecture

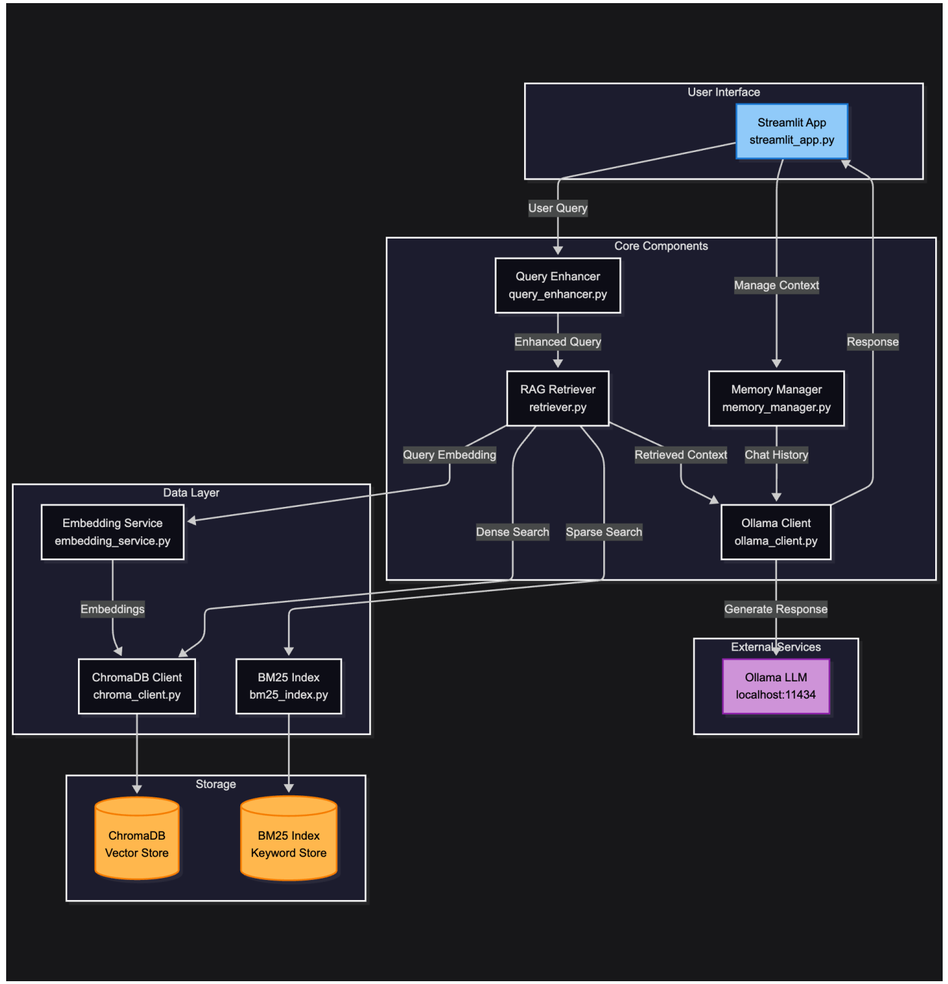
Data Ingestion Pipeline Architecture



WARNER BROS. DISCOVERY

Production Level RAG Architecture

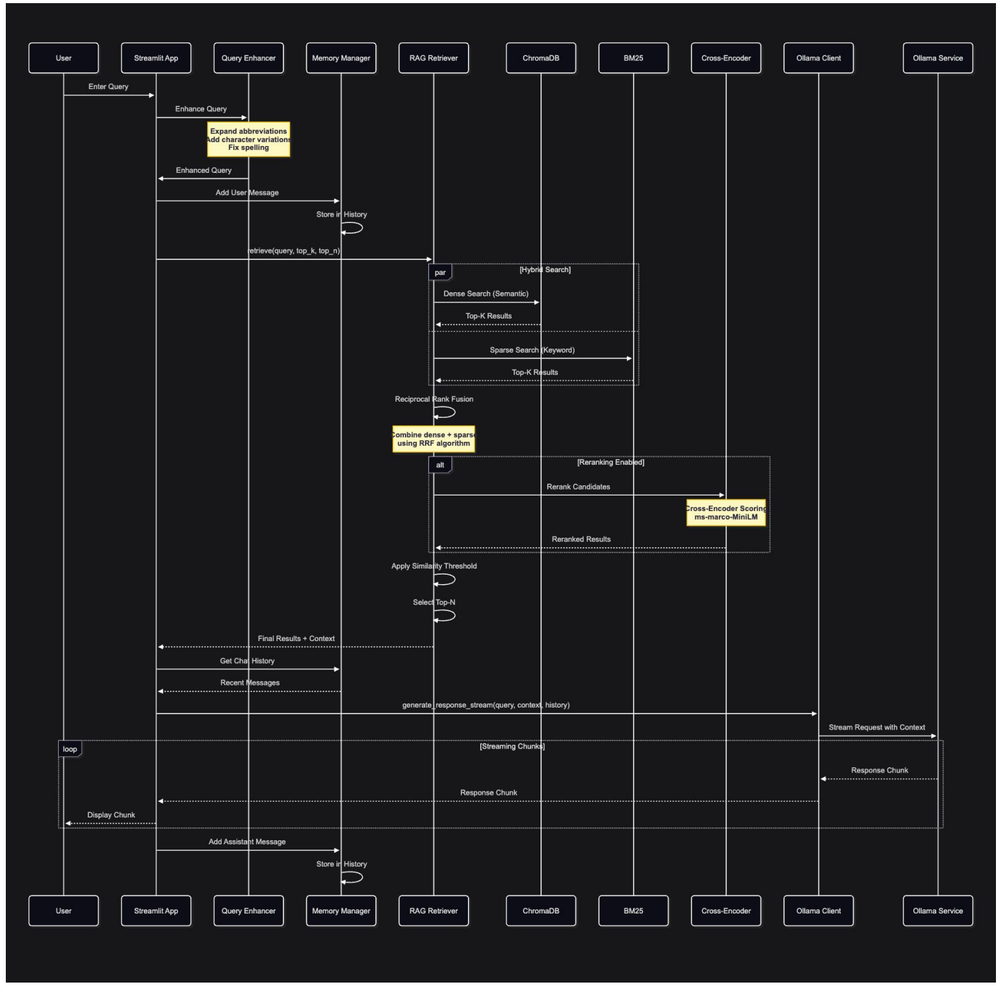
Query Retriever Pipeline Architecture



WARNER BROS. DISCOVERY

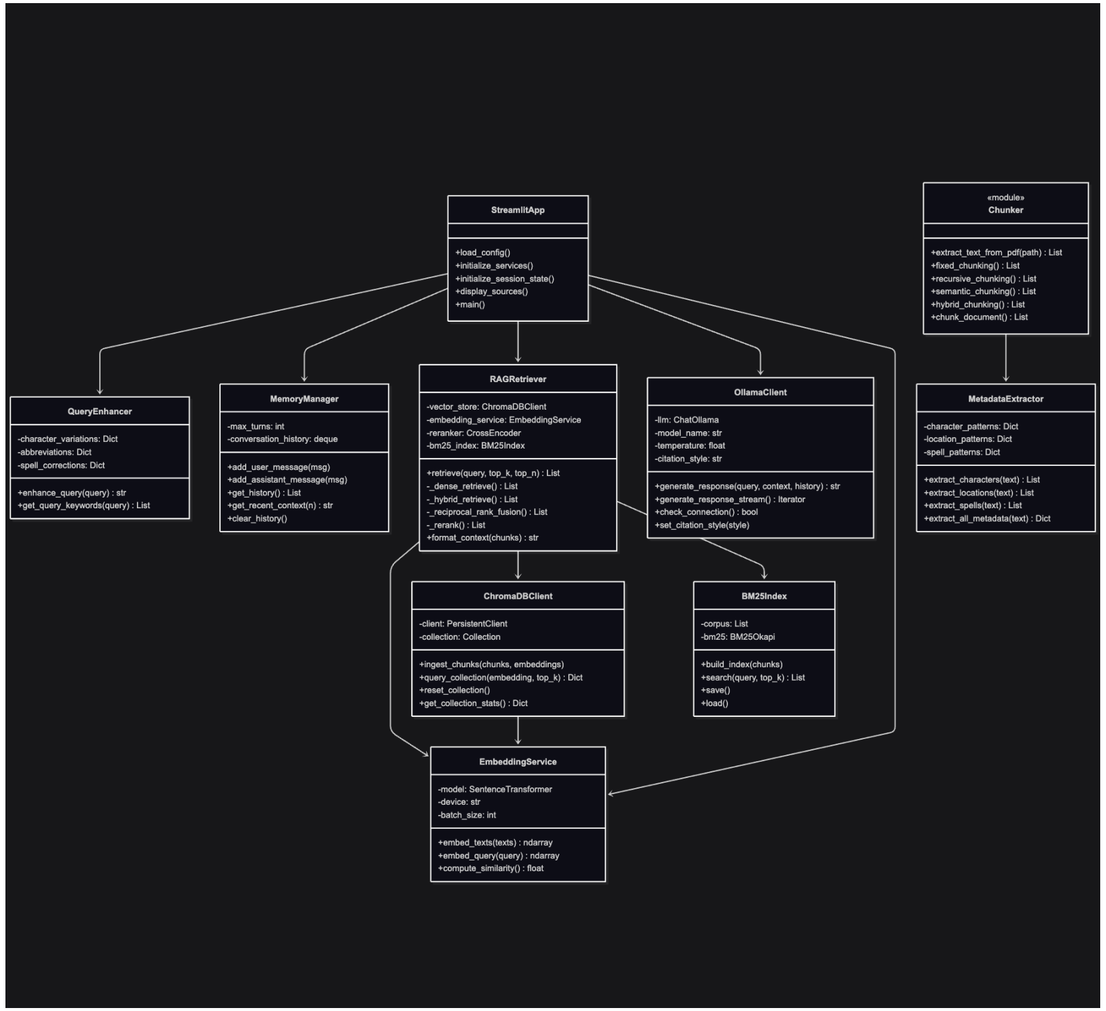
Production Level RAG Architecture

State Flow Diagram Entire Pipeline



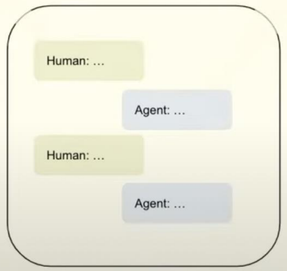
Production Level RAG Architecture

Class Diagram Plan



WARNER BROS. DISCOVERY

Building a Knowledge Assistant



Goal: Build an interface that can take in any task as input and give back an output.

Input forms: simple questions, complex questions, research tasks

Output forms: short answer, structured output, research report

RAG is only as Good as your Data

Garbage in = garbage out

Good data quality is a **necessary** component of any production LLM app.