LLM-Based Astrologer Recommendation Engine for Vedaz

Research Task Submission

July 23, 2025

1. Recommended LLM Stack

For building an LLM-based smart recommendation engine for Vedaz, I recommend using **Hugging Face**'s open-source models, specifically **Sentence Transformers** for generating embeddings and a fine-tuned **LLaMA 3.3 70B** or **DeepSeek R1** for semantic matching and recommendation logic.

Why Hugging Face and Open-Source Models?

- **Cost-Effectiveness**: Open-source models like LLaMA 3.3 or DeepSeek R1 are free to use, reducing licensing costs compared to proprietary models like OpenAI's GPT-4.1 or Claude 3.7, which charge per token.[](https://www.helicone.ai/blog/the-complete-llm-model-comparison-guide)
- **Customization**: Hugging Face provides over 60,000 open-source models, allowing fine-tuning on domain-specific data (e.g., astrology-related chat transcripts) for better relevance.[](https://futureagi.substack.com/p/best-llm-api-providers-2025-comparison)
- **Performance**: LLaMA 4 Scout and DeepSeek R1 offer competitive performance, with LLaMA 4 Scout excelling in multimodal tasks and DeepSeek R1 matching GPT-4.1 in some benchmarks.[](https://ai.meta.com/blog/llama-4-multimodal-intelligence/)[](https://www.reddit $HuggingFacesecosystemprovidesextensivedocumentationandcommunity-driventoolslikeTransfor$

Comparison with Alternatives:

- **OpenAI GPT-4.1**: Offers superior performance for general tasks but is expensive (e.g., $15 per million tokens) and lacks transparency due to its proprietary nature.[](https://futureagi.substack.com/p/best-llm-api-providers-2025-comparison)
- **Claude 3.7**: Strong in coding and reasoning but costly and less flexible for self-hosted customization.[](https://www.helicone.ai/blog/the-complete-llm-model-comparison-guide)

For Vedaz, the combination of Sentence Transformers for embeddings and a fine-tuned LLaMA 3.3 or DeepSeek R1 provides a balance of performance, cost, and control.

2. Hosting and Scaling

Hosting: I recommend self-hosting the LLM on **AWS** using **Amazon Bedrock** for managed infrastructure or **SageMaker** for custom deployments. AWS offers robust security, scalability, and integration with tools like **vLLM** for high-performance inference.[](https://futureagi.sub
llm-api-providers-2025-comparison)[](https://research.aimultiple.com/self-hosted-llm/)

Deployment Options:

- **vLLM**: A high-performance engine for efficient LLM serving, supporting distributed execution and NVIDIA/AMD hardware. It uses PagedAttention to reduce memory usage, ideal for handling 50,000 monthly active users.[](https://research.aimultiple.com/self-hosted-llm/)
- **Ollama**: Simplifies local deployment and integrates with Hugging Face models, suitable for initial prototyping and testing.[](https://research.aimultiple.com/self-hosted-llm/)
- **SageMaker**: For production, SageMaker supports model parallelism and auto-scaling, ensuring low latency for high user loads.

Scaling Strategy:

- **Auto-Scaling**: Configure SageMaker endpoints to scale based on request volume, ensuring low latency during peak usage.
- **Quantization**: Use 4-bit quantization (e.g., GGUF or GPTQ) to reduce memory and VRAM requirements, enabling efficient inference on smaller GPUs.[](https://research.aimultip
hosted-llm/)
- **Batching**: Implement continuous batching with vLLM to maximize throughput for concurrent user requests.[](https://research.aimultiple.com/self-hosted-llm/)

3. Monthly Cost Estimation

For 50,000 monthly active users, assuming each user makes 10 requests daily (500,000 requests/month), with an average input of 500 tokens and output of 100 tokens per request:

- **Model**: LLaMA 3.3 70B (4-bit quantized) requires 40 GB VRAM for inference.[](https://research
hosted-llm/)
- **Hardware**: AWS SageMaker with 4x NVIDIA A10G GPUs (24 GB VRAM each, $3.06/hour per instance). Assuming 2 instances for redundancy and load balancing: $\$3.06 \times 2 \times 730 hours = \$4,467.60/month$.**Storage and Networking** : $S3 for model storage (100GB, \$2.30/month) and data transfer costs (1TB, \$90/month)$.
- **Total Estimated Cost**: $4,560/month, excluding development and maintenance costs.

Using OpenAI's GPT-4.1 would cost $7,500/month at $15/million output tokens for 50 million output tokens, making the open-source approach significantly cheaper.[](https://futureagi.substa
llm-api-providers-2025-comparison)

4. Privacy and Safety Concerns

- **Data Privacy**: Self-hosting on AWS ensures user chat data remains on private infrastructure, avoiding data leaks associated with proprietary providers. Encrypt data at rest (AWS KMS) and in transit (TLS).[](https://www.datacamp.com/blog/top-open-source-llms)[](https://research.aimultiple.com/self-hosted-llm/)
- **User Consent**: Implement clear consent mechanisms for storing and analyzing chat history, complying with GDPR and CCPA.
- **Prompt Injection**: Mitigate risks by sanitizing user inputs and using guardrails (e.g., Hugging Face's SafePrompt) to prevent malicious prompts.[](https://futureagi.com/blogs/to 11-llm-api-providers-2025)
- **Bias and Fairness**: Fine-tune models on diverse astrology datasets to avoid biased recommendations. Regularly audit outputs for fairness.
- **Compliance**: Ensure SOC 2 Type II and GDPR compliance for enterprise-grade security, as supported by AWS Bedrock.[](https://www.helicone.ai/blog/the-complete-llm-model-comparison-guide)