Design for MoC AI Search Engine

1. Recommended Tech Stack and Models

The proposed technology stack is engineered to be entirely open-source, highly scalable, and completely deployable within an air-gapped, MeitY-empanelled on-premise data center without requiring any external internet connections or commercial API dependencies.

1.1 Core Infrastructure & Orchestration

- **Frontend Interface:** React (implementing secure Micro-Frontend/MFE architecture for seamless portal integration).
- **Backend Services:** FastAPI (handling asynchronous API gateway duties with native Active Directory RBAC integration).
- **Data Ingestion Engine:** Scrapy (for HTML scraping), Playwright (for dynamic SPA extraction), and Marker (for complex structured PDF parsing).
- **Vector Database:** Milvus. Configured to utilize Reciprocal Rank Fusion (RRF) to intelligently combine dense semantic embeddings with sparse keyword searches.
- **RAG Orchestration:** LlamaIndex (managing multi-modal chunking, vector projections, and hierarchical indexing pipelines).
- **High-Concurrency Inference:** vLLM (leveraging PagedAttention, Tensor Parallelism, and Data Parallelism to ensure maximum throughput under heavy concurrent user loads).
- **Dynamic LLM Orchestration:** Semantic Router (serving as the intelligent traffic controller, routing queries to the appropriate specialized model based on mathematical intent projection without invoking an LLM for the routing decision itself).

1.2 Observability & Analytics

- **LLM Telemetry & Tracing:** Langfuse. Selected for its open-source (MIT) license, allowing full deployment on air-gapped VPCs while providing exhaustive step-by-step tracing of the entire RAG pipeline, cost calculation, and latency monitoring.
- **Business Intelligence (BI) & Analytics:** Grafana. Plugs directly into the Langfuse PostgreSQL database and internal system metrics to provide dynamic, no-code visualization dashboards for non-technical stakeholders to analyze user search themes and infrastructure health.

1.3 Machine Learning Model Stack

To ensure optimal performance, low latency, and comprehensive Indic language coverage, a multi-model paradigm is deployed behind the Semantic Router:

- **Dense/Sparse Embedding:** BGE-M3 (for highly accurate, cross-lingual dense, sparse, and multi-vector text embeddings).
- **Vision Embedding:** SigLIP (for encoding scraped images and video keyframes into the

Milvus vector space for multimodal retrieval).

- **Speech-to-Text (Voice Search):** AI4Bharat Bhashini IndicConformer. An open-source STT model providing highly accurate real-time voice transcription across 22 scheduled Indian languages.
- **Dedicated Translation & Transliteration:** IndicTrans2. An open-source Neural Machine Translation (NMT) transformer that provides state-of-the-art, bidirectional translation and transliteration natively supporting all 22 scheduled Indic languages.
- **General Query LLM:** Meta Llama 3.1 8B Instruct. The primary workhorse for standard information retrieval and fast synthesis.
- **Long-Context LLM:** Mistral NeMo 12B. Utilized for processing massive archival documents requiring deep context windows (up to 128k tokens).
- **Multimodal & Complex Linguistic LLM:** Gemma 3. Utilized for reasoning over retrieved images, handling multimodal user queries, and providing robust cross-lingual support for advanced queries.

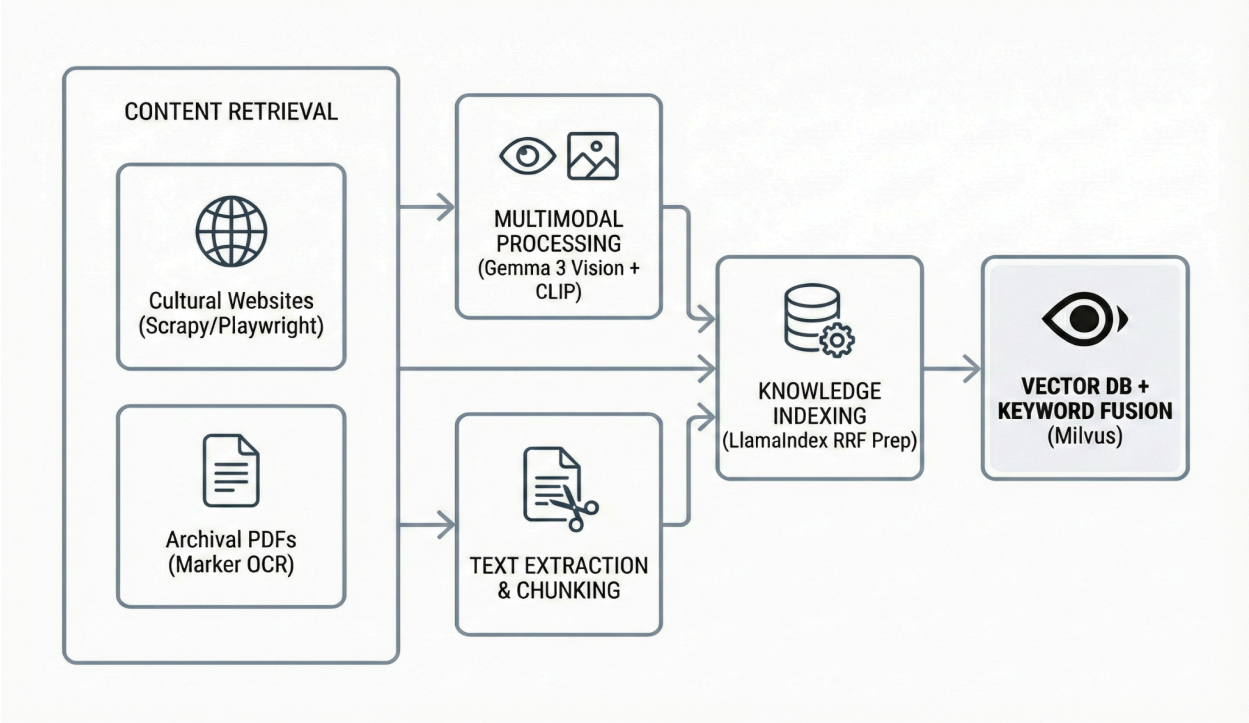
---

2. Architecture and Data Flow

The system operates across two primary lifecycles: the asynchronous background ingestion pipeline, and the real-time synchronous inference pipeline.

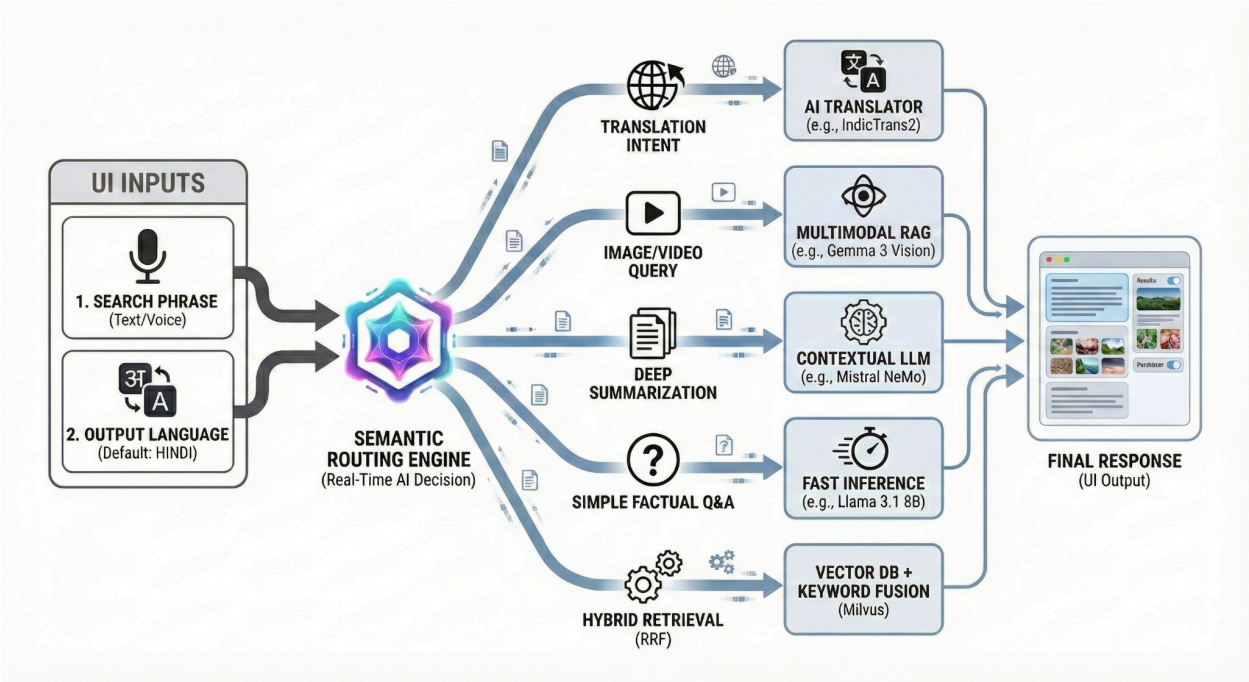
Phase 1: Ingestion and Knowledge Indexing

1. **Content Retrieval:** Scrapy targets static text across the 30 ring-fenced domains, while Playwright executes dynamic JavaScript to harvest modern portal text. Marker simultaneously processes archival PDFs, preserving tables and headers as structural Markdown.
2. **Multimodal Processing & Embedding:**
    - Extracted text is chunked hierarchically by LlamaIndex and embedded using the BGE-M3 model.
    - Images and video keyframes are extracted, optionally captioned by Gemma 3, and embedded using SigLIP.
3. **Knowledge Indexing in Vector DB:** Both textual and visual embeddings are stored within Milvus alongside comprehensive metadata (dates, origins, categories) to enable temporal pre-filtering and event discovery.



Phase 2: Real-Time RAG Query Execution

1. **User Input & STT:** A user submits a query via text or voice. If voice is used, the Bhashini IndicConformer transcribes the audio into Hindi or English text in real-time.
2. **Semantic Routing:** FastAPI passes the query to the Semantic Router. The router calculates the mathematical intent of the query and dynamically selects the processing path:
   - *Translation Required?* Routes to IndicTrans2.
   - *Heavy Document Summarization?* Routes to Mistral NeMo 12B.
   - *Multimodal/Visual Query?* Routes to Gemma 3.
   - *Standard Query?* Routes to Llama 3.1 8B.
3. **Hybrid Retrieval (RRF):** LlamaIndex queries Milvus. Milvus performs a Dense Semantic Search (for contextual meaning) and a Sparse BM25 Search (for exact keyword/name matching). These two distinct result sets are mathematically merged using Reciprocal Rank Fusion (RRF) to surface the most highly relevant multimodal contexts.
4. **Generative Inference:** The selected LLM, served via vLLM with Tensor/Data Parallelism, receives the RRF-fused context and the user query. It synthesizes a highly accurate, linguistically appropriate response.
5. **Observability Logging:** Simultaneously, Langfuse records the execution times, token usage, and retrieval chunks into a secure PostgreSQL database. Grafana polls this database to update the live BI analytics panel.



---

3. Fulfillment of Tender Requirements

This architectural layout has been strictly mapped to the requirements stipulated in the Ministry of Culture's RFP document.

RFP Requirement Clause	Quote from Tender Document	Architectural Resolution
Ring-fenced Retrieval	*"Semantic Search should be ring fenced to 30 websites of Ministry of Culture and its associated offices."* (Page 7)	The Scrapy/Playwright ingestion engines are strictly domain-locked to the 30 specified URLs, ensuring no external or hallucinated web data enters the Milvus vector database.
Multilingual Support	*"Comprehensive bilingual support (English/Hindi) for search queries, real-time translation, transliteration..."* (Page 21)	Native integration of **IndicTrans2** handles pure translation/transliteration , while **Gemma 3** and **BGE-M3** handle cross-lingual embeddings

		and generative tasks.
Multimedia Integration	*"Where applicable, search results must include relevant images and videos sourced from the originating websites."* (Page 7)	Addressed via **SigLIP** visual embeddings stored in Milvus, combined with **Gemma 3's** robust multimodal comprehension, enabling LlamaIndex to retrieve and reason over visual data natively alongside text.
Dynamic LLM Orchestration	*"AI control layer that automatically selects and manages multiple Large Language Models (LLMs) to deliver the best balance... (At least 3 models)."* (Page 21)	Fully satisfied by the **Semantic Router**, which dynamically dispatches queries across Llama 3.1 8B, Mistral NeMo 12B, Gemma 3, and IndicTrans2 based on zero-shot vector intent.
Event Discovery	*"The search functionality should also index and display event listings across all connected sites."* (Page 7)	Implemented using **LlamaIndex metadata filtering**. Dates and event tags are stored in Milvus, allowing the system to execute temporal filters before performing vector searches.
Auditability & Traceability	*"Every response must link to source documents, with logs for all queries and system changes."* (Page 21)	**Langfuse** acts as the dedicated, open-source tracing layer, logging every prompt, vector chunk retrieved, and generated response into an immutable database.
Analytics Panel	*"Provision of analytics panel to asses user queries by content theme, type, category which can help*	**Grafana** is deployed over the Langfuse telemetry database, offering administrators a

	training the model in better ways." (Page 20)	drag-and-drop BI dashboard to visually track themes, trends, and LLM bottlenecks.
Voice Search	*"Implement speech-to-text functionality with voice icon integration for intuitive voice search interactions."* (Page 21)	Satisfied by integrating the **AI4Bharat Bhashini IndicConformer**, specifically designed for low-latency voice-to-text transcription across Indian languages.
Latency and Concurrency	*"Maximum query response time shall not exceed 3 seconds... System must support projected concurrent users (1000)"* (Pages 22)	The **vLLM** inference engine uses PagedAttention and continuous batching to maximize hardware utilization. Combined with load balancers and GPU parallelization, it guarantees sub-3-second streaming generation even at high loads.
Security & Data Sovereignty	*"Compliance: Security audit certification through cert-in empanelled agencies, encryption at rest (AES-256)... deployed within a MeitY-empanelled Indian Data Centre"* (Pages 21)	Every proposed component (Langfuse, Grafana, vLLM, Milvus) is 100% open-source and deployable completely offline. Storage volumes utilize AES-256, ensuring absolute compliance with MeitY and Cert-In data residency mandates.