lyzr®

Private & Secure

AI Agents

for Enterprises

- Fully Integrated Private SDKs
- Runs On Your Cloud
- Meets AI Safety standards

Webinar #1

How Enterprises Are Adopting GenAI?

Date: 24th Jan 2024
Presenter: Siva Surendira, Founder, Lyzr AI

What’s in the menu today?

1. What is Generative AI?
2. What are the Pain Points faced by Enterprises?
3. Layers of an Enterprise Grade Generative AI Stack?
4. How to choose the right LLMs, Vector Stores, Agent Frameworks?
5. Decoding a S-O-T-A Chatbot Architecture
6. Popular Use cases

We spoke to Enterprise CIOs and Leading Tech Startups on GenAI Adoption

“Private Alternate For SaaS”

“I like the features of a GenAI SaaS product, like Chatbase, but prefer need a fully hosted solution that runs privately in our cloud or data center.”

“Enterprise alternate for open-source”

“Open-source frameworks are good for prototypes. But the enterprise production workloads should be more secure and meet AI Safety standards.”

“No Building Blocks, Please”

“We don’t want to spend months building a POC. We would like platforms that are mature, powerful and super abstracted, like Snowflake.”

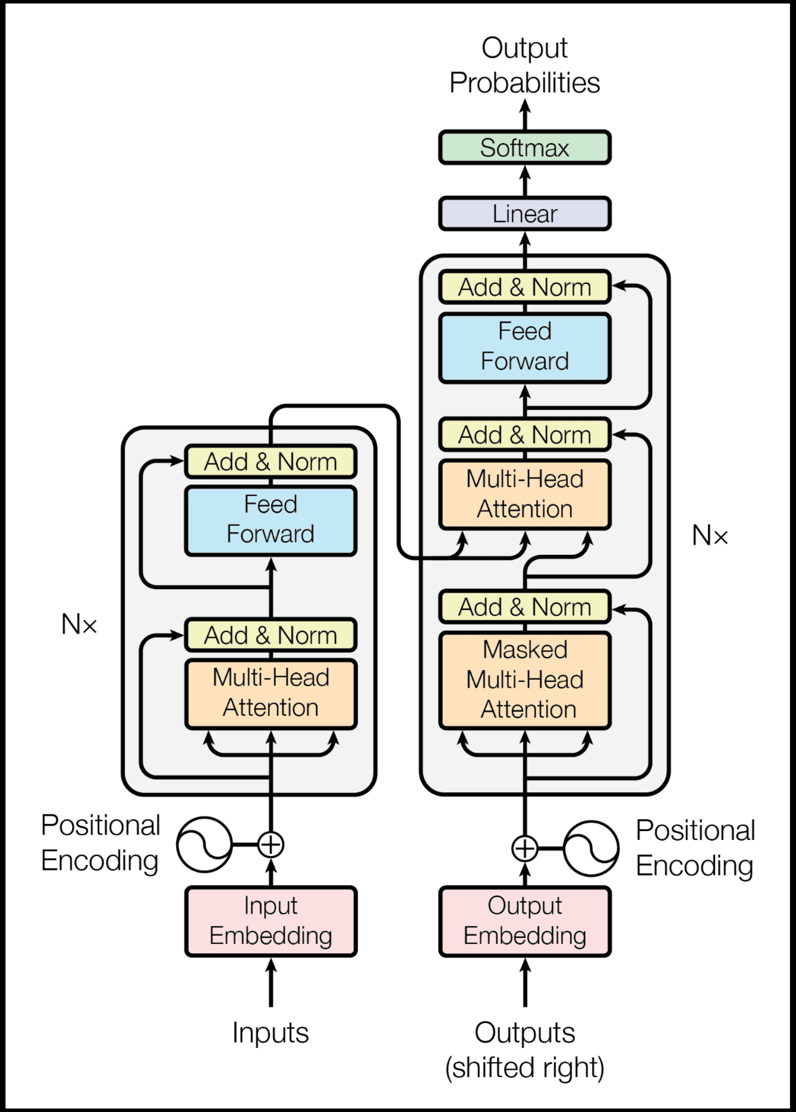
“Predictable Pricing with 24*7 Support”

“24*7 support and predictable pricing are key considerations when it comes to enterprises adopting new technologies.”

1. What is Generative AI?

2. How Transformers Work?

3. Why GPT-4 is the SOTA LLM?



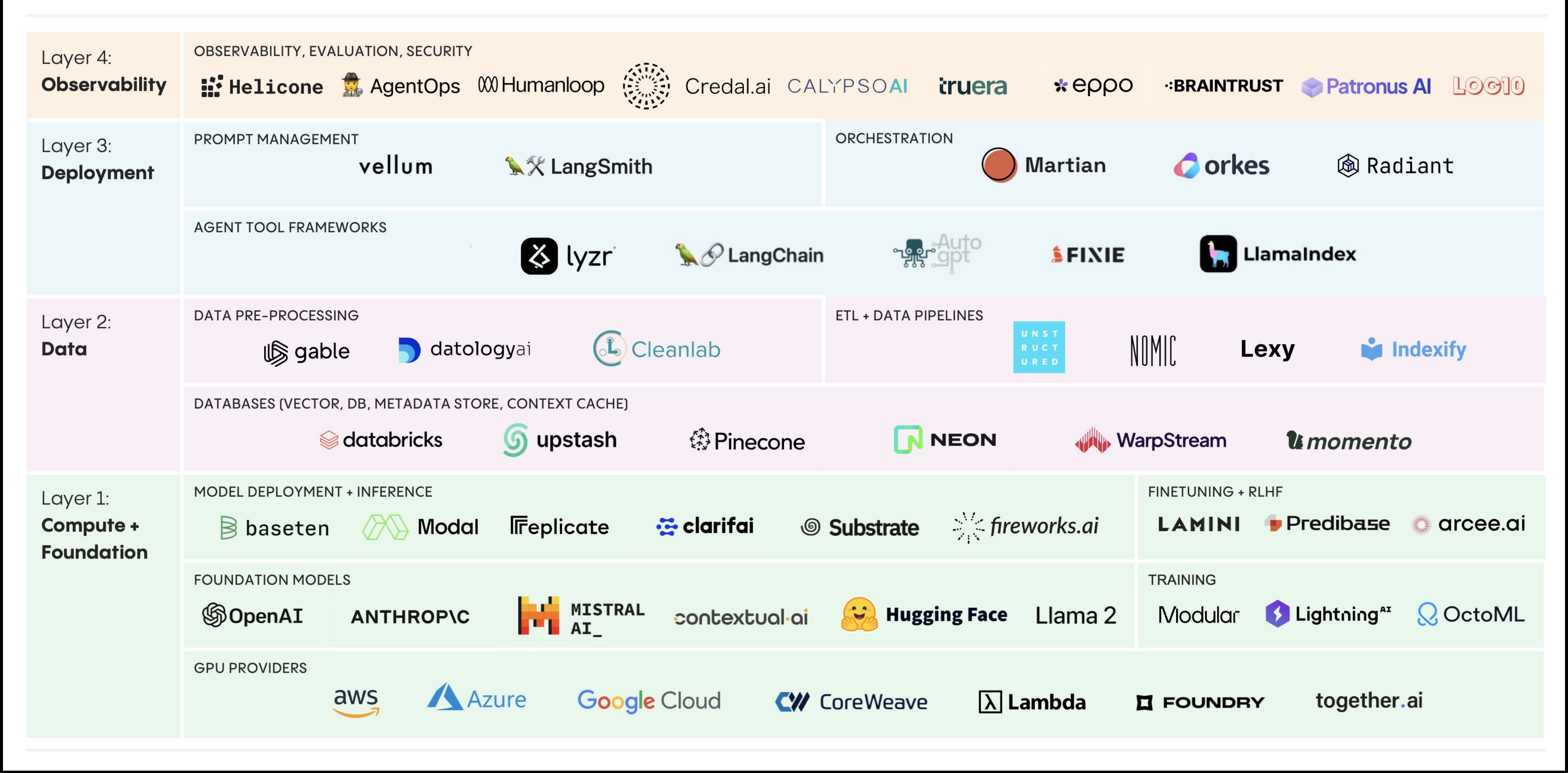
lyzr

Private Agent SDKs for Enterprise

Layers of an Enterprise-Grade Generative AI Stack

Compliance	Compliance and security standards	ISO 42001, SOC2, HIPAA, GDPR
Observability	Monitor and track GenAI apps performance. LLM operations in a nutshell.	Lyzr AIMS, Portkey, LLM.Report
Custom Workflows	The custom application logic and UI UX components of an application	Streamlit, React
Agent Framework	The tooling layer which interactions with LLMs and helps you build generative AI applications	Lyzr AI, Langchain, DSPy, LlamaIndex, Cohere
Data Stores	Specialized data stores for LLM apps like Vector Databases	Pinecone, Weaviate, PGVector, ChromaDB
Foundational Models	Large Language Models and Diffusion Models	GPT-4, Claude, Gemini, Llama2 7B, Mistral, Phi2.7, Bert 1B, MidJourney, Stability AI, DallE3
Infrastructure	The compute power required to run the models, apps, databases	AWS, Amazon Bedrock, Azure, Google Cloud, Google Vertex AI, Nvidia

The Growing Modern GenAI Stack



Layer 4: Observability

OBSERVABILITY, EVALUATION, SECURITY
- Helicone
- AgentOps
- Humanloop
- Credal.ai
- CALYPSOAI
- truera
- eppo
- BRAINTRUST
- Patronus AI
- LOG10

Layer 3: Deployment

PROMPT MANAGEMENT
- vellum
- LangSmith

ORCHESTRATION
- Martian
- orkes
- Radiant

AGENT TOOL FRAMEWORKS
- lyzr
- LangChain
- Auto gpt
- FIXIE
- LlamaIndex

Layer 2: Data

DATA PRE-PROCESSING
- gable
- datologyai
- Cleanlab

ETL + DATA PIPELINES
- UNSTRUCTURED
- NOMIC
- Lexy
- Indexify

DATABASES (VECTOR, DB, METADATA STORE, CONTEXT CACHE)
- databricks
- upstash
- Pinecone
- NEON
- WarpStream
- momento

Layer 1: Compute + Foundation

MODEL DEPLOYMENT + INFERENCE
- baseten
- Modal
- replicate
- clarifai
- Substrate
- fireworks.ai

FINETUNING + RLHF
- LAMINI
- Predibase
- arcee.ai

FOUNDATION MODELS
- OpenAI
- ANTHROPIC
- MISTRAL AI_
- contextual.ai
- Hugging Face
- Llama 2

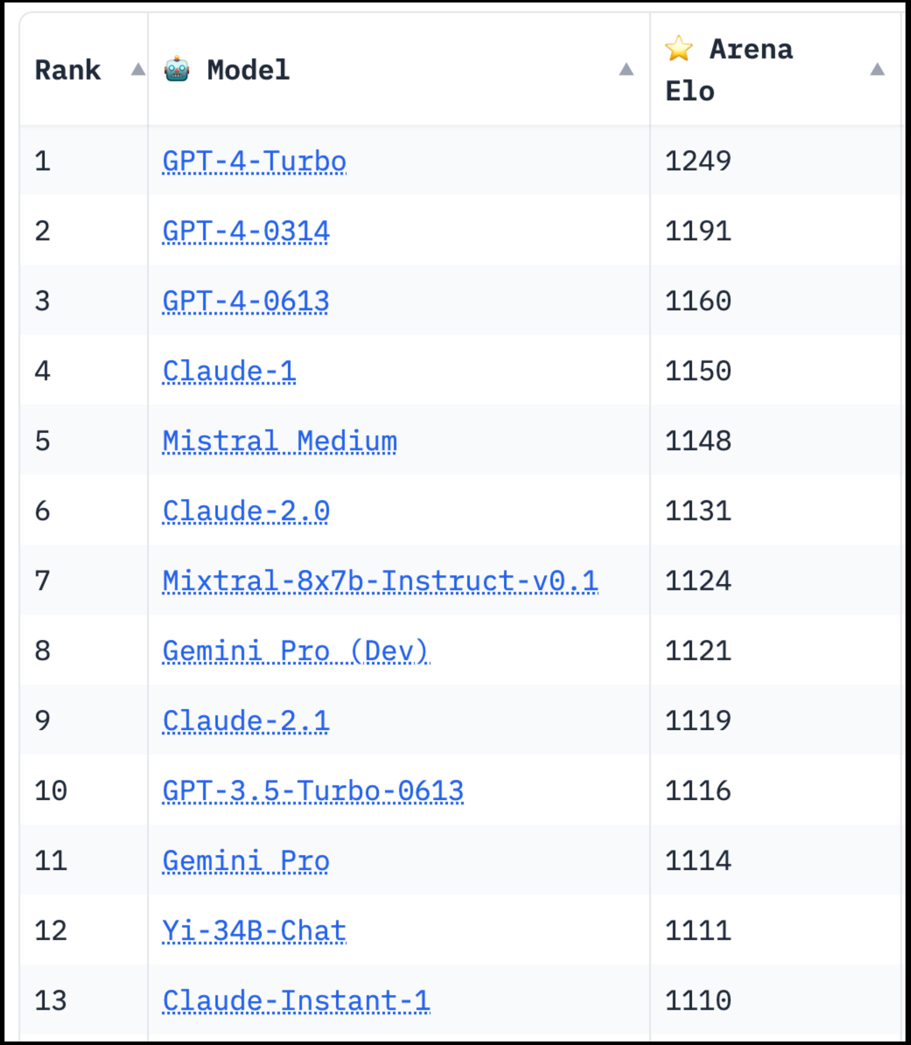
TRAINING
- Modular
- Lightning AI
- OctoML

GPU PROVIDERS
- aws
- Azure
- Google Cloud
- CoreWeave
- Lambda
- FOUNDRY
- together.ai

Choosing the Right AI Stack is Hard

How to choose the right LLMs?

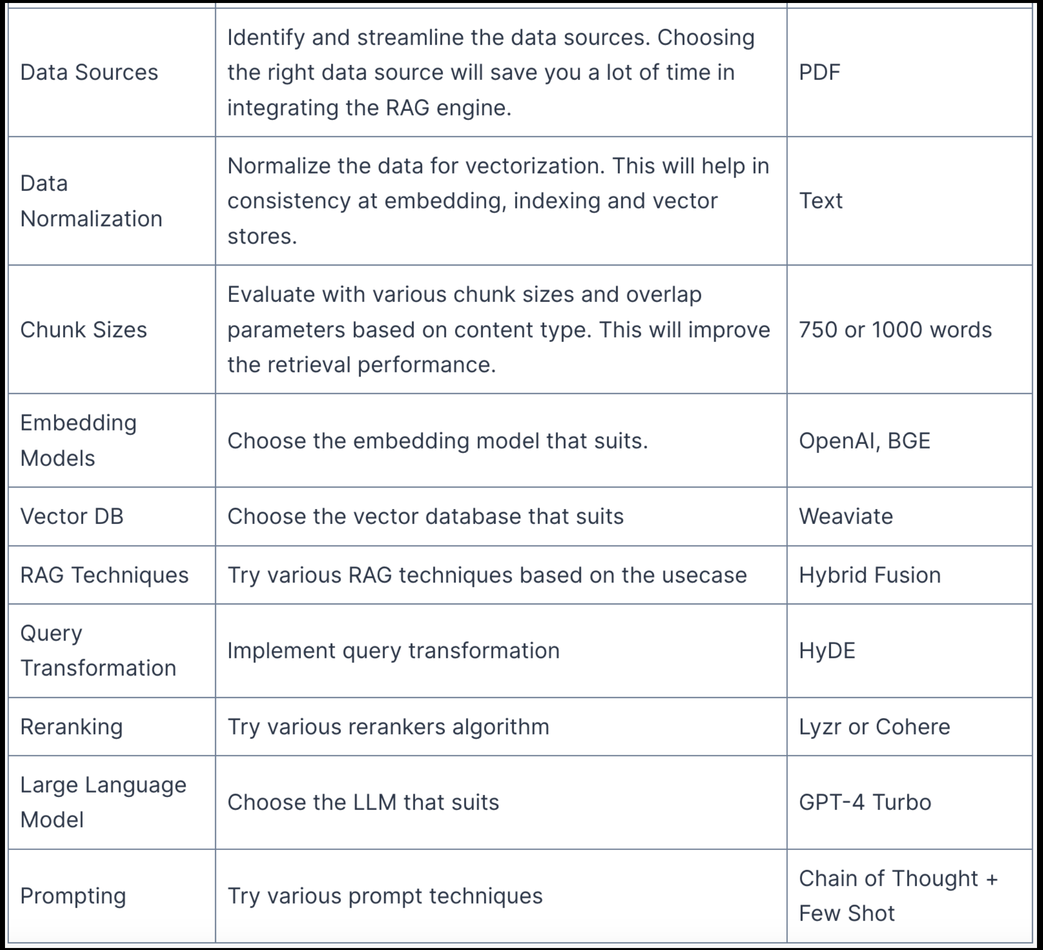
Large Knowledge Bank & Better at Reasoning	GPT-4
Good Reasoning Skills + Good at NLP	GPT 3.5 Claude Mixtral Medium
Data Privacy & Security + Good at NLP	Mixtral 8 x 7B Llama2 7B, 13B, 70B Falcon 180B
Responsible AI	BERT 1B Phi 2.7B
Good at Skills (Programming)	Code Llama 34B



Rank ▲	🤖 Model ▲	⭐ Arena Elo ▲
1	GPT-4-Turbo	1249
2	GPT-4-0314	1191
3	GPT-4-0613	1160
4	Claude-1	1150
5	Mistral Medium	1148
6	Claude-2.0	1131
7	Mixtral-8x7b-Instruct-v0.1	1124
8	Gemini Pro (Dev)	1121
9	Claude-2.1	1119
10	GPT-3.5-Turbo-0613	1116
11	Gemini Pro	1114
12	Yi-34B-Chat	1111
13	Claude-Instant-1	1110

How to choose the right Vector Database?

Data Source	Structured or Unstructured
Purpose of Usage	Search Content Generation Chat
RAG Techniques	Semantic Search Query Fusion
License	OpenSource Enterprise
Support	Enterprise Community



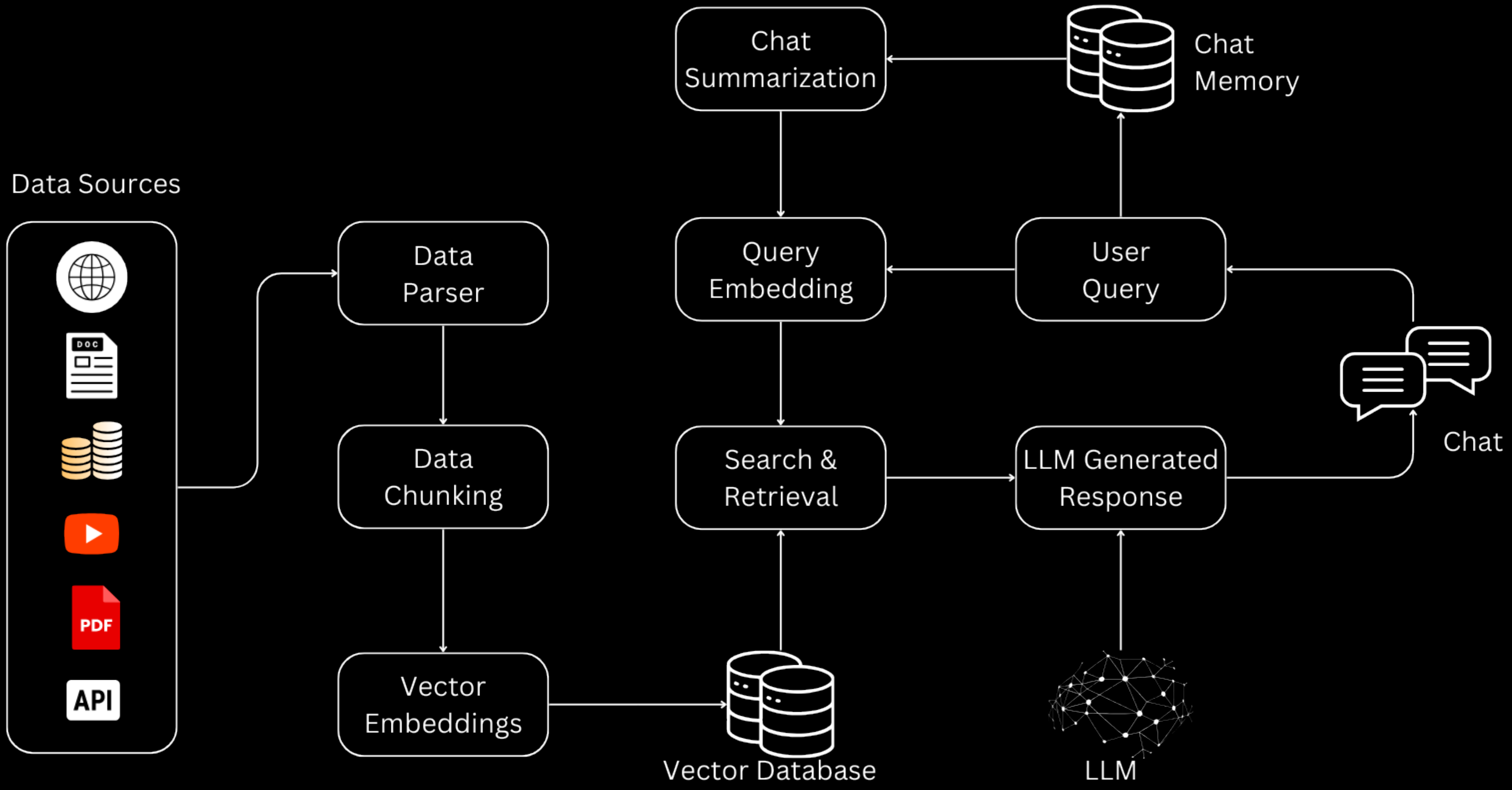
Data Sources	Identify and streamline the data sources. Choosing the right data source will save you a lot of time in integrating the RAG engine.	PDF
Data Normalization	Normalize the data for vectorization. This will help in consistency at embedding, indexing and vector stores.	Text
Chunk Sizes	Evaluate with various chunk sizes and overlap parameters based on content type. This will improve the retrieval performance.	750 or 1000 words
Embedding Models	Choose the embedding model that suits.	OpenAI, BGE
Vector DB	Choose the vector database that suits	Weaviate
RAG Techniques	Try various RAG techniques based on the usecase	Hybrid Fusion
Query Transformation	Implement query transformation	HyDE
Reranking	Try various rerankers algorithm	Lyzr or Cohere
Large Language Model	Choose the LLM that suits	GPT-4 Turbo
Prompting	Try various prompt techniques	Chain of Thought + Few Shot

How to choose the right Agent Tool Framework?

Langchain	Oldest and most popular stack. Known for chains & agents.
Lyzr AI	Low-code fully-integrated Agent SDKs for rapid development
LlamaIndex	RAG at its center. Moving to agents.
DSPy	A program first Stanford backed framework

ai-management
app-scalability development-time
functionality ai-security
complexity
data-privacy flexibility
enterprise-support
learning-curve integrations
success-stories llmops
agent-versatility

How does an Enterprise-Grade Customer Support Agent look like?



So, what are some of the popular use cases?

Chat	Customer Support, Lead Generation, Automated Helpdesk, Process Copilots
Search (RAG)	Enterprise Search, Context-Aware Apps, Document Search, Knowledge Base
Data	Synthetic Data Generation, Conversational Analytics, Data Annotation
Generators & Summarizers	Marketing Content, Product Docs, Customer Reports, Contracts, Email Templates
Automation	Meeting Summarizer, Auto SDR, Auto Blog Writer, Auto Claims Processing,

Exclusive Webinar Offer

Your first Generative AI MVP development is on us. Here is what you get,

- An MVP built for you on your idea
- Includes custom app development
- 1-year Lyzr Business License

Contact

siva@lyzr.ai
ani@lyzr.ai

Website

www.lyzr.ai

Time for some Q&A

Contact

siva@lyzr.ai
ani@lyzr.ai

Website

www.lyzr.ai