**Vedaz LLM-Based Smart Astrologer Recommendation Engine**

Prepared by: Pratik Guha Roy  
LinkedIn: [linkedin.com/in/pratikfromindia](https://www.linkedin.com/in/pratikfromindia)  
Date: 23/07/2025

1. What We Built

A smart recommendation engine that uses modern LLM techniques to match users with astrologers based on natural language input. Users describe their problem or area of concern in plain English (or Hindi), and the system recommends the most suitable astrologers based on semantic similarity, profile tags, and LLM summarization.

2. Tech Stack Used

Embedding & Similarity Search

* Model: all-mpnet-base-v2 from sentence-transformers for user query and astrologer profile embeddings
* Why? It offers high-quality semantic representation and performs better than MiniLM for nuanced similarity tasks.

LLM Summarization / Explanation

* Provider: Together.ai
* Model: mistralai/Mixtral-8x7B-Instruct-v0.1
* Use: To explain why a recommendation was made in human-readable language
* Access: via Together AI API using Python client (together SDK with .env key handling)

Vector Store

* Used FAISS (in-memory, open-source) for fast nearest neighbor search on embedded vectors.

Frontend (Optional)

* While this version uses CLI / Jupyter interface, it is ready to be extended to a React or Streamlit frontend with FastAPI as backend.

3. Sample Flow

1. Embed user query

query = "I want to talk to someone about career, spirituality and guidance"

query\_embedding = model.encode(query)

2. Compute cosine similarity with precomputed astrologer profile embeddings

similarities = cosine\_similarity([query\_embedding], all\_astrologer\_embeddings)

3. Select top 3 matches and prepare summary using Together AI

summary\_prompt = f"User asked: {query}. Recommend an astrologer among {top\_3\_names} and explain why."

response = together.chat.completions.create(

model="mistralai/Mixtral-8x7B-Instruct-v0.1",

messages=[{"role": "user", "content": summary\_prompt}]

)

4. Hosting & Cost Plan

| Component | Platform | Monthly Cost Estimate |
| --- | --- | --- |
| LLM API | Together AI | Rs. 0–1000 (free tier + overage) |
| Embedding /  FAISS Index | Local /  lightweight VM | Free / minimal |
| Backend API (Python) | Render /  Railway | Rs. 300-600 |
| Optional Frontend | React /  Streamlit | Free (static) |

Total Estimated Cost: Under Rs. 1500/month for up to 50,000 queries with caching and smart LLM usage.

5. Privacy & Ethical Guardrails

* No user chats stored permanently unless explicit consent is given.
* Sensitive topics like mental health or legal issues are filtered out.
* Bias mitigation: Recommendations avoid prioritizing astrologers based on gender, region, or popularity.
* LLM guardrails can be added using prompt shaping or moderation APIs.

Conclusion

This system offers explainable AI-based recommendations for Vedaz that combine classical NLP (embeddings + similarity search) with modern generative AI (Mixtral via Together AI). It balances accuracy, cost-efficiency, and user trust, and is modular enough to be scaled further with fine-tuning, multilingual support, or richer analytics.