LLM-Powered Astrologer Recommendation Engine for Vedaz: Key Considerations

I. LLM Stack Recommendation and Justification

For Vedaz's astrologer recommendation engine, a **fine-tuned open-source LLM (Mistral 7B or LLaMA 3 8B) deployed via Hugging Face Inference Endpoints** is recommended. This choice prioritizes data privacy, deep customization, and cost-efficiency.

- **Data Privacy & Control:** Sensitive user data (chat history, profile) requires stringent control. Open-source models offer complete data sovereignty, while Hugging Face Inference Endpoints provide SOC2 Type 2 certification and GDPR compliance, with explicit 30-day log retention policies, superior to proprietary APIs' default data retention.
- **Deep Customization:** Fine-tuning on Vedaz's proprietary data is crucial for domain-specific understanding and accurate astrologer recommendations, creating a unique AI asset. Mistral 7B and LLaMA 3 8B are efficient and suitable for fine-tuning using PEFT methods like LoRA/QLoRA, which reduce memory needs.
- **Scalable Cost-Efficiency:** For 50,000 monthly active users, managed open-source deployments are more cost-effective long-term than token-based proprietary APIs. Hugging Face Inference Endpoints offer autoscaling and "scale-to-zero" features, optimizing costs during fluctuating traffic.

Table 1: LLM Stack Comparison

Feature	OpenAI (Proprietary)	Hugging Face (Managed Open-Source)
Data Privacy	Default 30-day retention/ZDR option	30-day log retention/private endpoints
Customization	Limited API Fine-tuning	LoRA/QLoRA
Cost Model	Pay-per-token	Hourly/Pay-as-you-go
Setup Time	Minutes	Hours

II. Hosting and Scaling Strategy

Effective hosting and scaling are vital for 50,000 MAU.

- **Cloud Provider:** AWS, Azure, and GCP offer robust ML services and GPU instances (NVIDIA A10G, L4) suitable for LLM inference. Cost optimization involves selecting efficient GPUs and leveraging multi-model endpoints to share resources.
- **Deployment:** Managed Inference Endpoints (Hugging Face, SageMaker, Vertex AI) offer simplicity and integrated autoscaling. Self-hosting on Kubernetes with optimized frameworks

like Ray Serve or Triton Inference Server (using vLLM for "continuous batching") provides more control and throughput.

- **Scaling:** Horizontal Pod Autoscalers (HPAs) manage unpredictable loads, while continuous batching maximizes GPU utilization by dynamically processing requests. Multi-model endpoints further optimize resource use.

III. Monthly Cost Estimation for 50,000 Monthly Active User

Assumptions: 50,000 MAU, 1 interaction/user/day (1.5M interactions/month), 500 input tokens + 100 output tokens = 600 total tokens/interaction. Total: 900 Million tokens/month.

Table 2: Estimated Monthly LLM Inference Costs

LLM Stack/Hosting Strategy	Core Cost Driver	Estimated Monthly Cost
OpenAI GPT-4o Mini (API)	Token-based pricing	~$2,700.00
Hugging Face Inference Endpoints (Mistral 7B)	GPU-hours (managed service)	~$8,760.00 (before scale-to-zero)
Self-Hosted AWS EC2 (LLaMA 3 8B)	GPU-hours + Operational Overhead	~$7,300 - $8,000+

Note: Costs are highly sensitive to usage; token optimisation (concise prompts, context trimming) is crucial.

IV. Privacy and Safety Concerns

Addressing privacy and safety is fundamental for user trust and ethical AI.

- **Data Privacy & PII Handling:** Sensitive user data requires robust PII masking and anonymization (detection, masking, optional restoration) using tools like regex and Named Entity Recognition (NER). Compliance with India's Digital Personal Data Protection Act, 2023 (DPDP Act) is paramount, requiring explicit consent and lawful purpose for data processing.
- **Ethical AI Considerations:**
  - **Bias & Fairness:** LLMs can propagate societal biases. Mitigation involves unbiased prompting, representative training data, and continuous monitoring.
  - **Hallucinations:** LLMs can generate incorrect content. Retrieval Augmented Generation (RAG) can reduce this by grounding responses in trusted data.
  - **Transparency:** Explainable AI (XAI) helps users understand recommendations, building trust.
  - **Content Moderation:** Filter harmful content in queries and responses, leveraging tools like Google Cloud's Natural Language API.
- **Security Best Practices:** Secure API key management, data encryption (at rest and in transit), abuse monitoring, and mitigation of LLM-specific threats (data poisoning, prompt injection, data leakage) are essential.