THE NATIONAL
COLLEGE
JAYANAGAR

×

HACK
WITH INDIA

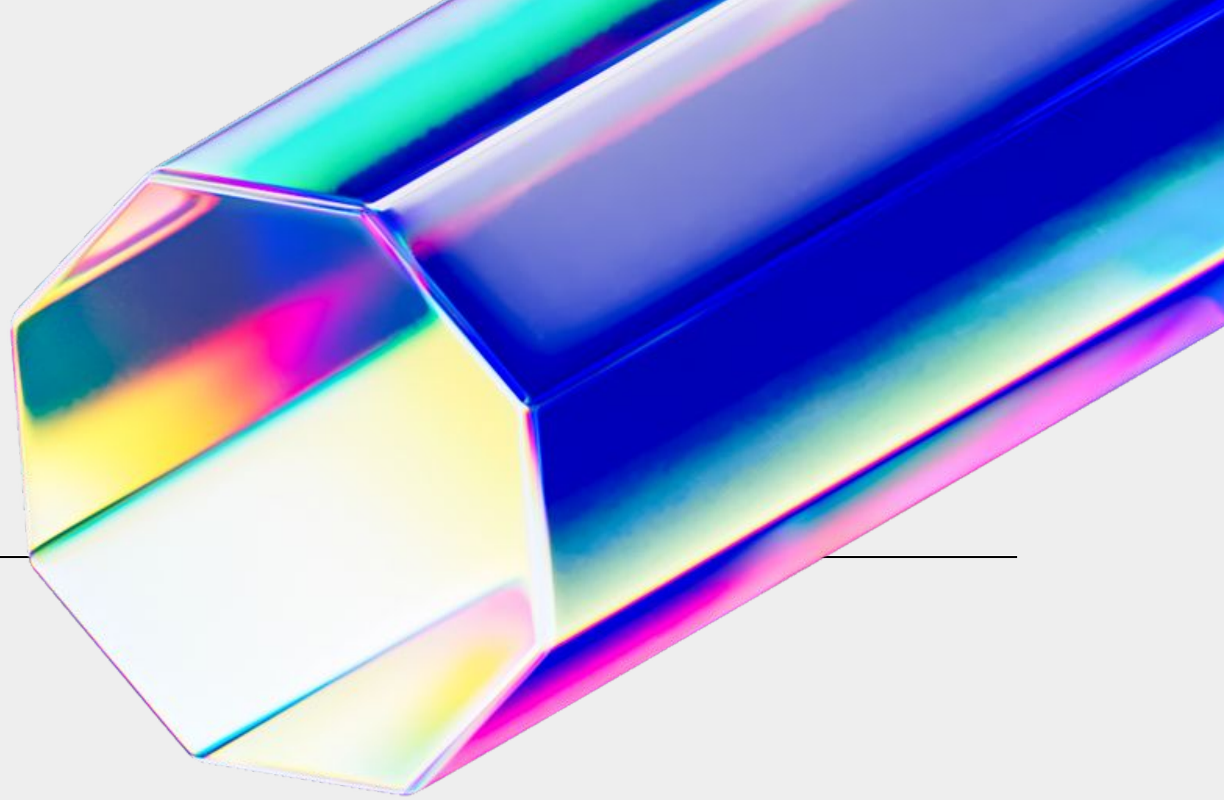
Import
{hackademia}

TEAM: ALPHA

TEAM MEMBERS:

Akshat Kumar
Beeravalli Anjali
Rishikeshwaran M

PROBLEM STATEMENT



The Challenge

Despite advancements in AI, voice-based querying of complex, domain-specific knowledge bases, such as those in healthcare, remains largely ineffective. Clinicians, researchers, and even patients often struggle to access accurate and relevant information due to outdated, inefficient search methods.

Why This Matters

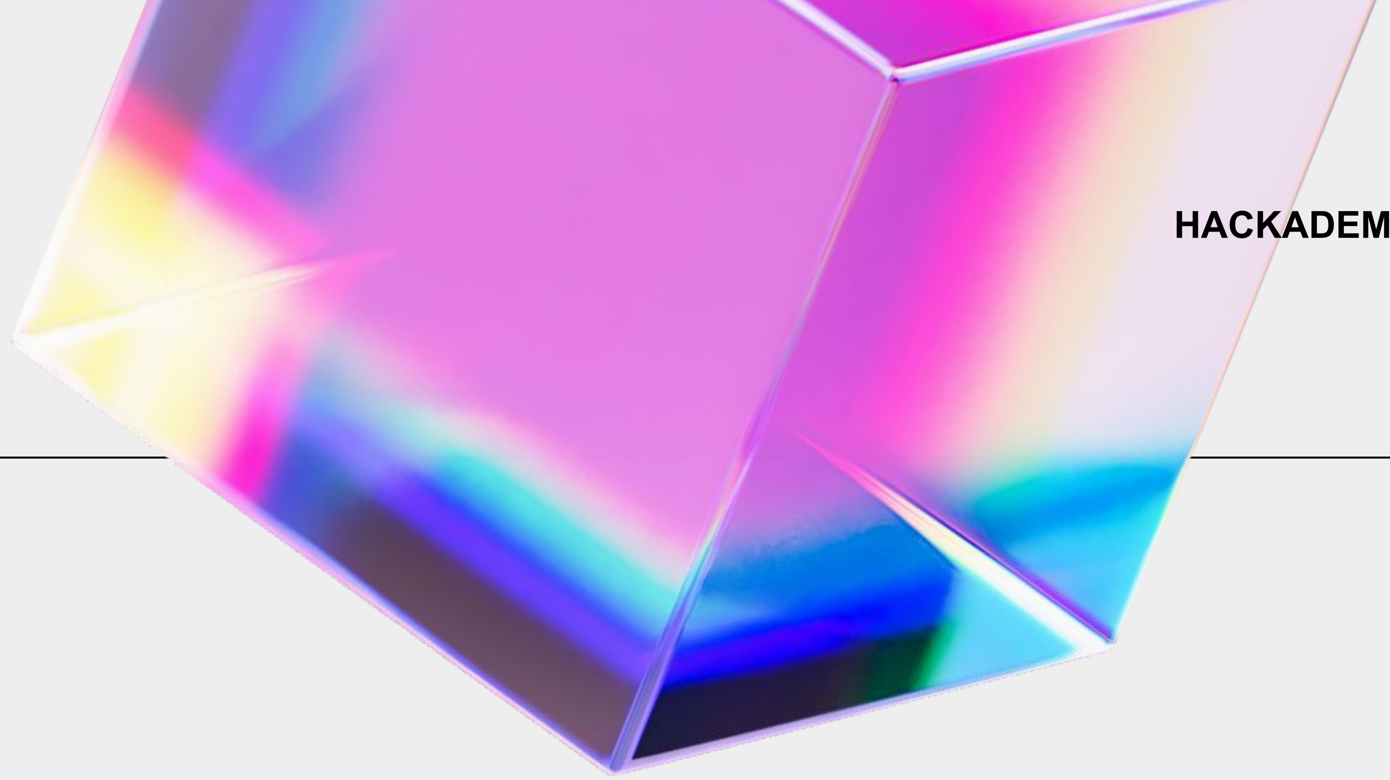
Traditional keyword-based searches are not only time-consuming but also mentally demanding, particularly in fast-paced or hands-free environments like clinical settings. Interrupting workflows to type queries disrupts productivity, especially during voice-first scenarios such as medical rounds or on-the-go consultations.

The Reality Gap

Current Experience	Ideal Experience
Repetitive clicks, forms, and manual filters	Seamless, real-time voice interaction
Information overload and irrelevant results	Context-aware, targeted document retrieval
Risk of AI hallucinations and unverifiable data	Trusted, evidence-backed answers with clear sources

This disconnect highlights the need for intuitive, voice-first systems that provide fast, reliable, and context-sensitive access to critical knowledge—bridging the gap between information overload and actionable insight.

SOLUTION



High-Level System Architecture

The system enables end-to-end voice-based querying of domain-specific knowledge bases. Spoken input is converted to text using an automatic speech recognition (ASR) model (e.g., Whisper), followed by intent parsing. A retriever module embeds the query using a domain-tuned model (e.g., BioBERT) and searches a vector database (e.g., via LlamaIndex) to identify relevant documents or passages. The retrieved context is combined with the query and passed to a large language model (LLM), which generates a grounded response. Guardrails are applied to minimize hallucinations, enforce source fidelity, and ensure safe completions. The response is then delivered back to the user through text-to-speech (TTS), optionally citing its source.

Flow: Voice Input → ASR → Retriever → LLM + Guardrails → TTS→ Happy User

Data & Knowledge Pipeline

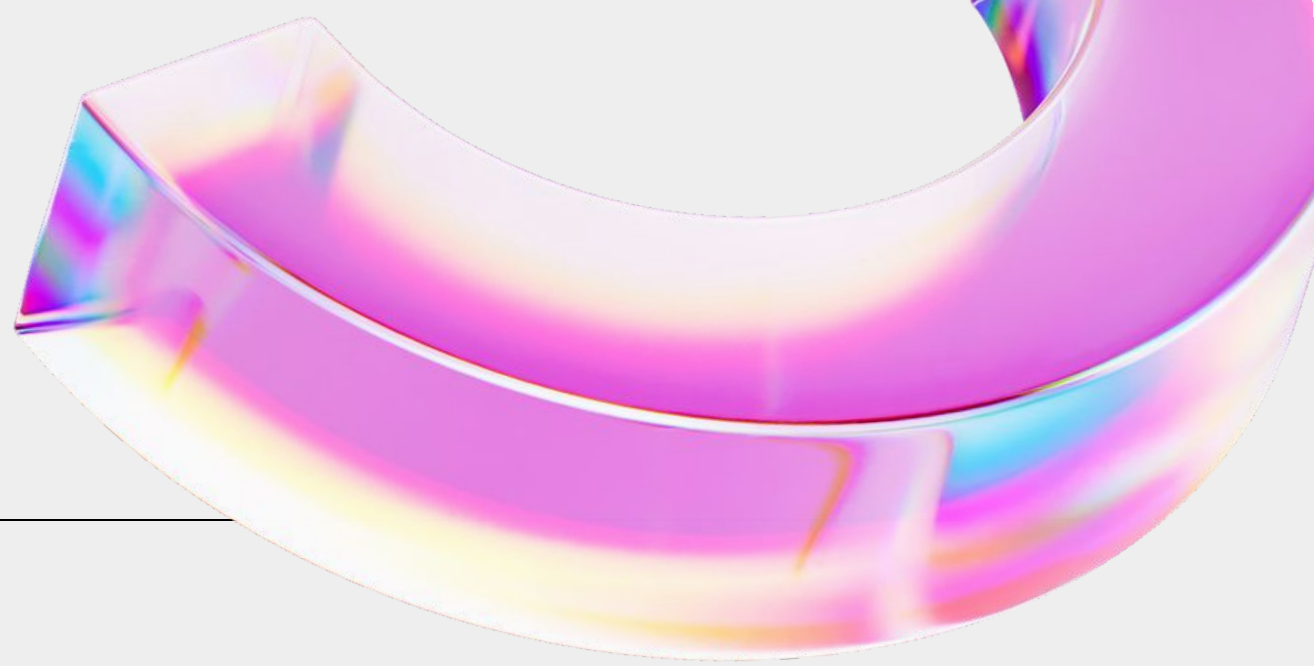
Structured clinical content—such as guidelines, FAQs, or internal documentation—is ingested, converted to text, chunked, and embedded into a vector database. Domain-specific embeddings (e.g., PubMedBERT) improve semantic relevance, while sources are ranked by credibility to support high-quality retrieval.

Guardrails & Transparency

To ensure safety, prompts are constrained to use only retrieved content. Responses with unverifiable claims are blocked or flagged, and uncertainty disclaimers are provided when appropriate. Source citations, audit logs, and response traceability enhance explainability and compliance.

Feedback & Continuous Learning

A human-in-the-loop approach enables clinicians to validate and correct responses. Feedback is looped back into the system to refine retrieval ranking and tighten safety guardrails. Evaluation metrics—including retrieval precision, response latency, and guardrail exceptions—are monitored via performance dashboards.



PROCESS/WORKFLOW (IMAGES ARE PREFERRED)

Benefits & Value Propositions

- **Efficiency:** Voice-first interaction streamlines information access.
- **Credibility:** Responses are grounded in vetted, domain-relevant sources.
- **Compliance:** Built-in guardrails support safe, auditable communication.
- **Maintainability:** Lightweight updates via knowledge ingestion eliminate the need for frequent LLM retraining.

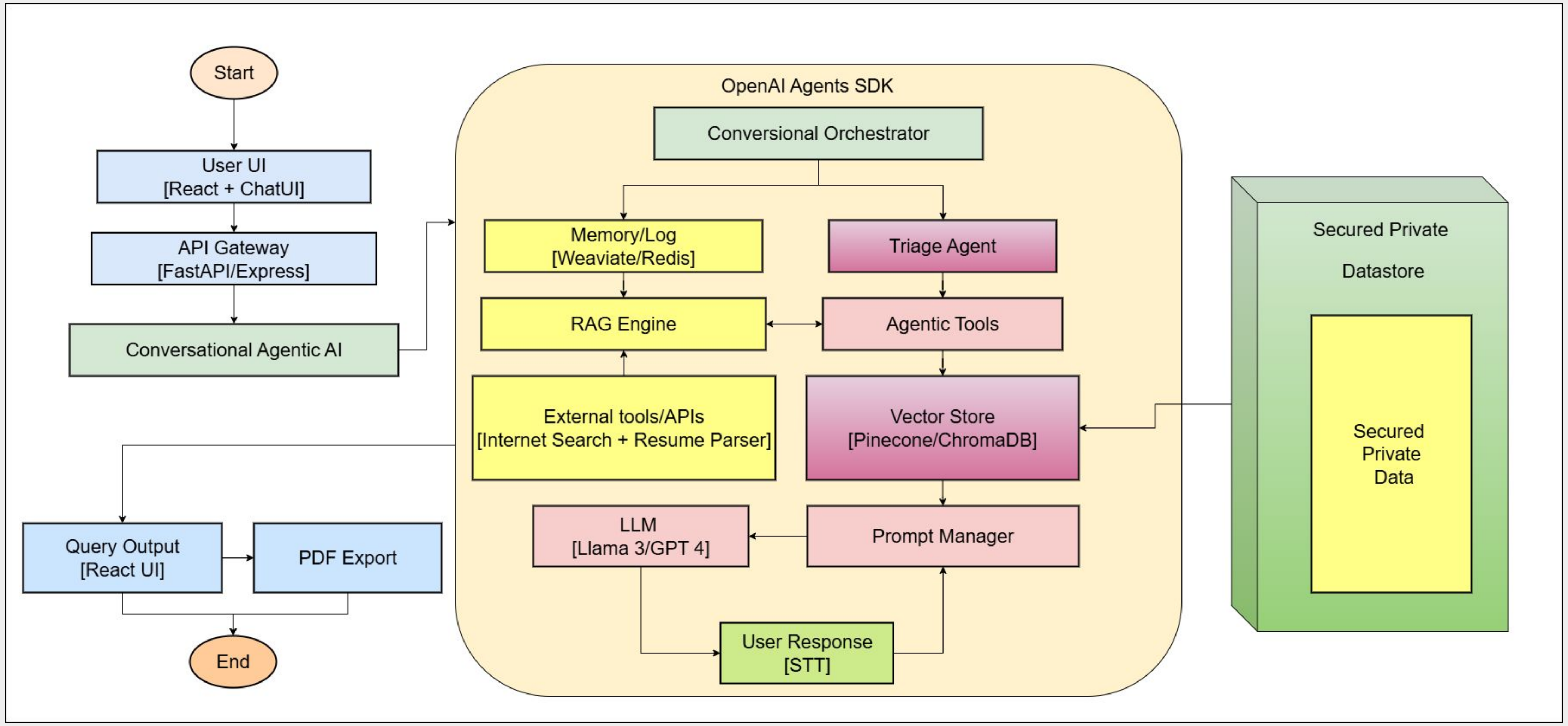
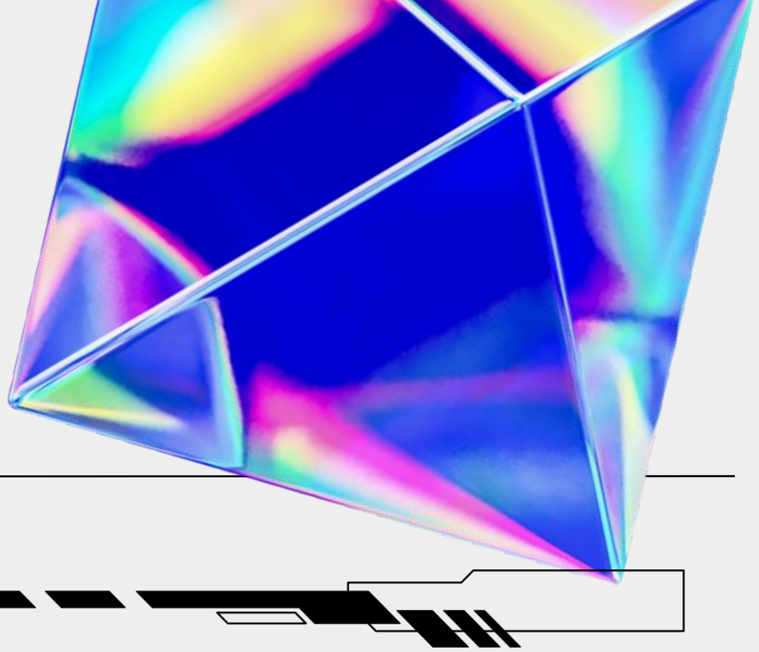


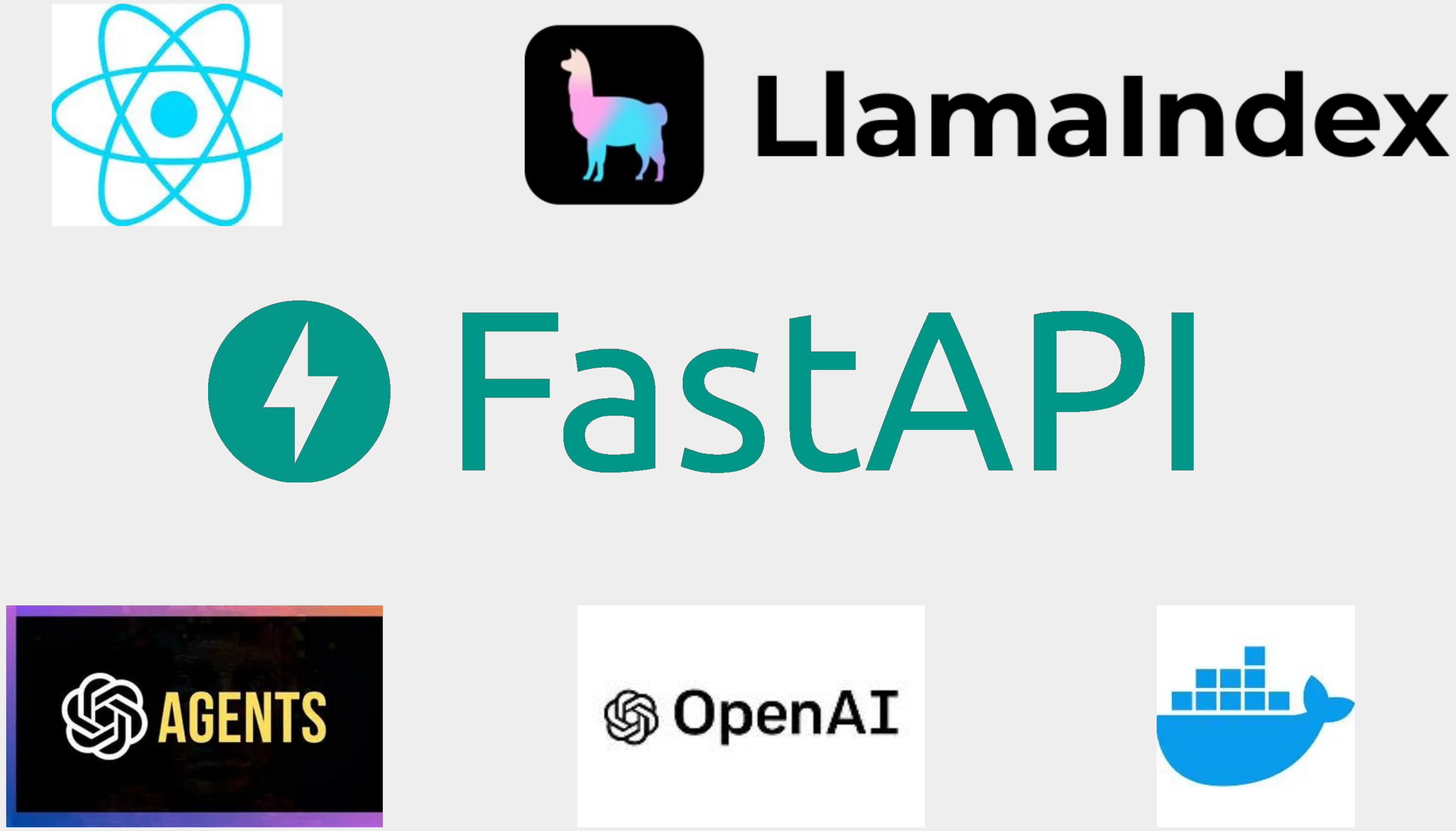
Fig: Architecture Diagram

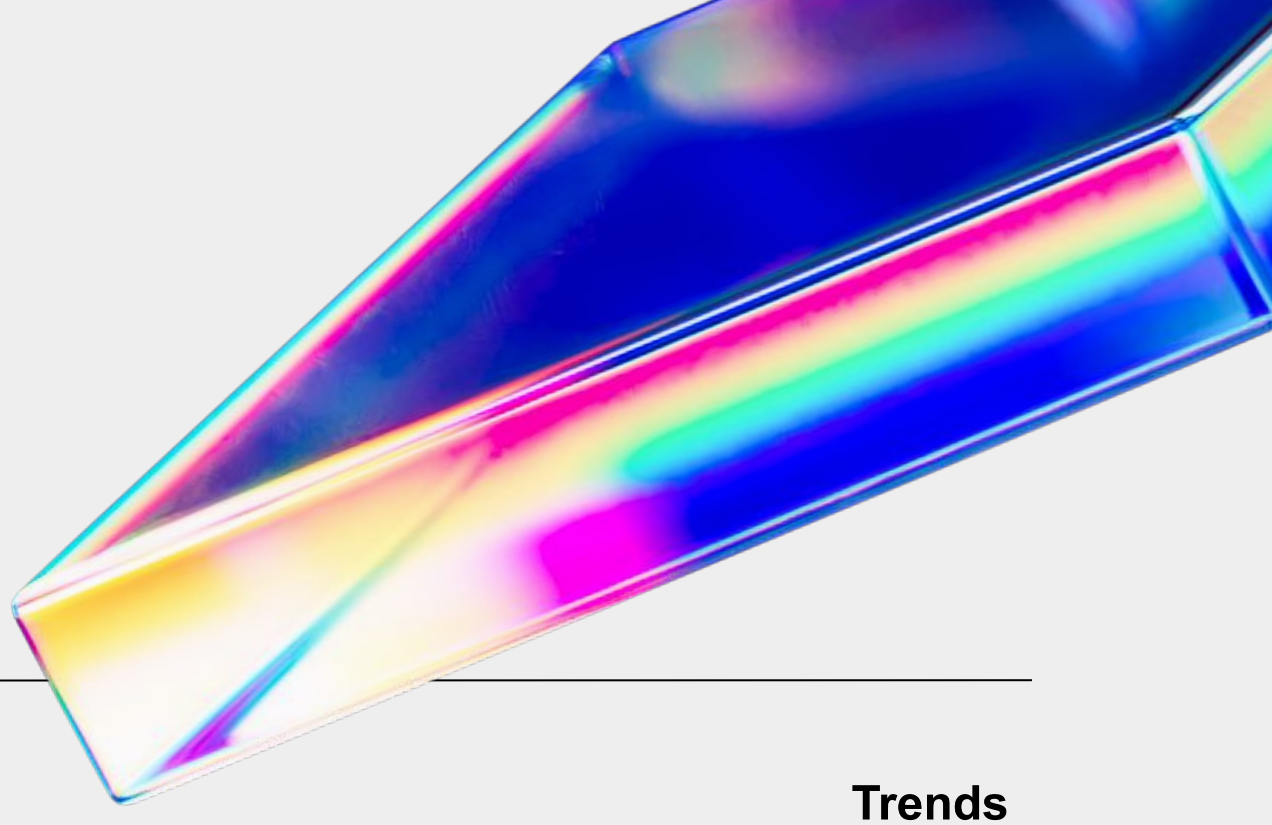
TECH STACK



Technology Stack

- **Frontend:** React.js with a voice interface for input, playback, and user feedback.
- **Backend:** FastAPI manages ASR, retrieval, generation, and TTS via RESTful endpoints.
- **Retrieval/LLM Layer:** Powered by LlamaIndex or OpenAI Agent SDK to orchestrate RAG workflows.
- **ASR/TTS:** Implemented via Whisper or commercial speech services.
- **Deployment:** Containerized via Docker with orchestration support (e.g., Kubernetes), integrated with a vector database and CI/CD pipeline.





REAL WORLD USAGE /FEASIBILITY

Market & Adoption Trends

Voice-enabled RAG (retrieval-augmented generation) systems are gaining traction in healthcare, improving efficiency and patient engagement. Institutions like Northwell Health and Boston Children's Hospital have automated triage and reduced call center loads. Infinitus's "Eva" voice agent handles insurance queries at scale, replacing the equivalent of 100+ FTEs. Apollo Hospitals reported a 46% increase in provider productivity using Augnito's Spectra voice platform.

Patient-Facing Applications

Voice assistants are increasingly used for post-discharge care, medication reminders, and symptom escalation. One clinical study reported over 99% triage accuracy without adverse events. These systems are multilingual and personalized, helping improve patient understanding and adherence to care plans.

Specialized Use Cases & Research

Research prototypes show promise across chronic care and emergency response. AsthmaBot supports multimodal asthma management with low hallucination risk. A diabetes screening assistant used voice to detect early symptoms in older adults, while CognitiveEMS delivers real-time, voice-guided support to emergency responders under latency constraints.

Feasibility & Operational Readiness

- **Governance & Compliance:** Integration with provider-managed content ensures auditability and minimizes hallucination.
- **Scalability:** Vector-based indexing allows rapid updates without retraining LLMs.
- **Human Oversight:** Safeguards include audit sampling and escalation protocols for ambiguous queries.
- **Deployment:** Containerized services (FastAPI, ASR/TTS, vector DBs) support clinical-scale use with response times within 1–2 seconds.