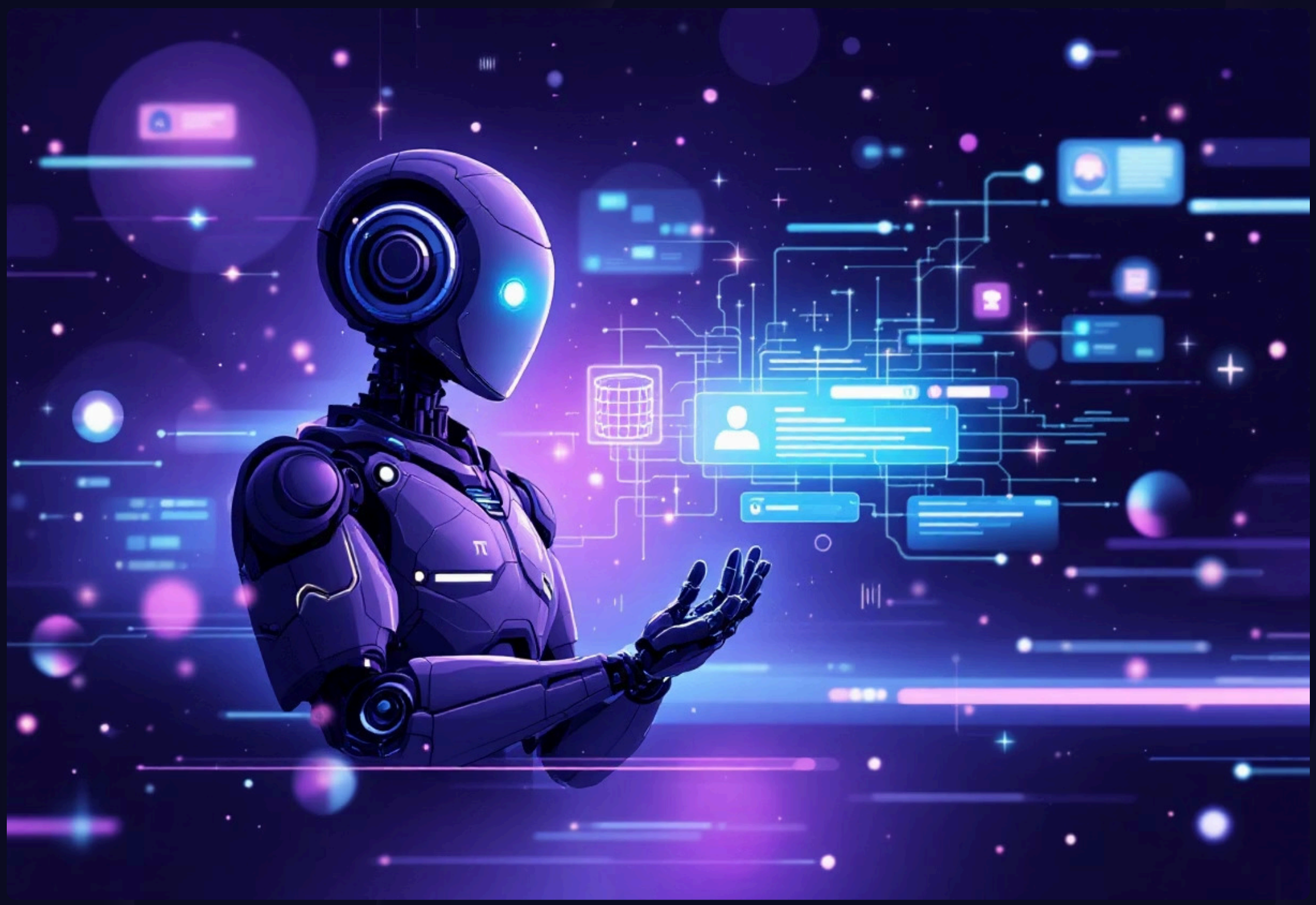
# Aura Guide: An AI Aura-Soma Chatbot for Velana.net

An innovative AI chatbot powered by Retrieval-Augmented Generation, specifically designed to offer intelligent and accurate guidance on

Aura-Soma for Velana.net

# Project Overview & Hypothesis



This project aimed to develop an AI Aura-Soma  Chatbot for Velana.net, focusing on high-fidelity user assistance.

## My Hypothesis

By creating a specialized knowledge base from structured data and enriching it with unstructured sources (like video transcripts), we could build a RAG-based Chatbot capable of answering user questions with **over 90% semantic accuracy**.

### Result: Hypothesis Strongly Supported

The RAG-based Chatbot successfully demonstrated high semantic accuracy, validating our approach.

# Interactive Project Architecture

## 1. RAG & LLM Architecture

This diagram illustrates the main information flow. The bot ingests data, stores it in a searchable database, retrieves relevant information based on a user's query, and uses an LLM to generate a final, context-aware response.

### Data Sources

JSON files and YouTube video transcripts form the knowledge base.

### Pinecone DB

A vector database where text is stored as numerical embeddings for fast similarity search.

### Retrieval

The user's query is used to find the most relevant text chunks from Pinecone.

### LLM + Prompt

LLM: gpt-3.5-turbo-0125. Embedding: text-embedding-3-large. The retrieved context and query are sent with a system prompt for generation.

## 2. Agent & Tool Logic

To improve accuracy, a routing system classifies each query and selects a specialized tool. Each tool uses specific metadata filters to narrow down the search, ensuring the LLM receives highly relevant context.

### User Query

The initial question from the user.

### Router

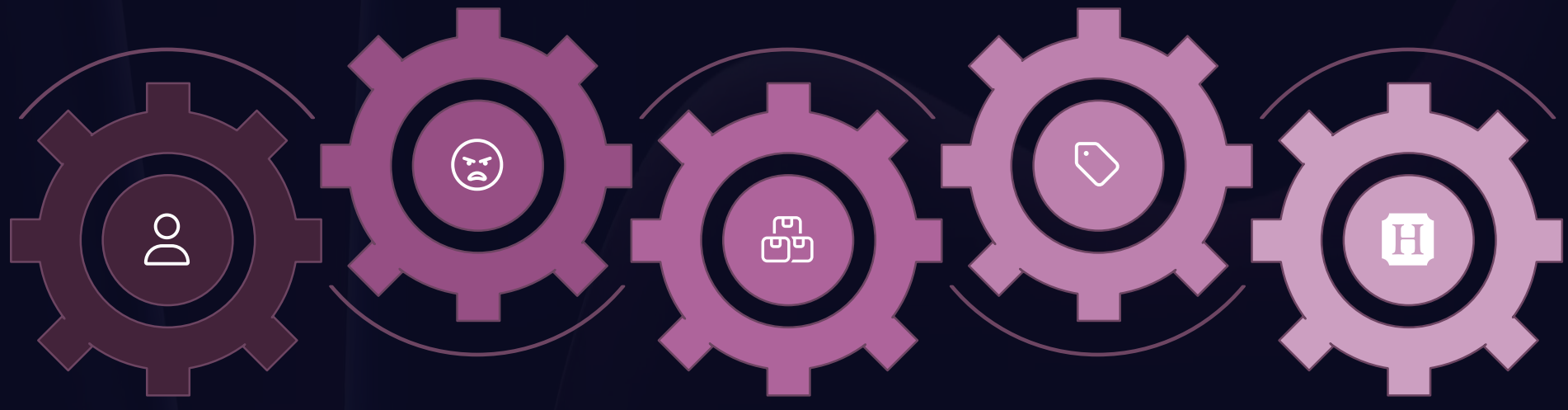
An LLM call that classifies the query's intent (e.g., pricing, product).

### Specialized Tools

- ProductTool - Filters for `topic: "products"`
- PricingTool - Filters for `topic: "pricing"`
- FAQTool - Filters for general topics
- MeetAuraSomaTool - Filters for `doc: "transcript"`
- BookingRedirectTool - Appends booking link

**AI Router**
Classifies intent and routes

**PricingTool**
Provides pricing details



**User Query**
Incoming question from user

**ProductTool**
Handles product information

**FAQ & Redirect**
FAQ, bookings, or external redirects

# 3. Evaluation & Deployment

A robust evaluation loop ensures quality control by automatically comparing bot responses to a 'golden dataset'. The system is designed for deployment as a standalone Streamlit application.

## Evaluation Loop

### Bot Response

The AI generates a response to the user query.

### Compare to Golden Dataset

The bot's response is automatically compared against a set of predefined ideal responses.

### Evaluator LLM (GPT-4o)

A powerful LLM acts as an impartial judge to score the accuracy and relevance of the bot's output.

### Log Score to LangSmith

Evaluation scores and metrics are logged for continuous monitoring and improvement.

## App Structure

### User Interface

The UI is embedded on the live Velana website, which is hosted on the Tilda platform.

### Streamlit App

This component handles the user interface, manages application state, and makes calls to the RAG logic.

### RAG Bot Logic

The core Retrieval-Augmented Generation system processes queries and synthesizes responses.

# Process & Challenges

### Data Enrichment

To add depth, I transcribed YouTube videos using OpenAI's Whisper API. A unique cleaning technique grouped text by speaker, ensuring contextual coherence for the RAG system.

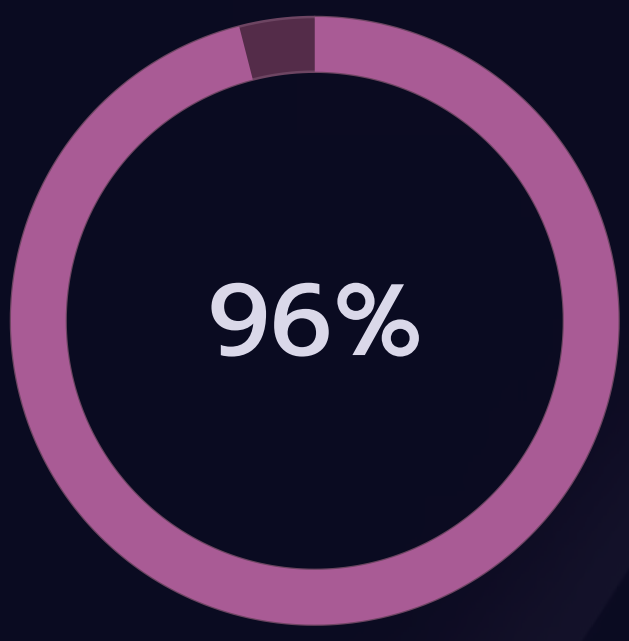
### Clearing Untracked File Cache

A major challenge arose when untracked files in my local Git cache caused sync conflicts with GitHub. Resolving this required clearing the Git cache to remove these ghost files, which was a critical step to ensure a clean and successful deployment.

### Initial Evaluator Strictness

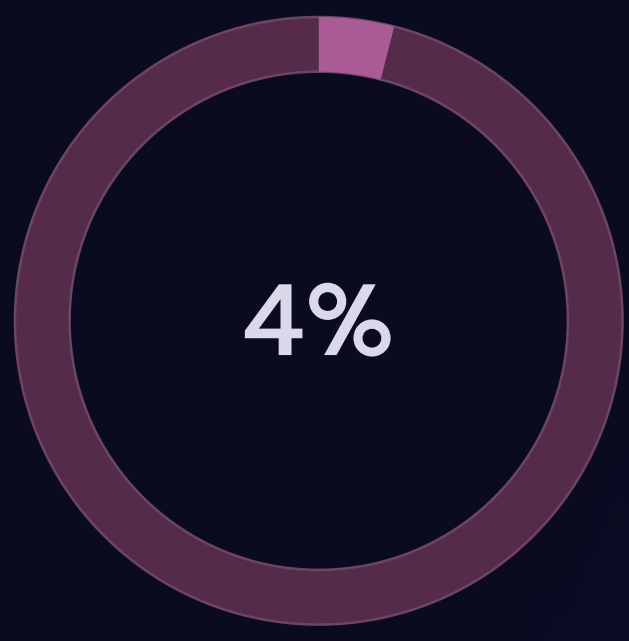
Initially, the `gpt-4o` evaluator did not score for semantic accuracy as expected, often giving low scores to answers that were correct. The issue was its literal interpretation. I engineered a more robust, hybrid system with a detailed rubric and a keyword-matching fallback to align the AI's judgment more closely with human assessment.

# Results & Key Insights



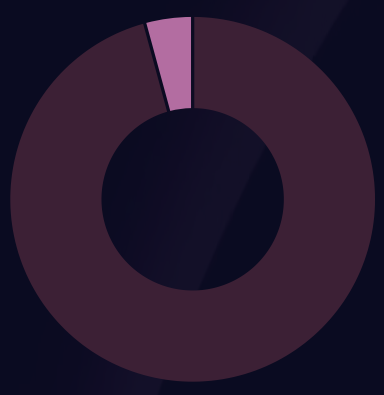

Semantic Accuracy

The bot achieved a high score, confirming our hypothesis.




Margin for Improvement

Identified areas for further refinement in the evaluation process.




Final score based on a hybrid of LLM evaluation and keyword matching.

### Hypothesis Confirmed

The bot achieved a **96% semantic accuracy score**, strongly supporting my initial hypothesis.

### Surprising Insight

Unstructured data from YouTube transcripts was incredibly effective, providing the bot with a more authentic, philosophical voice that structured data lacked.

### Main Learning: Evaluation is Key

A robust, automated evaluation pipeline is the most critical component for building a reliable and trustworthy AI system.

### Future Work

The next step is a beta launch to test the bot against unpredictable, real-world user queries and use that data to further improve its knowledge and accuracy.

---

## Aura Guide Bot Project

By Lovely Ibañez

# Thank You