FOUNDATIONALLM

FoundationalLLM

Introduction

Joel Hulen

General Manager – Solliance Training

AI Architect - FoundationaLLM

Lino Tadros

AI Solution Architect– Solliance

Chief Evangelist- FoundationaLLM

FOUNDATIONALLM

A platform for deploying Gen AI in the enterprise.

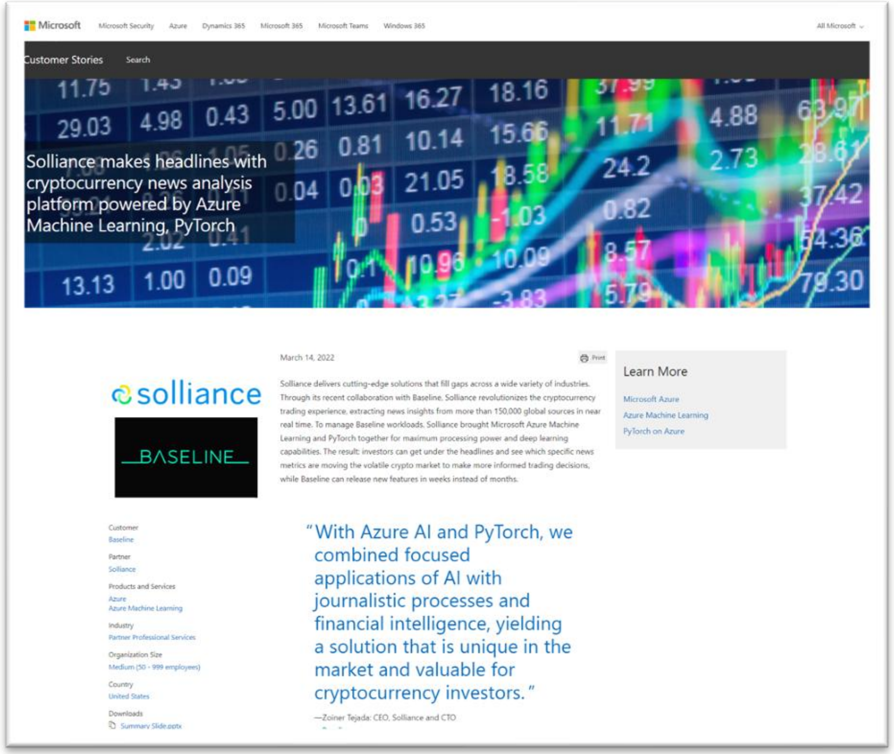
scaling

securing

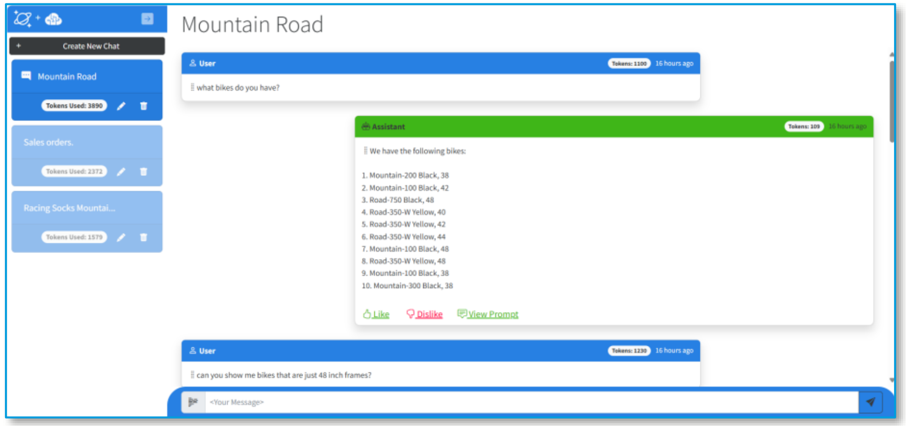
governing

Origin Story

Building GenAI solutions well before ChatGPT moment



Partner to MS Engineering



We created the Bring Your Own Copilot accelerator for the CosmosDB team

Reference solution for how to build your own production-ready Copilot in Azure

[GitHub: Azure/BuildYourOwnCopilot]

The Problems

Enterprise GenAI is complicated, time consuming & capital intensive.

Enterprises need comprehensive platforms that are compliant for high-risk data
- not components, toys or accelerators.

Enterprises want short time to real value, not more proof of concepts.

Problem #1 - Misconception

Leveraging large language models is all about prompt engineering, it's as easy as:

Your app sends a well-crafted prompt to an API.

ChatGPT model responds with a completion.

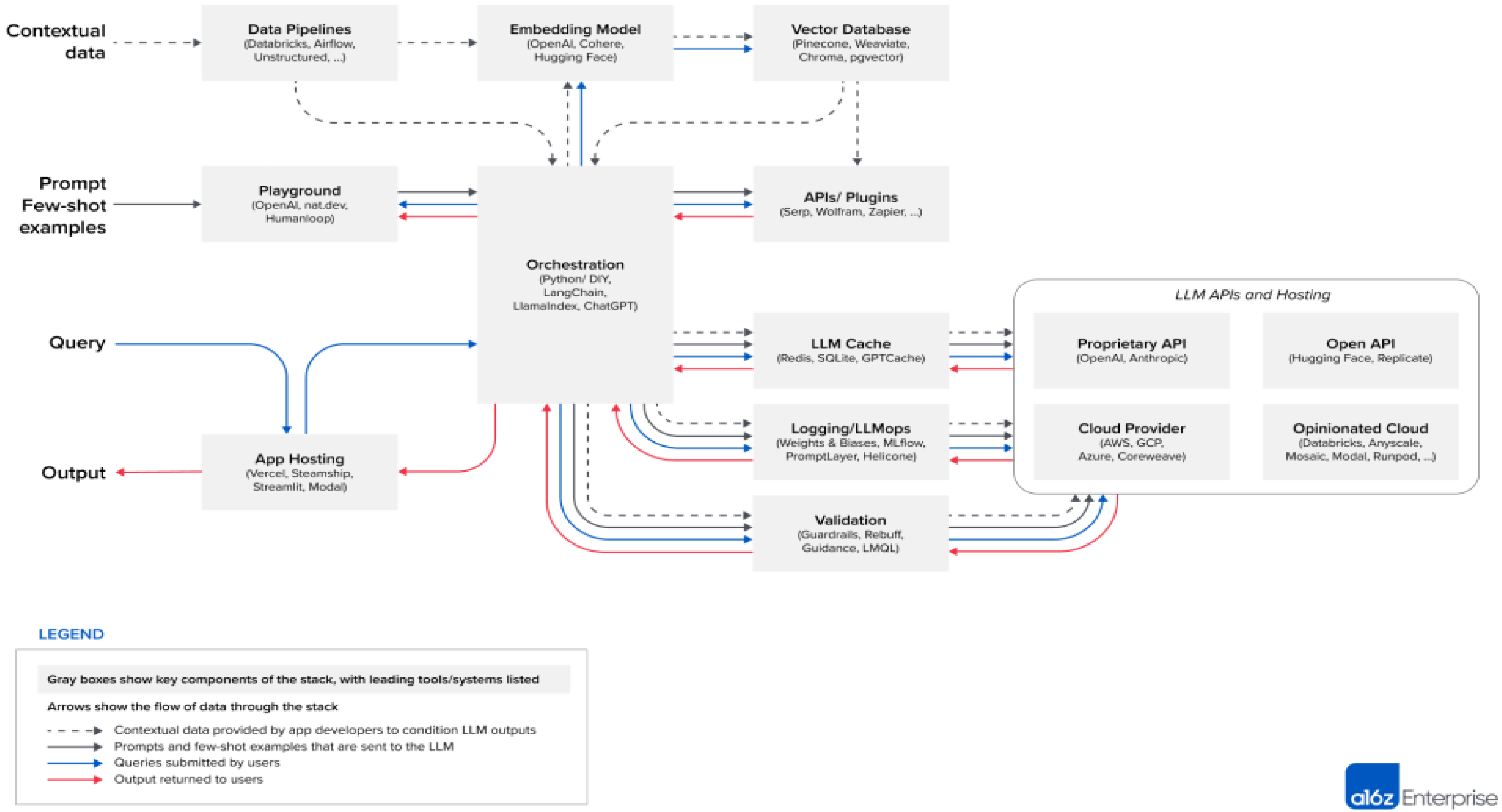
All done.

As an enterprise, how do I?

- provide a branded chat user experience atop this?
- manage conversation history on a per user basis?
- enable API integration into my automated workflows?
- use more advanced interaction patterns that are recursive, create code or use tools?
- scale and batch load thousands to millions of documents as knowledge for the model?
- provide all the data from my enterprise data estate?
- optimize the vectorization approach to best suit the content?
- measure and optimize completion quality?
- leverage other LLM's (Llama 2, Mistral) alongside ChatGPT models?
- leverage the LLM orchestrators (LangChain, Semantic Kernel, Prompt Flow) I want to use?
- self-host LLMs so my data doesn't leave my environment?
- govern token use down to the user or app level?
- keep sensitive knowledge sources private to authorized users?
- keep sensitive data (PII, trade secrets) from being sent to the LLM?
- manage having multiple AI agents?
- scale the solution to support 100k users or multiple data sovereignty regions?

Problem #2 - A complex reality

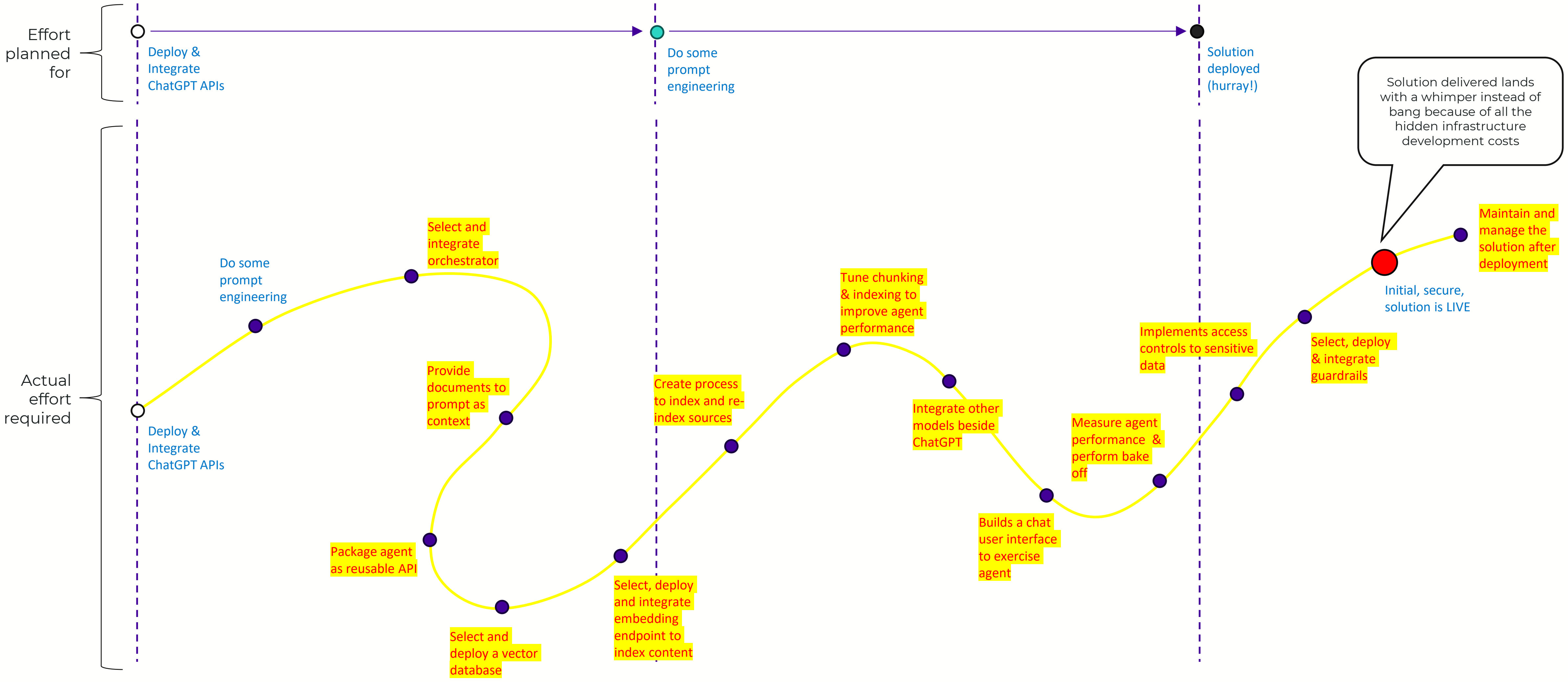
In reality, an enterprise grade LLM powered solution has a lot of moving parts (but don't just take our word for it):



a16z Enterprise

Source: https://a16z.com/2023/06/20/emerging-architectures-for-llm-applications

Problem #2 - It can be a long and frustrating customer



Problem #3 – High Price Tag

Me: "Starting from scratch, generative AI initiatives take about 6 months, and high six to low seven figures."

Big-4 : "Agree with everything you said, but actually, double those."

Generative AI endeavours...are costing from $5 million to $20 million in upfront investments.

Source: Gartner

	Consume Commercial GenAI apps	Embed GenAI APIs in custom apps	Extend GenAI models via data retrieval	Customize GenAI models via fine-tuning	Build Custom models from scratch
Use case	Coding assistants	Personalized sales content creation	Document search with RAG	Virtual assistant	Medical, insurance or financial services LLMs
Upfront costs	~$100K to $200k	~$750K to $1M	~$750K to $1M	~$5M to $6.5M	~$8M to $20M
Recurring costs (per user per year)	~$280 to $550	~$790 to $1.2K	~$1.3K to $11K	~$8K to $11K	~$11K to $21K

Source: Gartner

Problem #4 – AI Model Sprawl

Marketing
- OpenAI GPT-4o
- OpenAI DALLE-3
- Stable Diffusion

Finance
- OpenAI GPT-4o
- OpenAI GPT-o1

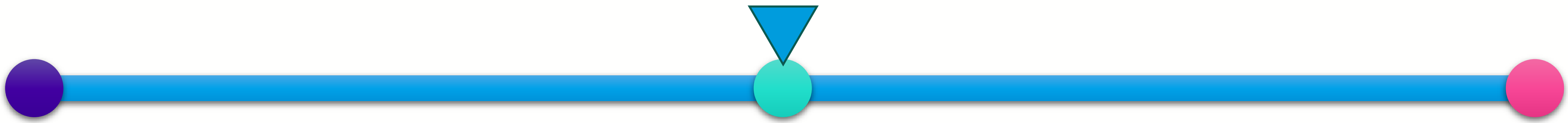
Legal
- Claude 3.5 Sonnet
- OpenAI GPT-4o

Operations
- Claude 3.5 Haiku
- OpenAI GPT 4o-mini

Developers
- Google Gemini 1.5
- Meta Llama 3
- Mistral
- OpenAI GPT-4o
- OpenAI GPT-o1

Strike a balance between self service and build from scratch

FoundationaLLM



Overview of LLM Platform Landscape

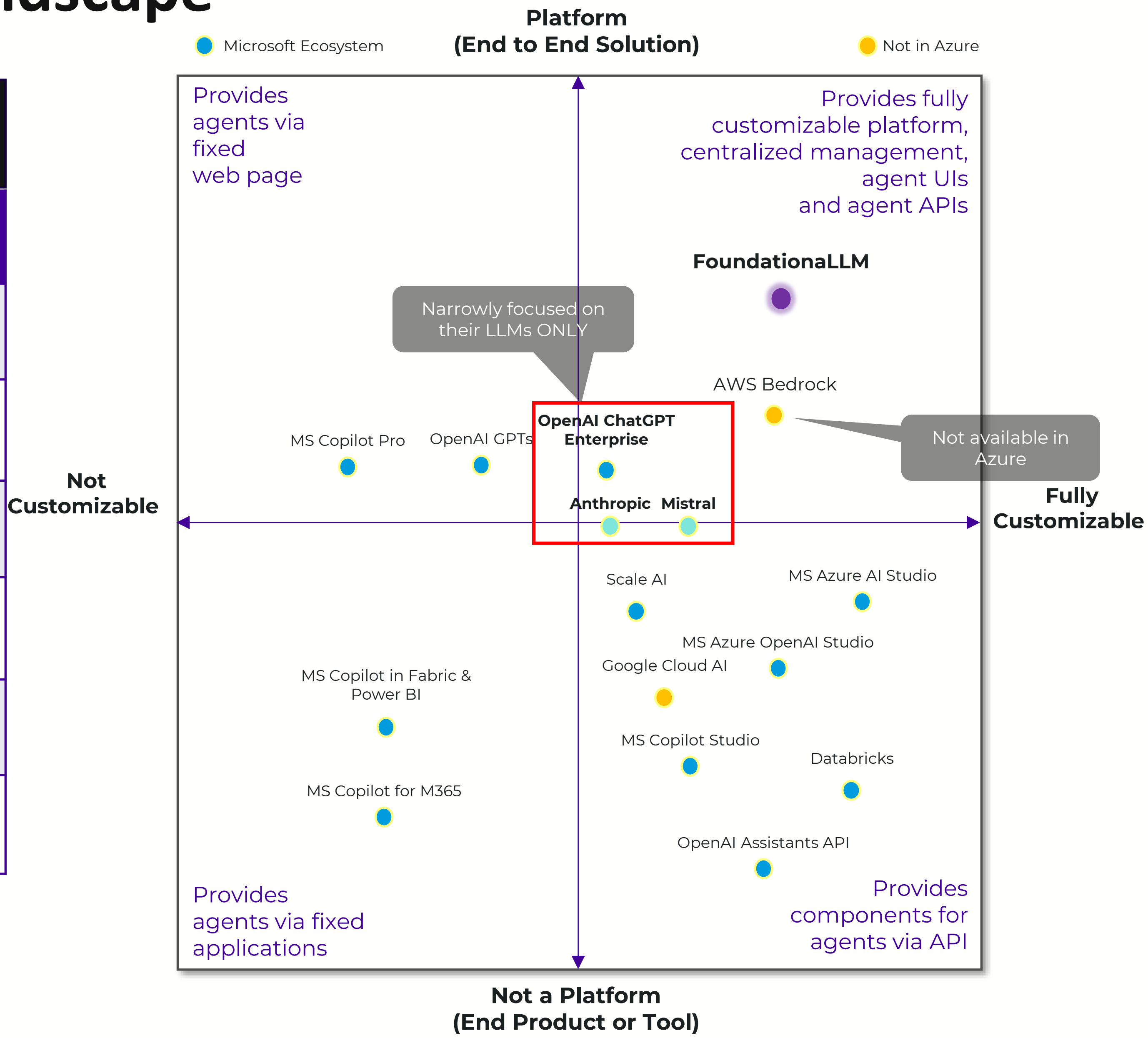
Create LLMs

Company	Product Description
Anthropic	Claude family of models
Cohere	Command family of models
Databricks	DBRX family of models
Google Cloud AI	Gemma & Gemini family of models
Open AI	GPT family of models
Meta	Llama family of models
Mistral	Mistral family of models

Not what we do.

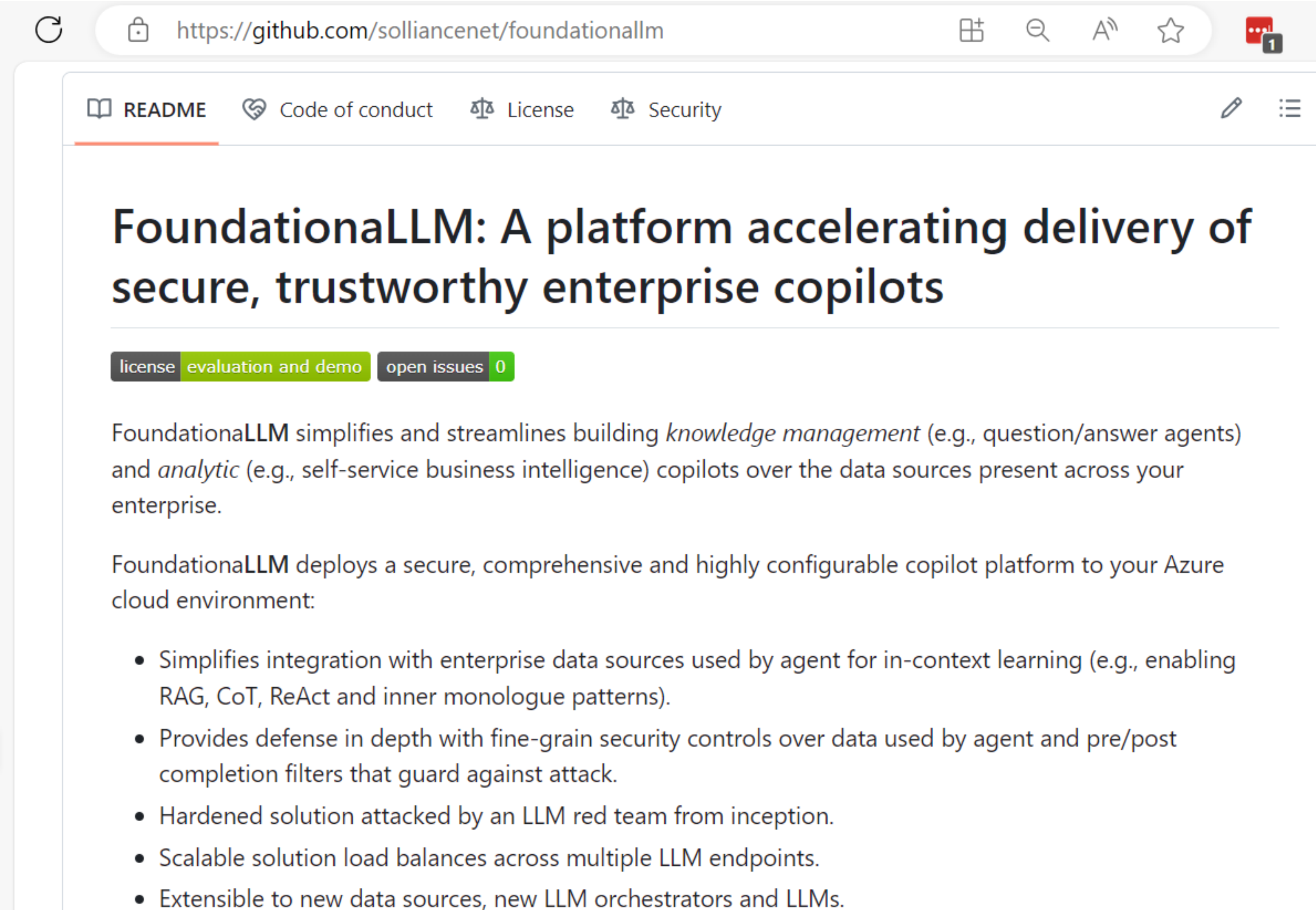
Create Software for Using LLMs

Company	Product Description
FoundationaLLM	FoundationaLLM
Microsoft	Azure AI Studio Azure OpenAI Studio Copilot Studio
AWS	AWS Bedrock
Databricks	Databricks Mosaic AI
Open AI	Enterprise ChatGPT
Scale AI	Scale GenAI Platform



FoundationaLLM source code is out in the open.

See it for yourself on GitHub.


https://github.com/solliancenet/foundationallm

README | Code of conduct | License | Security

FoundationaLLM: A platform accelerating delivery of secure, trustworthy enterprise copilots

license evaluation and demo | open issues 0

FoundationaLLM simplifies and streamlines building *knowledge management* (e.g., question/answer agents) and *analytic* (e.g., self-service business intelligence) copilots over the data sources present across your enterprise.

FoundationaLLM deploys a secure, comprehensive and highly configurable copilot platform to your Azure cloud environment:

- Simplifies integration with enterprise data sources used by agent for in-context learning (e.g., enabling RAG, CoT, ReAct and inner monologue patterns).
- Provides defense in depth with fine-grain security controls over data used by agent and pre/post completion filters that guard against attack.
- Hardened solution attacked by an LLM red team from inception.
- Scalable solution load balances across multiple LLM endpoints.
- Extensible to new data sources, new LLM orchestrators and LLMs.

FoundationaLLM provides the platform for deploying, scaling, securing and governing generative AI in the enterprise.

Enable customers to create these bespoke AI solutions:

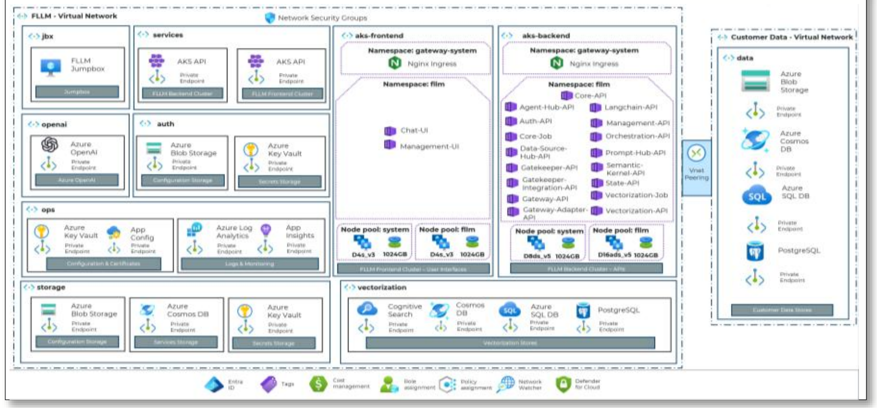
Chat Bots	Agents	Copilots
Converse with users	Execute complex tasks	Assist users within apps

FOUNDATIONALLM

APIs	Chat Portal	Management Portal
Easy AI agent integration	Full featured chat experience	Single pane of glass for AI

Platform

Hosted in Customer Tenant, LLM Model & Orchestrator Agnostic

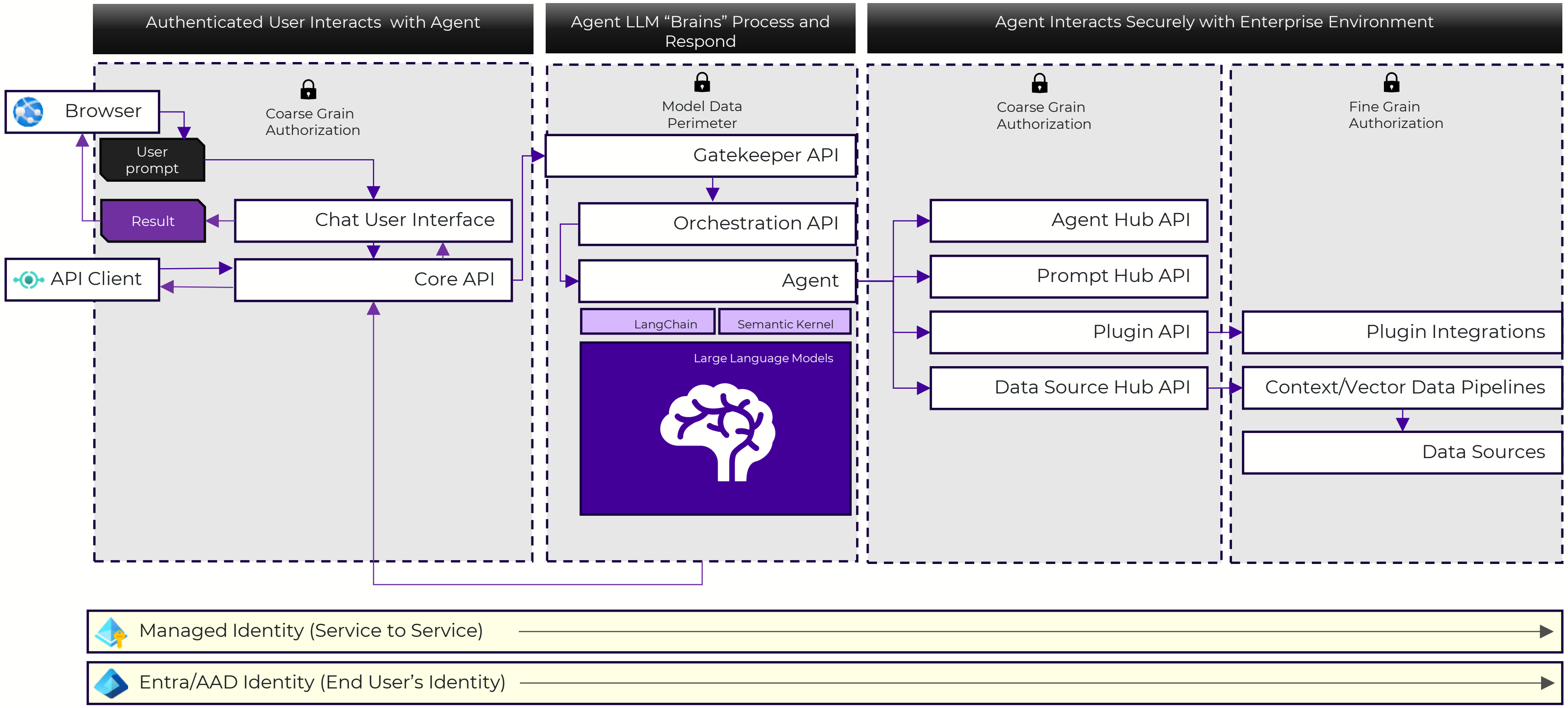


Differentiator #1 – Comprehensive, secure and flexible platform for chatbots, copilots and agentic solutions that deploys into customer Azure subscription.

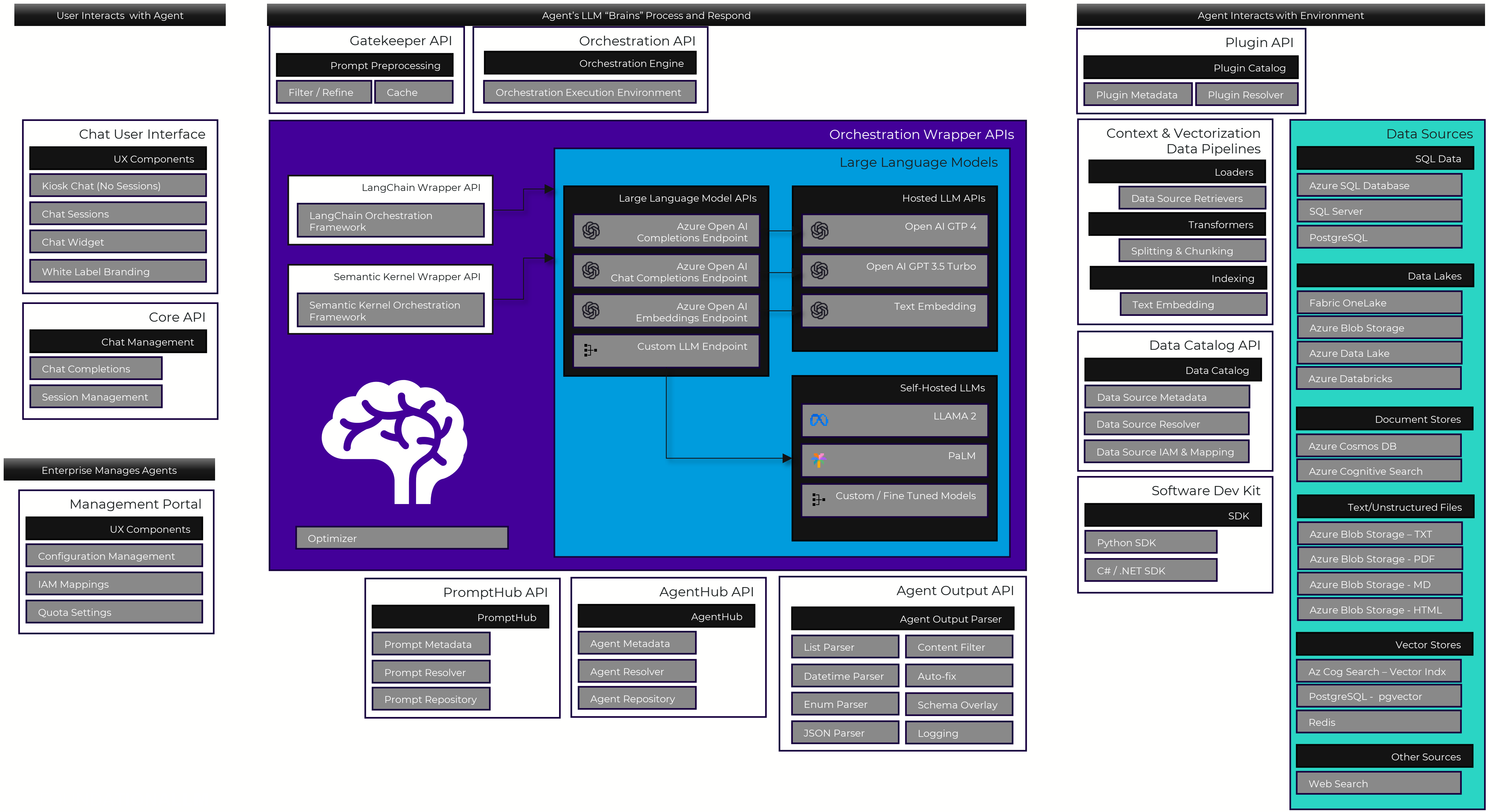
Differentiator #2 – API first platform designed for easy integration and provides a rich chat user interface ready to use out of the box.

Differentiator #3 – Secure. Designed from the ground up for handling high risk data.

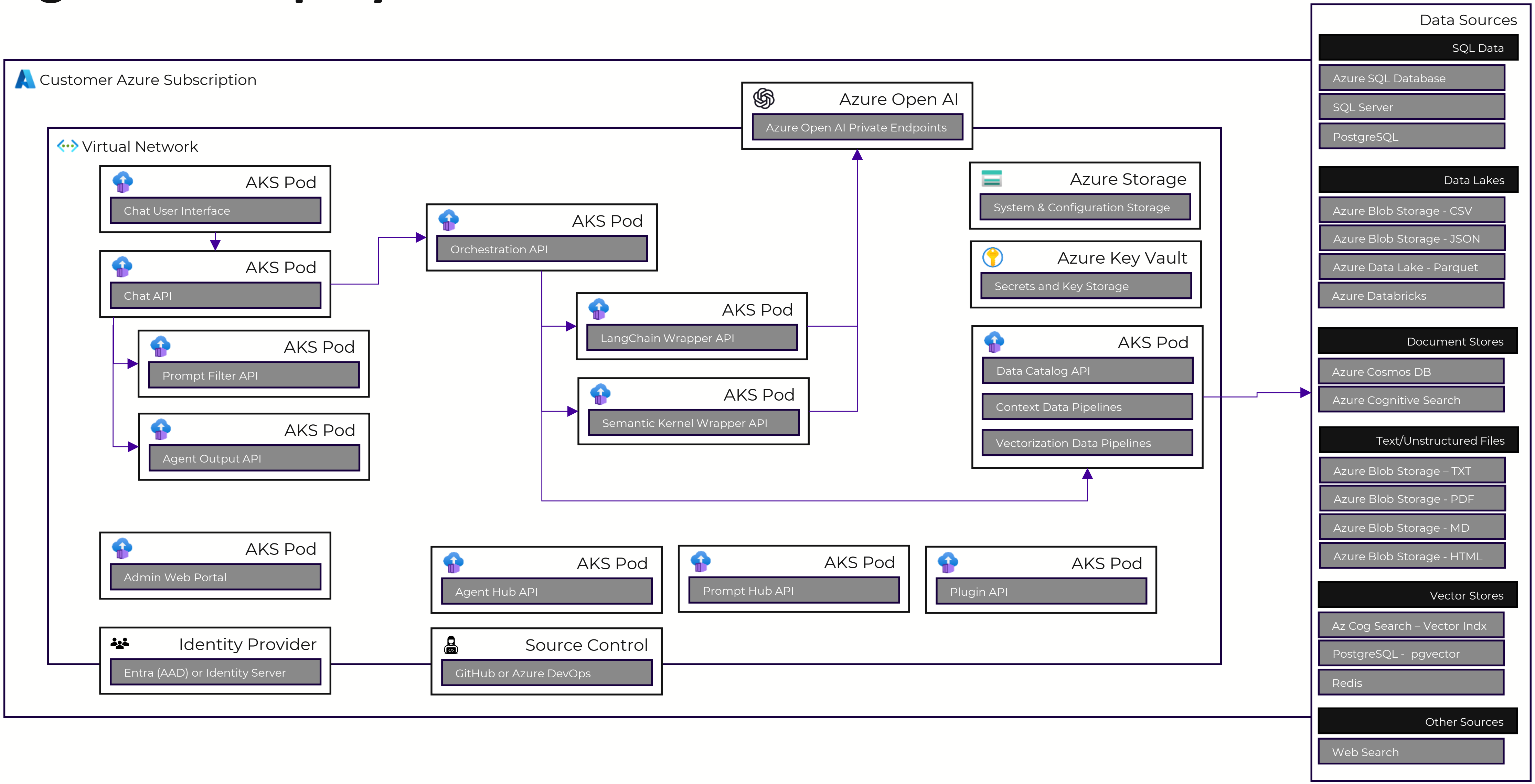
FoundationaLLM provides the platform for deploying, scaling, securing and governing generative AI in the enterprise.



FoundationaLLM enables trustworthy models grounded in enterprise data



High level deployment architecture

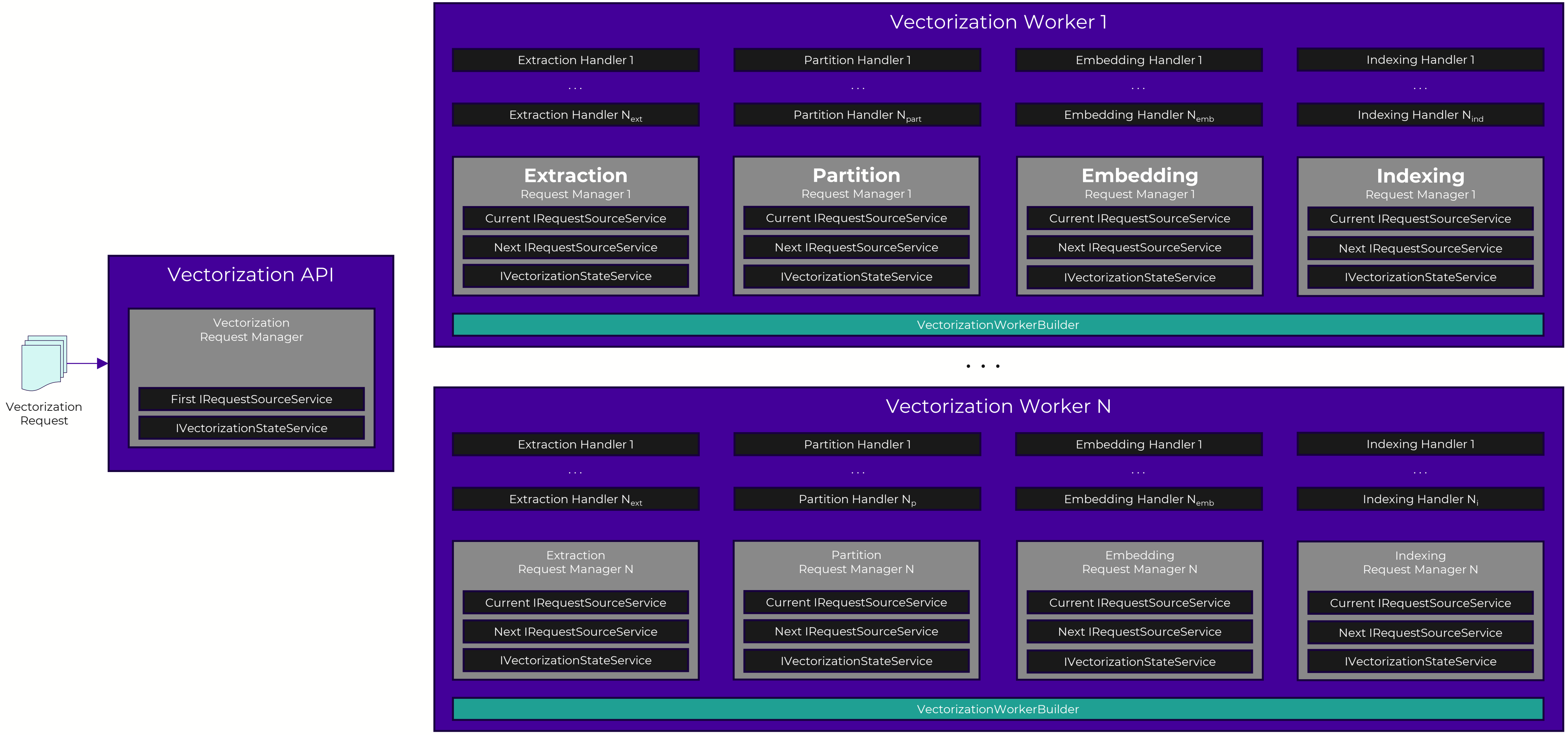


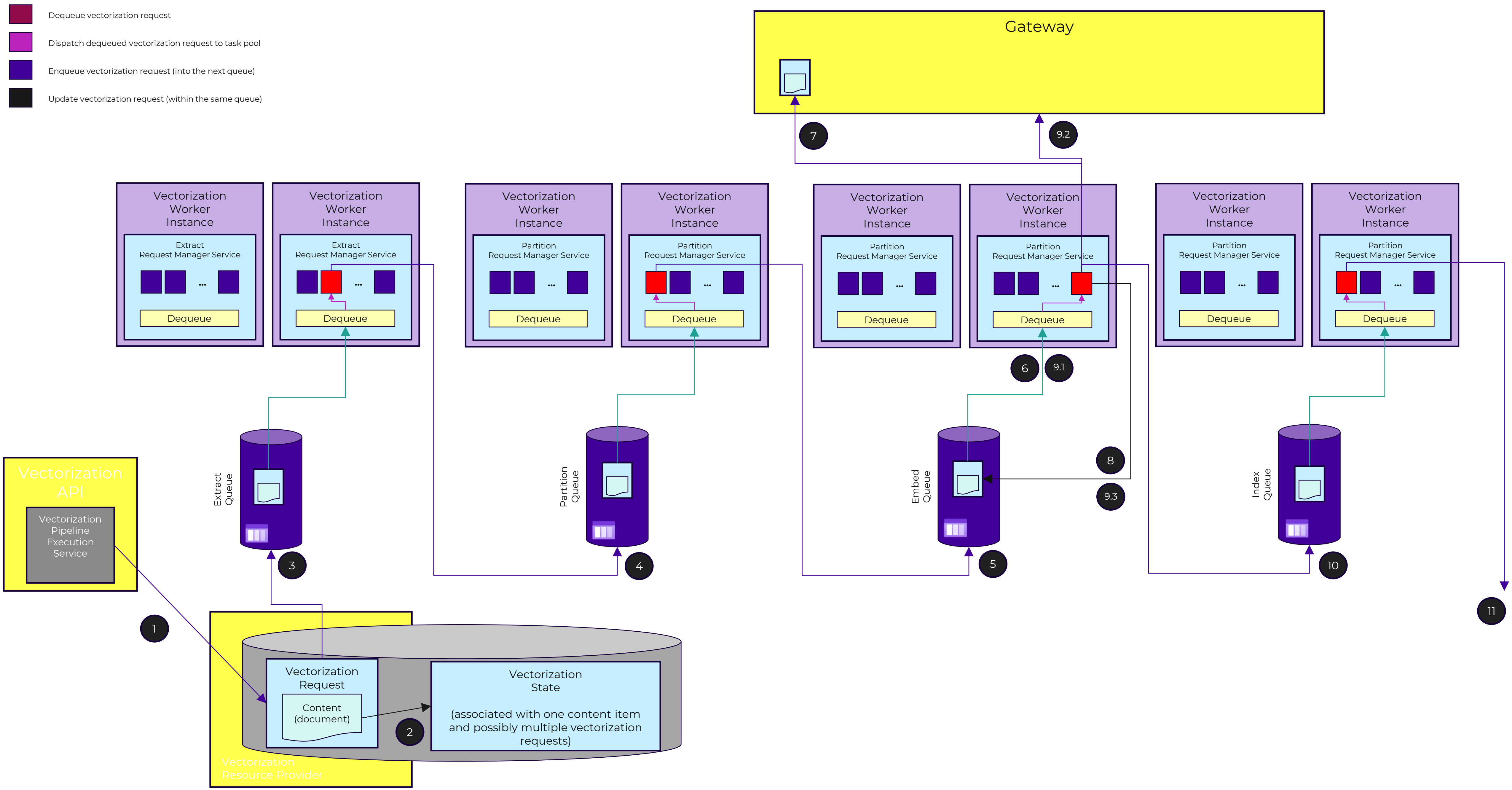
Vectorization

- Read next request from request source
- Invoke appropriate handler
- Create new request in the next request source
- Update the state of the vectorization pipeline

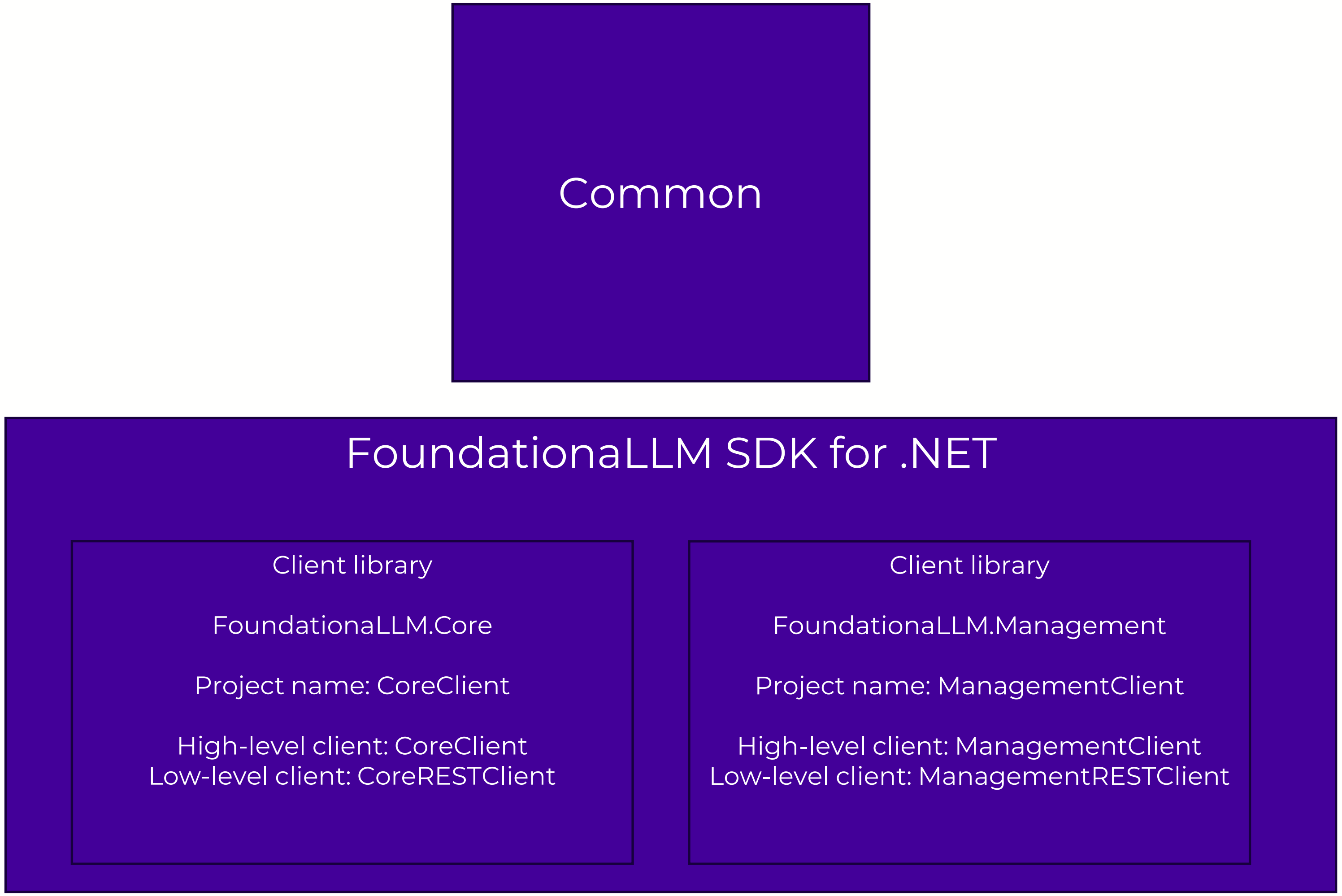
Creates all request manager instances based on the vectorization worker configuration

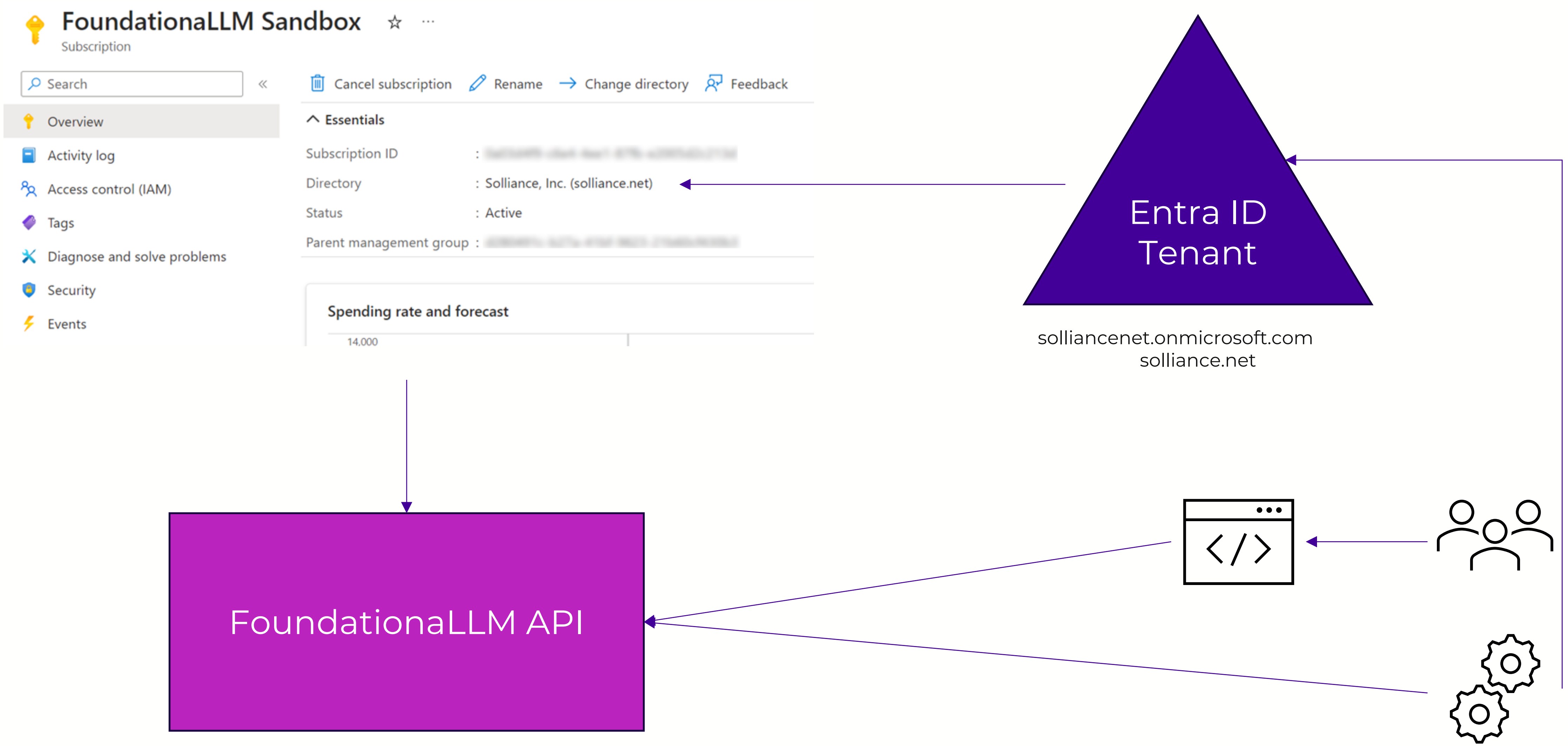
Vectorization

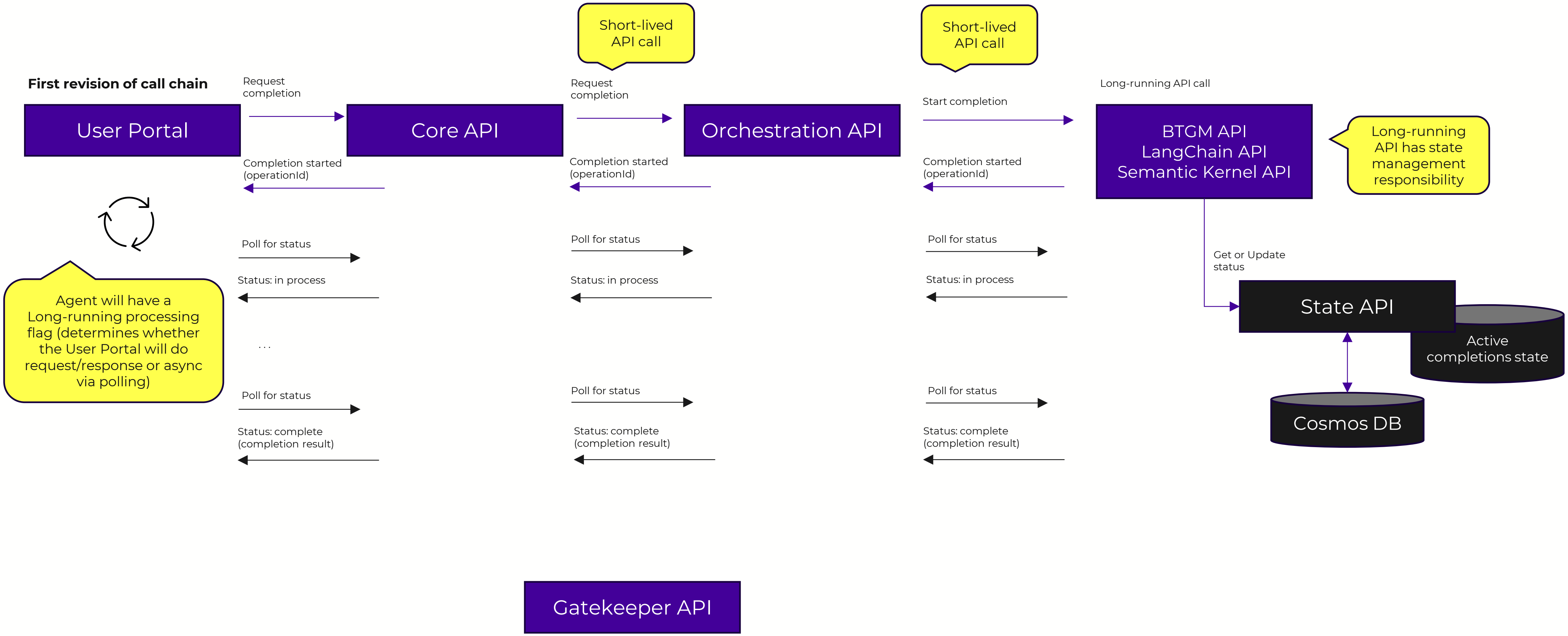




Dequeue vectorization request
Dispatch dequeued vectorization request to task pool
Enqueue vectorization request (into the next queue)
Update vectorization request (within the same queue)
Gateway
7
9.2
Vectorization Worker Instance
Extract Request Manager Service
...
Dequeue
Vectorization Worker Instance
Extract Request Manager Service
...
Dequeue
Vectorization Worker Instance
Partition Request Manager Service
...
Dequeue
Vectorization Worker Instance
Partition Request Manager Service
...
Dequeue
Vectorization Worker Instance
Partition Request Manager Service
...
Dequeue
Vectorization Worker Instance
Partition Request Manager Service
...
Dequeue
6
9.1
Vectorization Worker Instance
Partition Request Manager Service
...
Dequeue
Vectorization Worker Instance
Partition Request Manager Service
...
Dequeue
Extract Queue
Partition Queue
Embed Queue
Index Queue
8
9.3
3
4
5
10
11
Vectorization API
Vectorization Pipeline Execution Service
1
Vectorization Request
Content (document)
2
Vectorization State
(associated with one content item and possibly multiple vectorization requests)
Vectorization Resource Provider


Common
FoundationalLLM SDK for .NET
Client library
FoundationalLLM.Core
Project name: CoreClient
High-level client: CoreClient
Low-level client: CoreRESTClient
Client library
FoundationalLLM.Management
Project name: ManagementClient
High-level client: ManagementClient
Low-level client: ManagementRESTClient


FoundationalLLM Sandbox
Subscription
Search
Cancel subscription
Rename
Change directory
Feedback
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Security
Events
Essentials
Subscription ID :
Directory : Solliance, Inc. (solliance.net)
Status : Active
Parent management group :
Spending rate and forecast
14,000
Entra ID Tenant
sollIancenet.onmicrosoft.com
solliance.net
FoundationalLLM API



First revision of call chain
User Portal
Request
completion
Core API
Short-lived
API call
Request
completion
Orchestration API
Short-lived
API call
Start completion
Long-running API call
BTGM API
LangChain API
Semantic Kernel API
Long-running
API has state
management
responsibility
Completion started
(operationId)
Completion started
(operationId)
Completion started
(operationId)
Poll for status
Status: in process
Poll for status
Status: in process
Poll for status
Status: in process
Get or Update
status
State API
Active
completions state
Cosmos DB
Agent will have a
Long-running processing
flag (determines whether
the User Portal will do
request/response or async
via polling)
. . .
Poll for status
Status: complete
(completion result)
Poll for status
Status: complete
(completion result)
Poll for status
Status: complete
(completion result)
Gatekeeper API

Role Based Access Control (RBAC) and Policy Based Access Control (PBAC)

RBAC role definitions:

Reader
Contributor
User Access Administrator
Owner

Attachments Contributor
Conversations Contributor
Role Based Access Control Administrator

RBAC role assignments:

Instance
Resource

PBAC policy definitions:

User Principal Name (UPN) match

PBAC policy assignments:

Resource Type

All role definitions: https://github.com/solliancenet/foundationallm/blob/main/src/dotnet/Common/Constants/Data/RoleDefinitions.json
All authorizable actions: https://github.com/solliancenet/foundationallm/blob/main/src/dotnet/Common/Constants/Data/AuthorizableActions.json
All policy definitions: https://github.com/solliancenet/foundationallm/blob/main/src/dotnet/Common/Constants/Data/PolicyDefinitions.json

Role Based Access Control (RBAC) and Policy Based Access Control (PBAC)

Agent security requirements:

Reader role assignment to Entra ID Security Group associated to agent

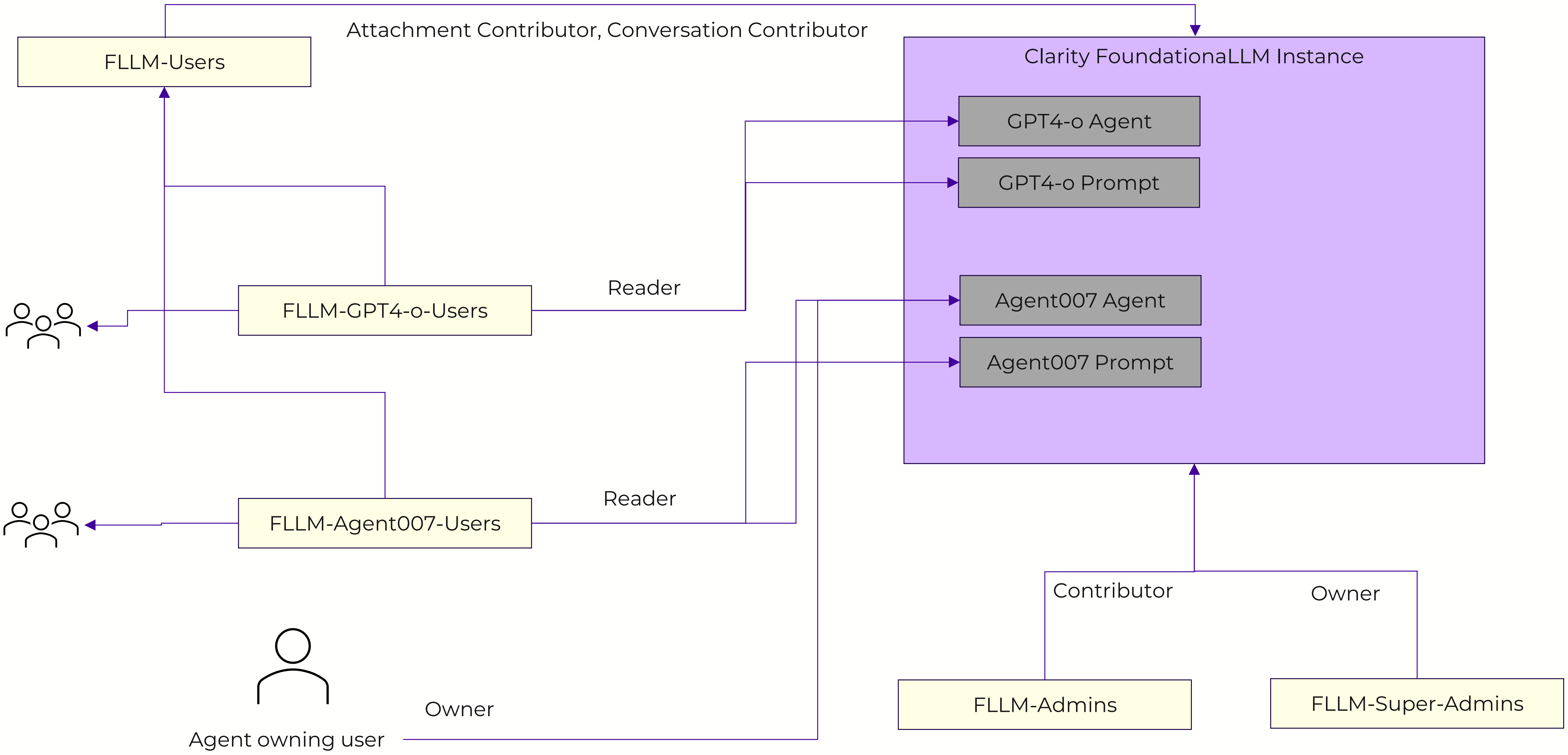
- How does the Security Group get created?
- How is membership for the Security Group managed?
- Does the Management Portal play any role in the permissions management process?
- Do we allow multiple owners on an agent?

User security requirements:

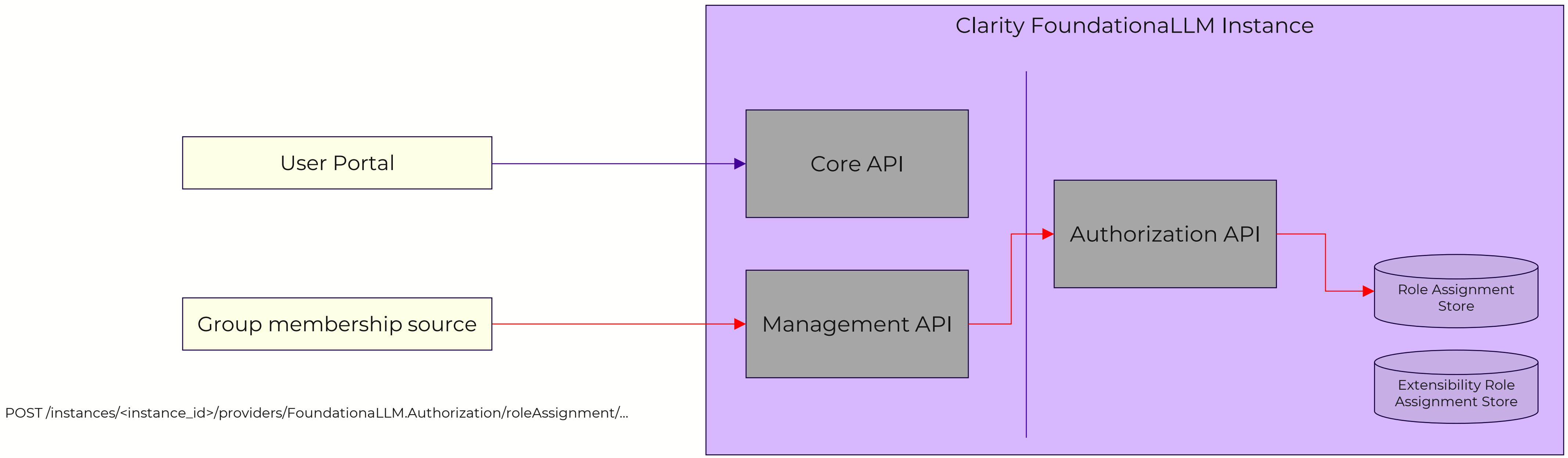
Reader role assignment to the user scoped at the agent
Conversation Contributor assignment to the user scoped at the FoundationaLLM instance level
Attachment Contributor assignment to the user scoped at the FoundationaLLM instance level

Reader role assignment to two specific configuration items
(CompletionResponsePollingIntervalSeconds, MaxUploadsPerMessage)

Role Based Access Control (RBAC) and Policy Based Access Control (PBAC)



Security groups membership integration

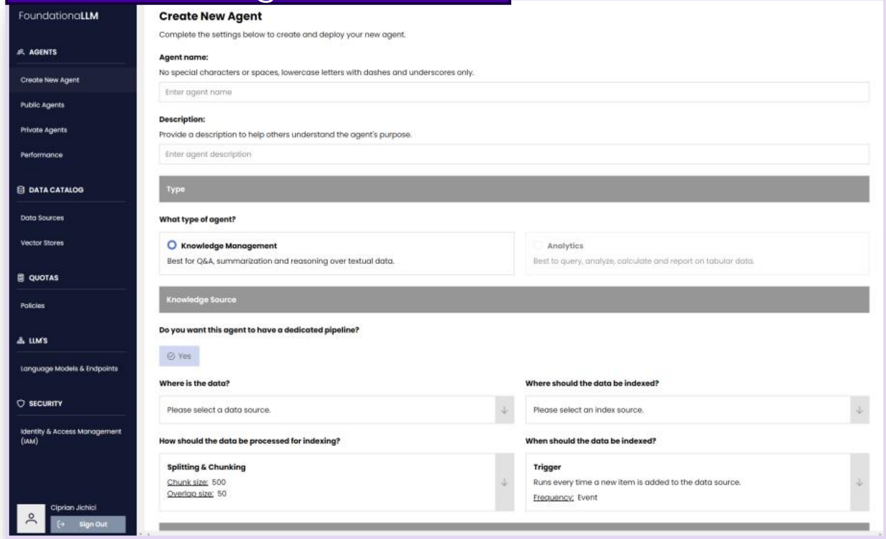


Management Portal

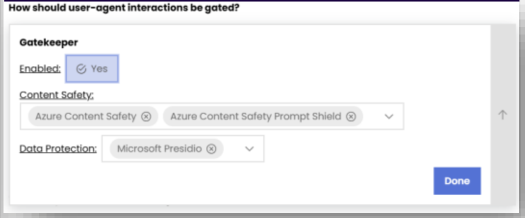
Manage all AI agents using a single pane of glass

- Self-service deployment of AI agents
- Single place to view and configure all agents
- Configure who has access to the agents with role based security
- Choose from a wide array of content safety and guardrails solutions
- Create complex agentic solutions that use enterprise data
- Monitor and manage costs

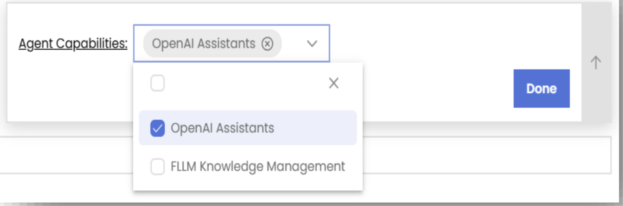
Self-Service Agent Creation



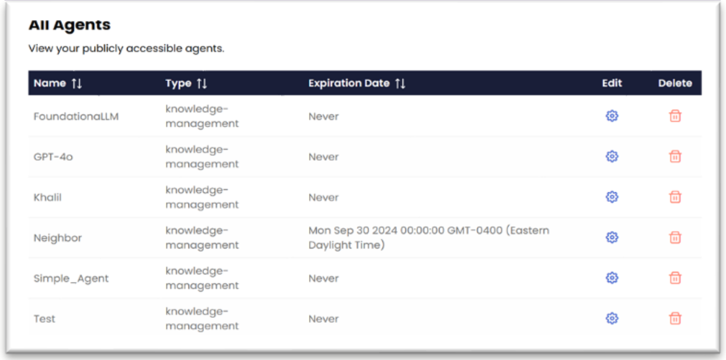
Configure Content Safety & Data Protection



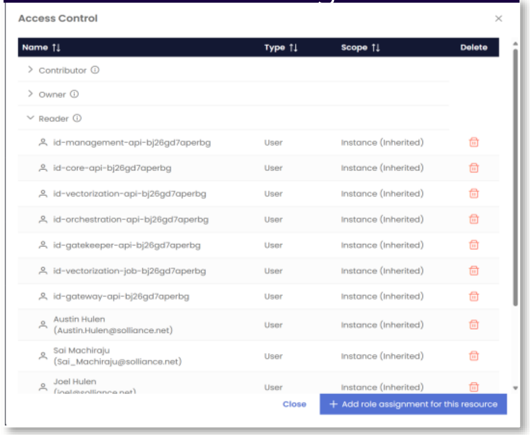
Easily create agents using OpenAI Assistants API



Manage All Deployed Agents



Configure Role Based Access Control on Agents

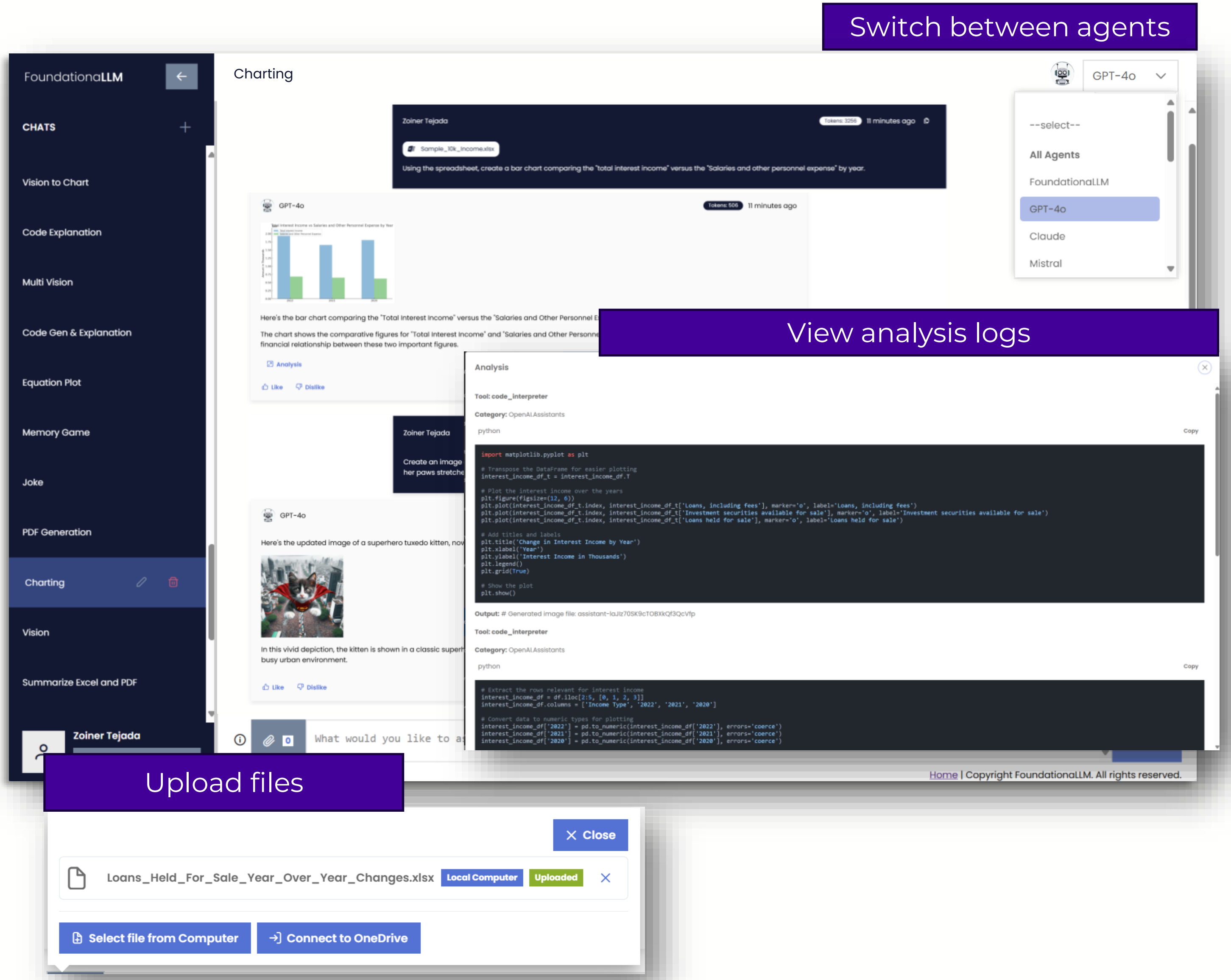


Chat User Interface

Provides advanced interactions across the range of Large Language Models & Orchestrators

- Feature rich, accessible and responsive chat application that works out of the box
- Single sign-on using enterprise credentials
- Choose from any agent to which user has access
- Maintains history of chat sessions per user

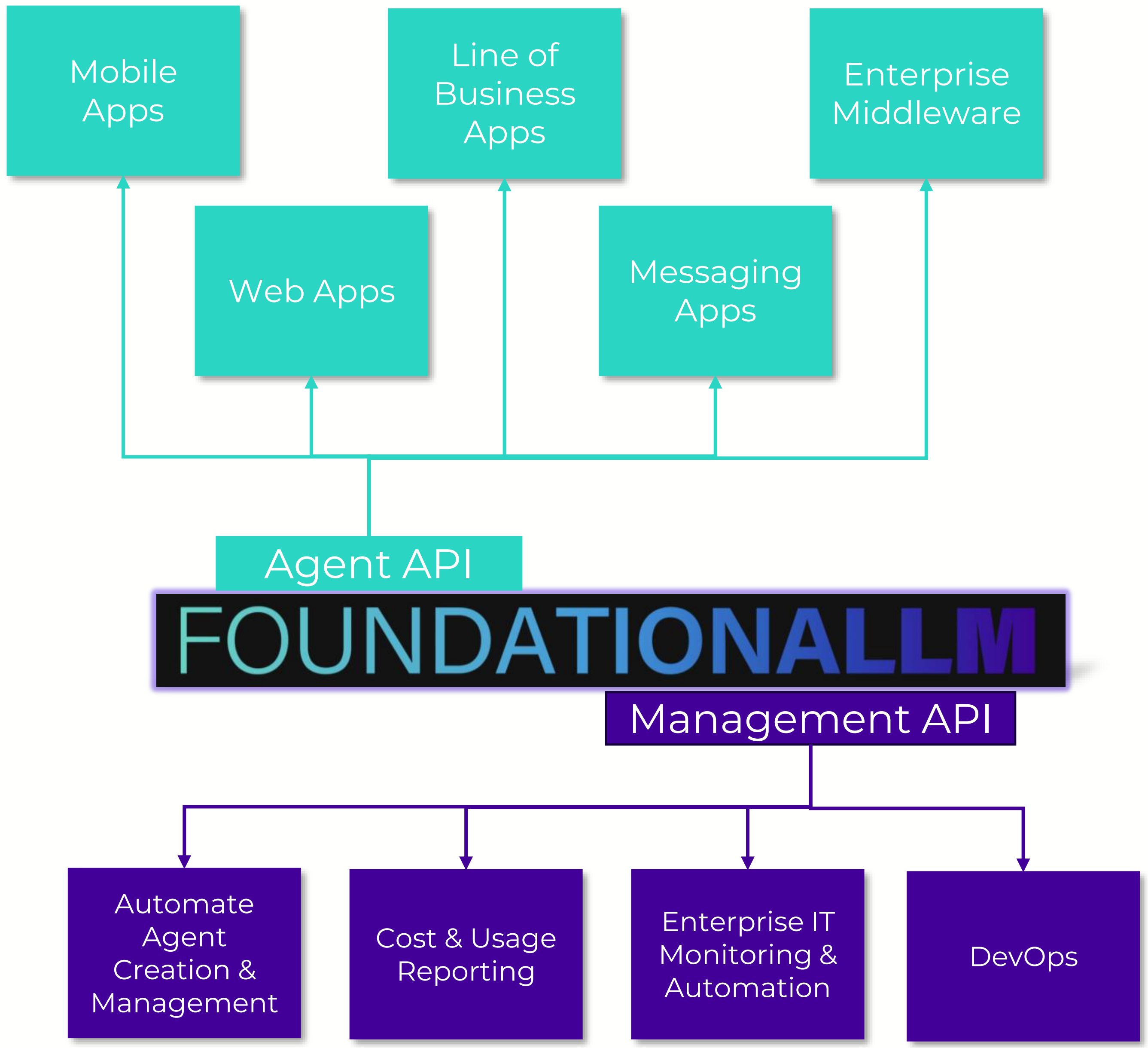
FoundationaLLM Features	
Branding	**Multiple Agents**
Configuration	**Agent to agent (@mentions)**
Accessibility	**Citations**
Chat History	**Charts & Tables (static & interactive)**
Sharing	**Inline images**
Upload local file	**Inline HTML**
Upload from OneDrive	**Code Syntax Highlighting**
Prompt Inspection	**LaTeX Equation Rendering**
Token Consumption	**Markdown formatting**
Assistants API Support (Analysis)	**File Download**
Production Use	**Responsive (Mobile Friendly)**



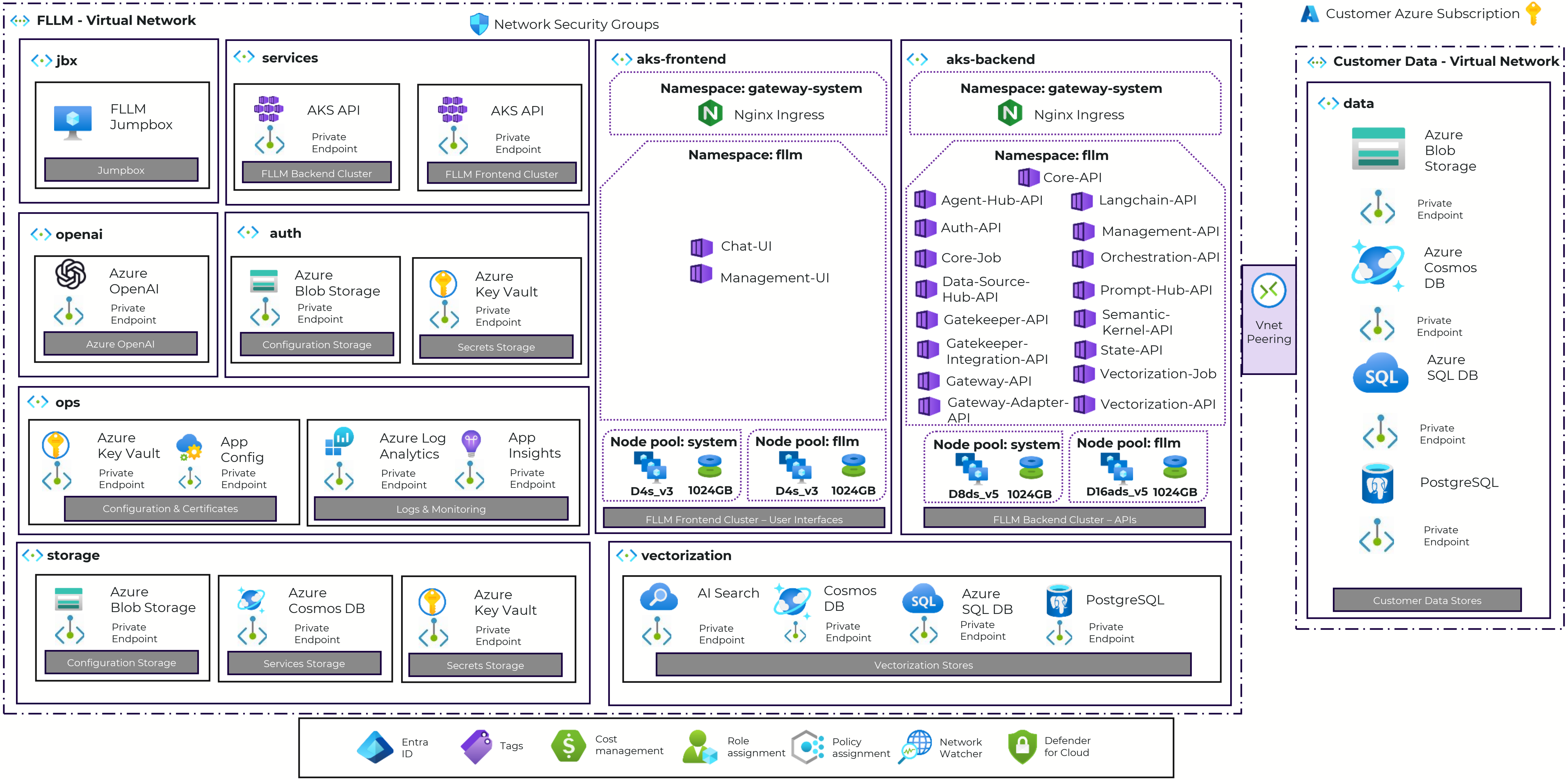
API First

Enable AI agent integration across the enterprise and allow platform management with existing IT systems.

- Agent API enables chat interactions to be embedded within any web page, application or middleware
- Management API enables management of the platform from external applications and IT systems
- Vectorization API for at scale indexing in RAG scenarios supporting front-end and back-end knowledge sources



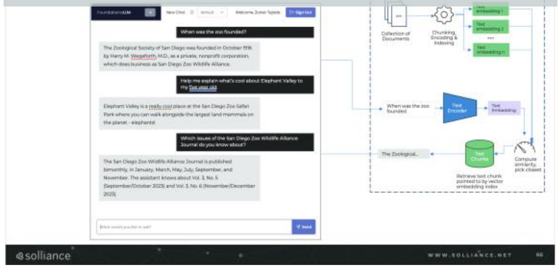
FLLM Standard Deployment Components



Powering top uses cases

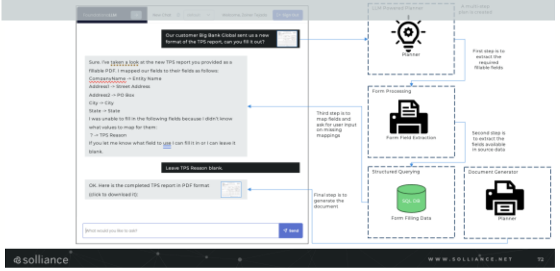
See all use cases:
https://foundationallm.ai/use-cases

Q&A from Documents & PDFs

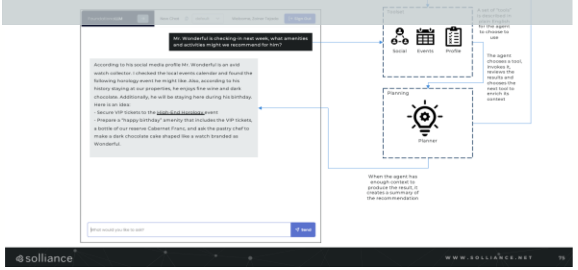


Anomaly Detection for Structured Data

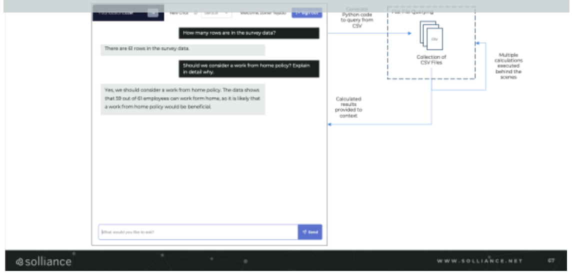
Generative form Filling & Document Creation



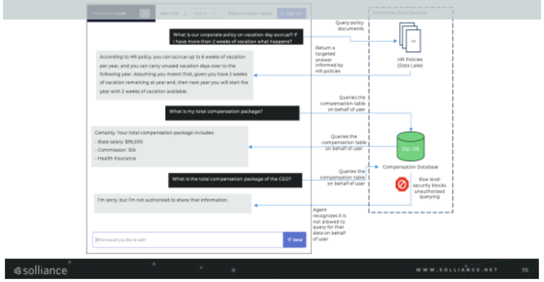
Dynamic Tool Use



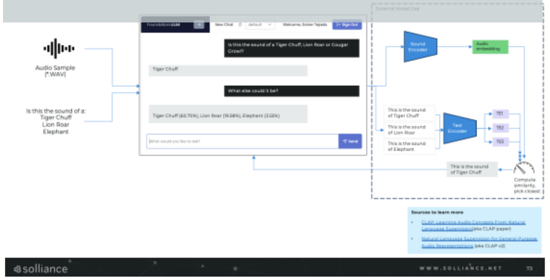
Recommendations from Flat File Data



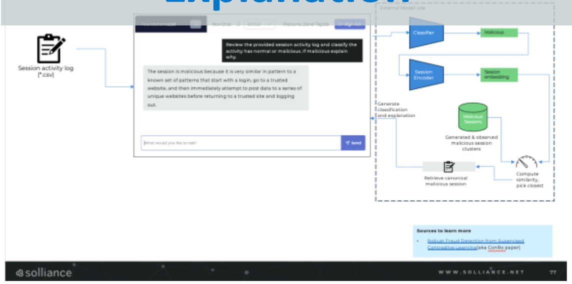
Security Aware Agents



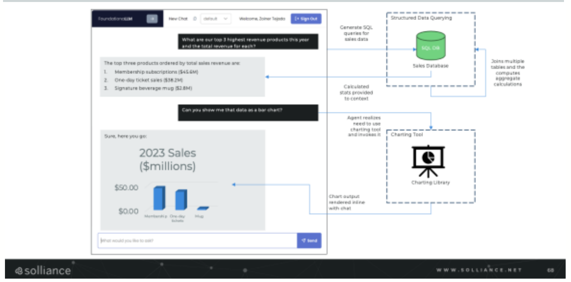
Zero Shot Audio Classification



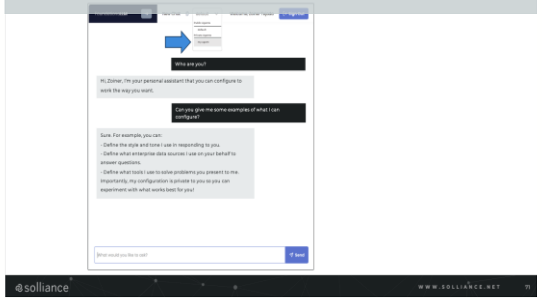
Fraud Detection and Explanation



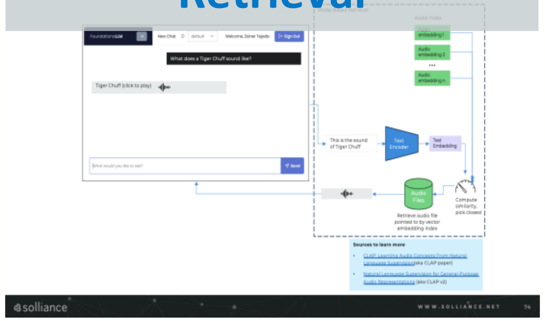
Self-service Analytics



Private Agents

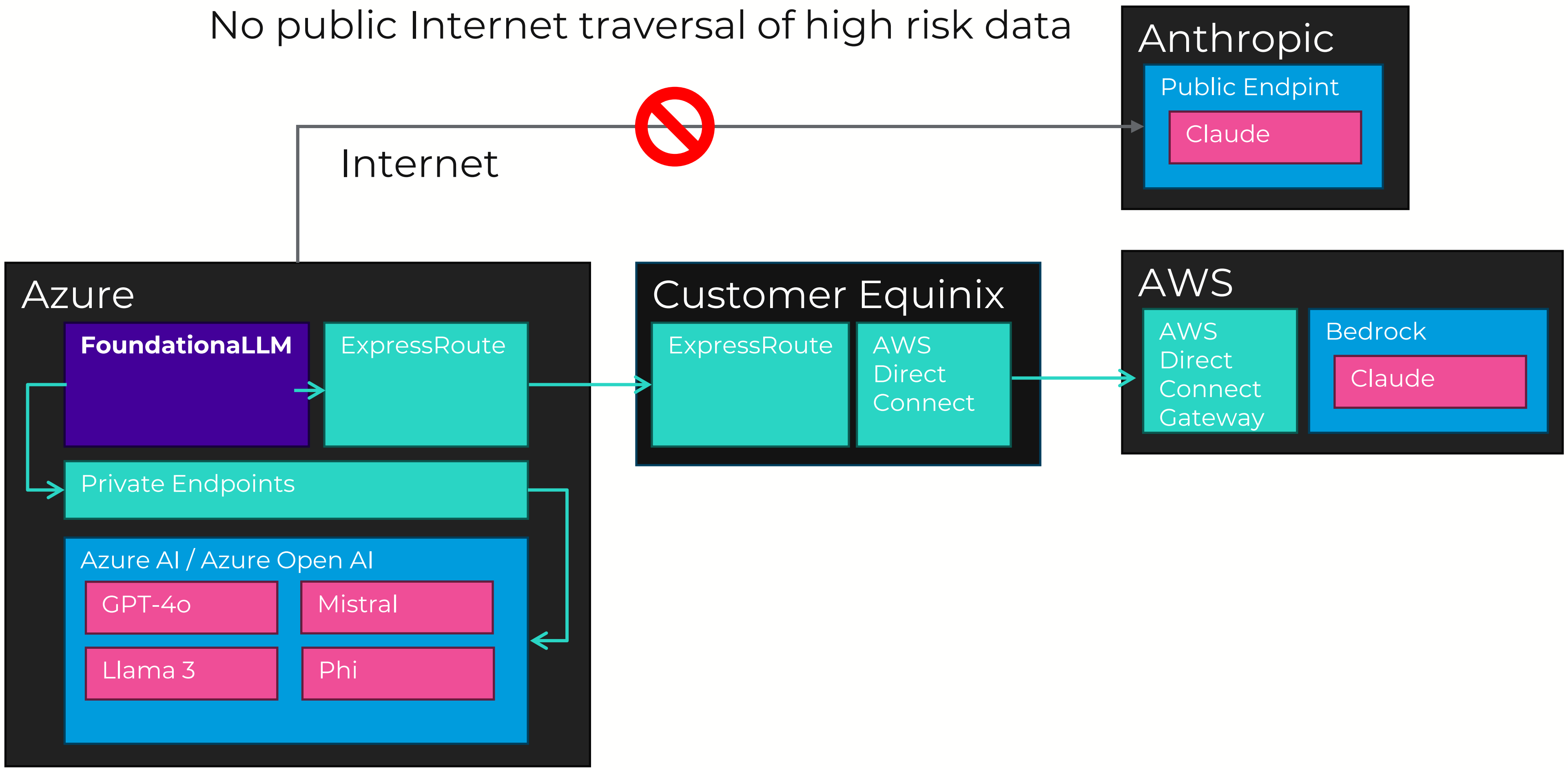


Zero Shot Audio Retrieval



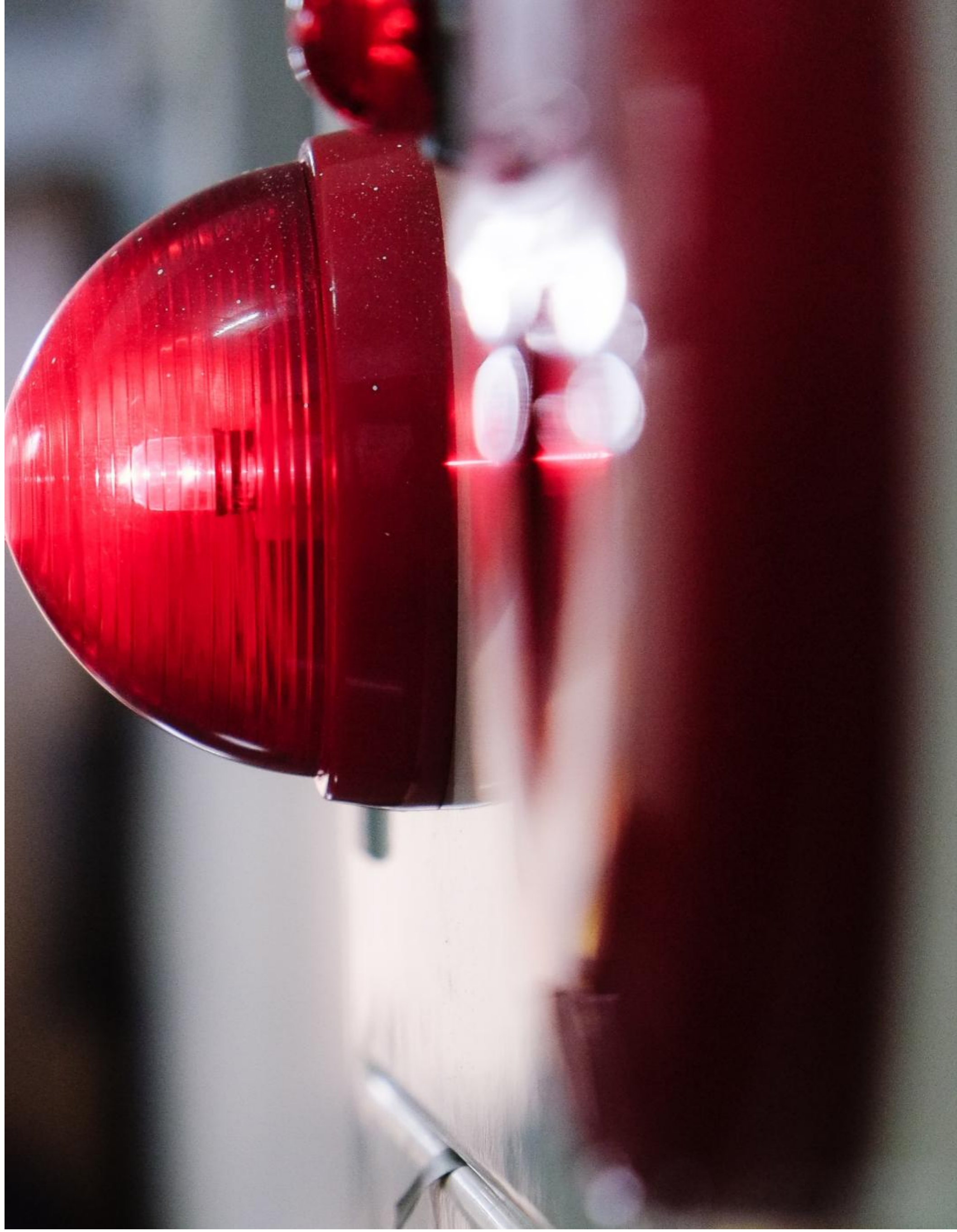
Speech Powered Agents

We deploy for maximum network security

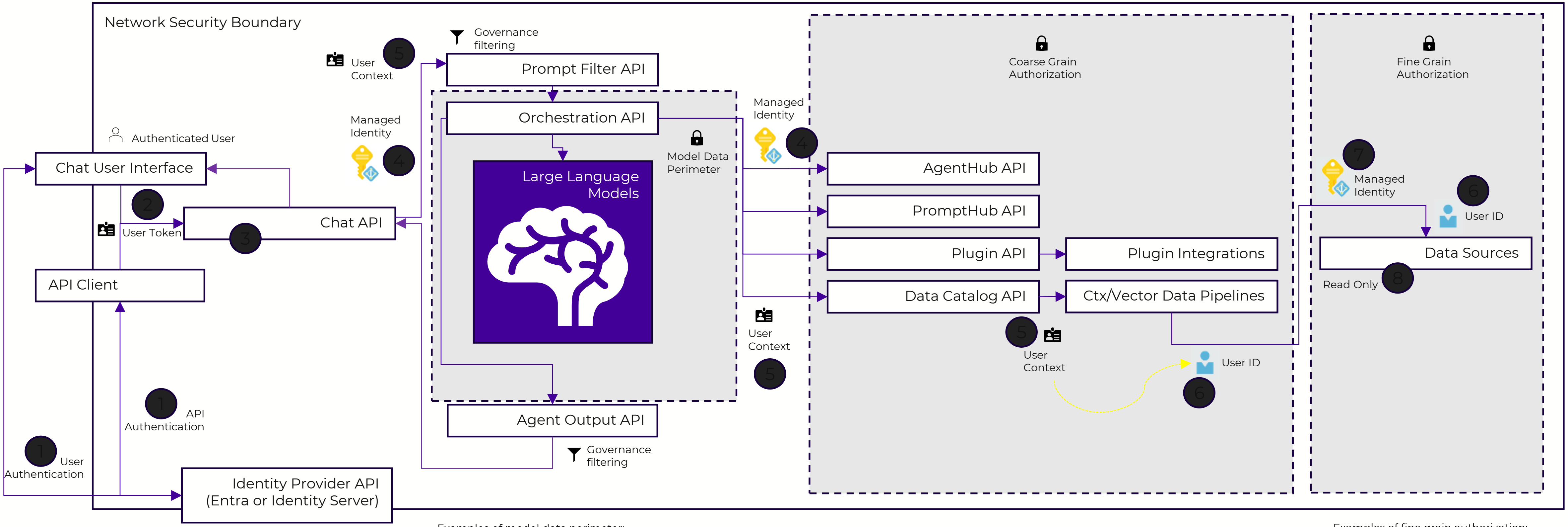


FoundationaLLM Helps to Mitigate These “Top 5” LLM Attacks

- ✓ Prompt Injection
- ✓ Excessive Agency
- ✓ Model Denial of Service
- ✓ Sensitive Information Disclosure
- ✓ Insecure Output Handling



The Foundation to a Secure and Compliant AI Solution



Examples of model data perimeter:

- Preventing exfiltration of PHI or sensitive data
- Disallowing calls to external models
- Blocking prompt injection attacks
- Model communication constrained to VNET and private endpoints

Examples of coarse grain authorization:

- Data source not listed in catalog
- Plugin not visible to orchestration
- Orchestration unable to retrieve connection string to database

Examples of fine grain authorization:

- Row level security
- Column level security
- File/folder level security

1 Users are authenticated using standard protocols (OpenID Connect, OAuth2) with their Identity Provider of choice. Supports single sign-on.

2 APIs authorize access based on authenticated user tokens (JWT's).

3 Chat API performs token validation.

4 Managed Identities identify each calling tier or component. No public API access beyond the UI layer APIs.

5 Each API tier requires a user context built upon the user token provided, enriched with attributes like token quotas, licenses and role membership.

6 Data sources identify users for fine grained access control. When needed, data source identities are mapped from user context.

7 As appropriate to the data source, a Managed Identity or other system level identity identifies calling tier or component to connect to the data source.

8 Data sources enforce read only access to prevent agents performing updates by accident for LLM generated queries.

Quantitative Measures

We can evaluate agent all-up performance along these dimensions

Does the answer read naturally?

Fluency
Measure the language proficiency of a generative AI's predicted answer.

Coherence Recommended
Measures how well the language model can produce output that flows smoothly, reads naturally, and resembles human-like language.

Did it answer the question?

Relevance Recommended
Measures the extent to which the model's generated responses are pertinent and directly related to the given questions.

Did it use the knowledge provided in the answer?

Groundedness Recommended
Measures how well the model's generated answers align with information from the input source.

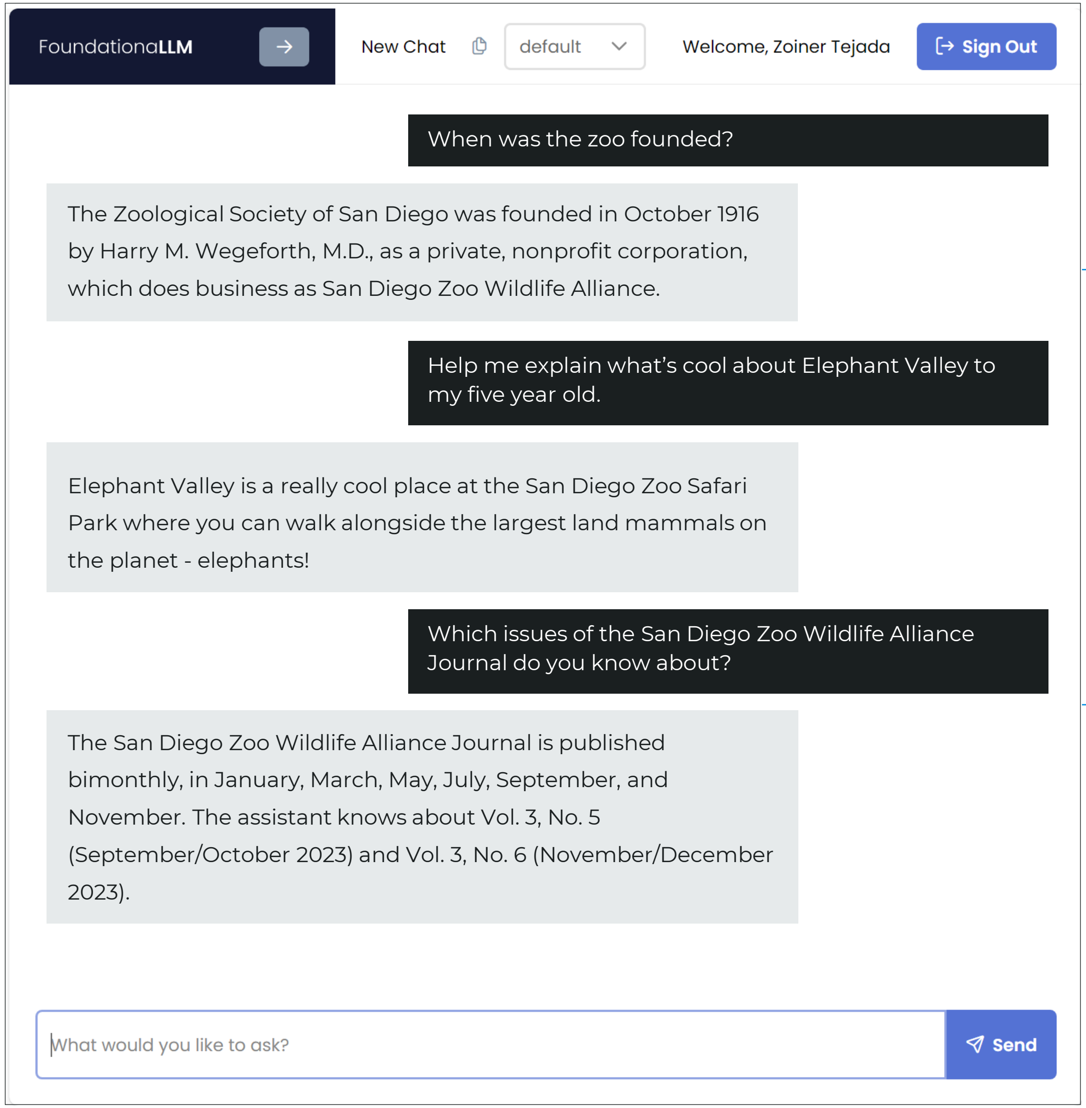
How good was the answer compared to a supplied best answer?

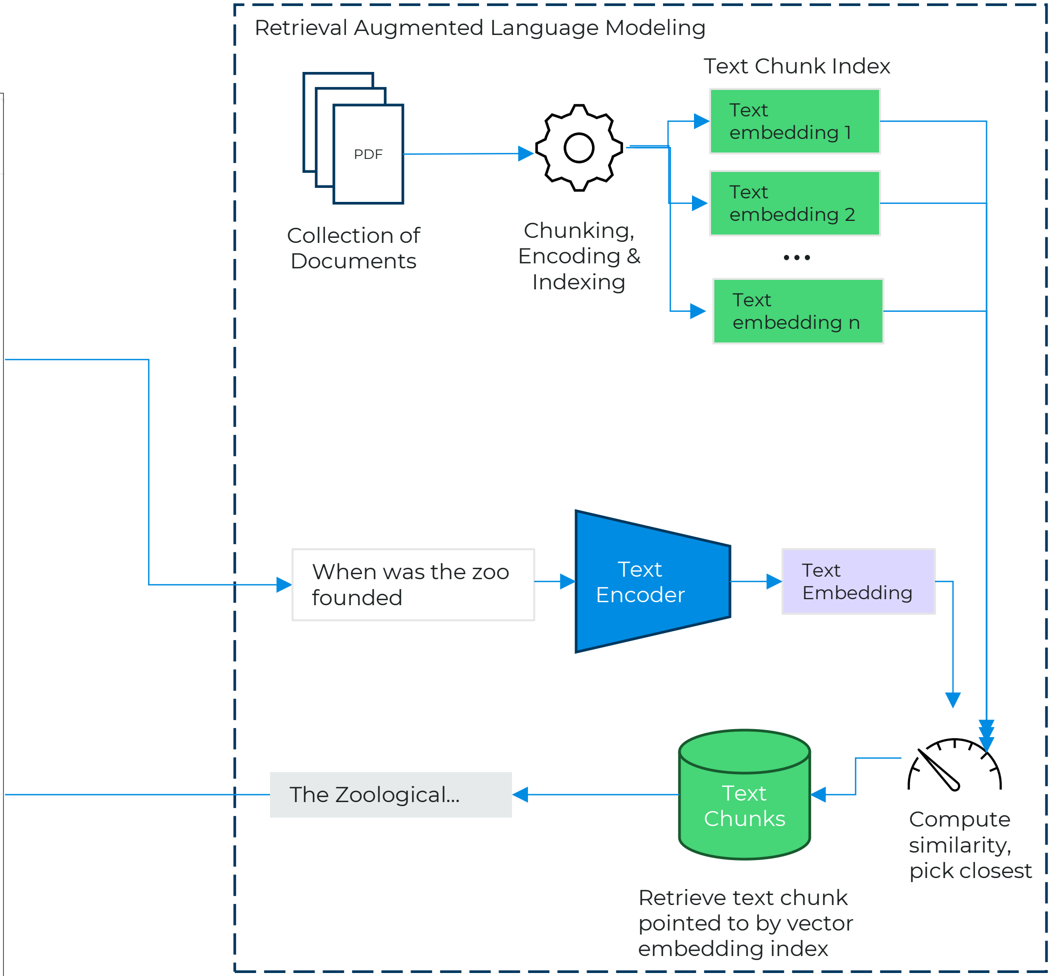
GPT similarity
Measures the similarity between a source data (ground truth) sentence and the generated response by a GPT-based AI model.

F1 score
Measures the ratio of the number of shared words between the model prediction and the source data (ground truth).

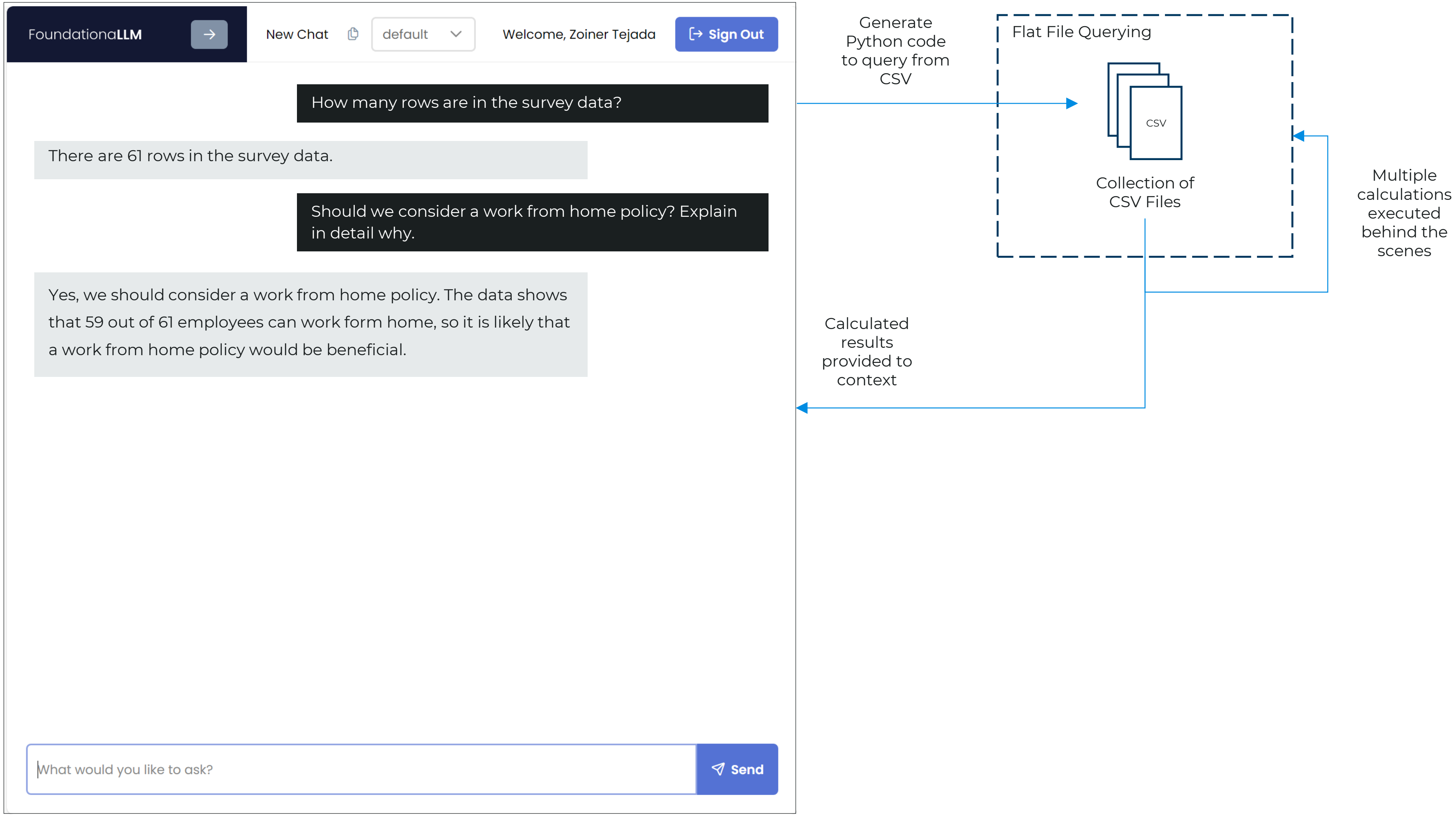
FoundationaLLM
Use Cases

Q&A from Documents & PDFs

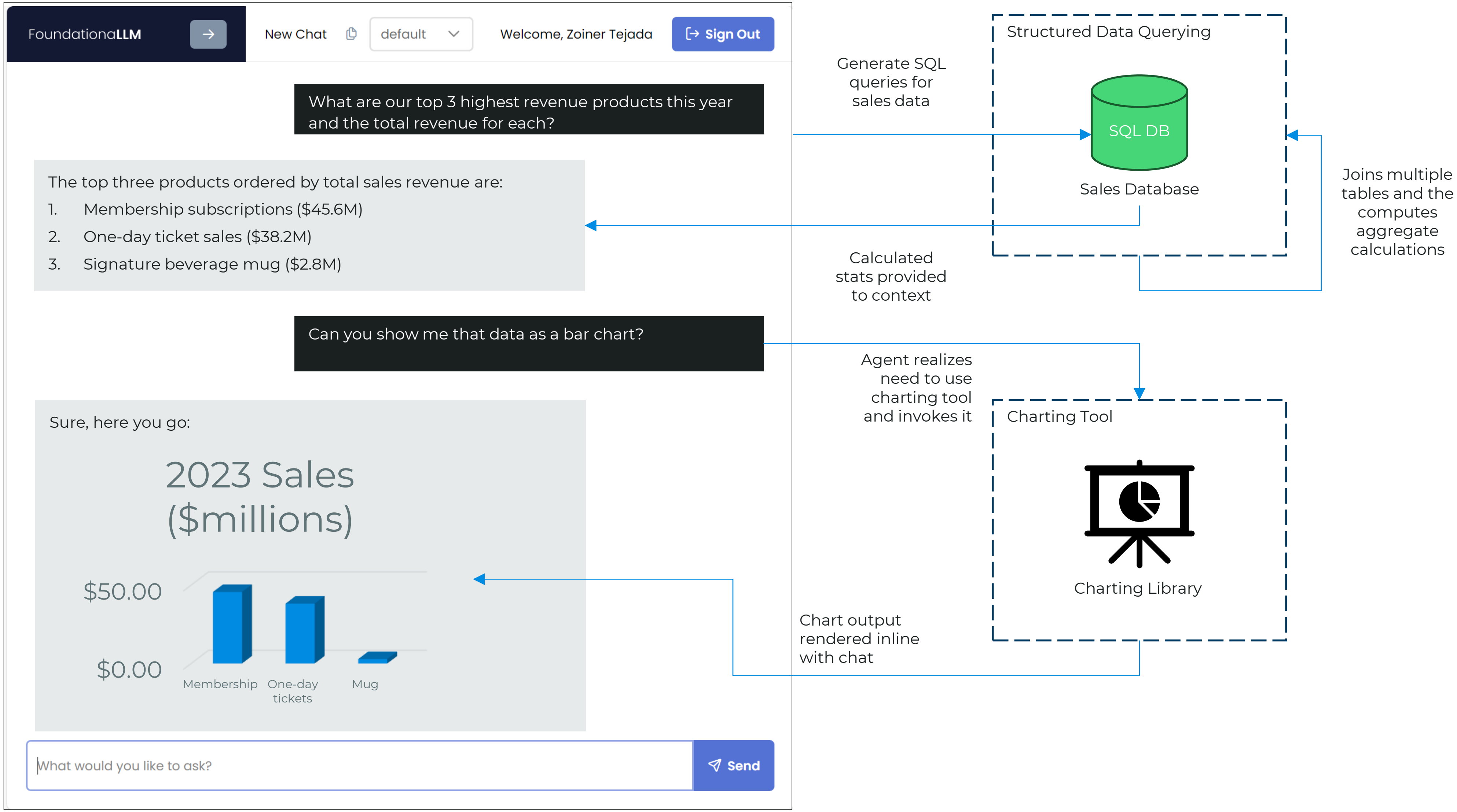




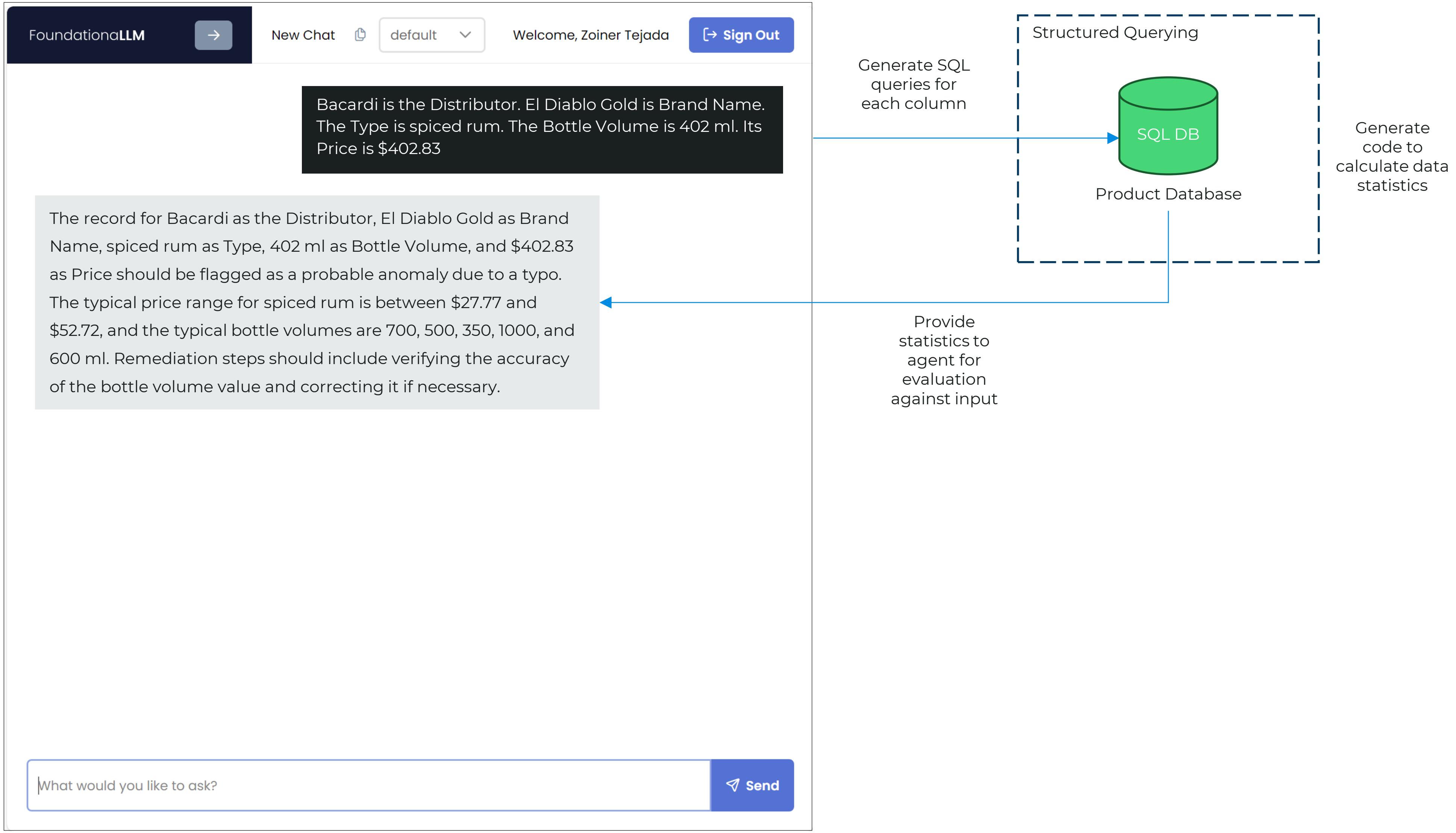
Recommendations from Flat File Data



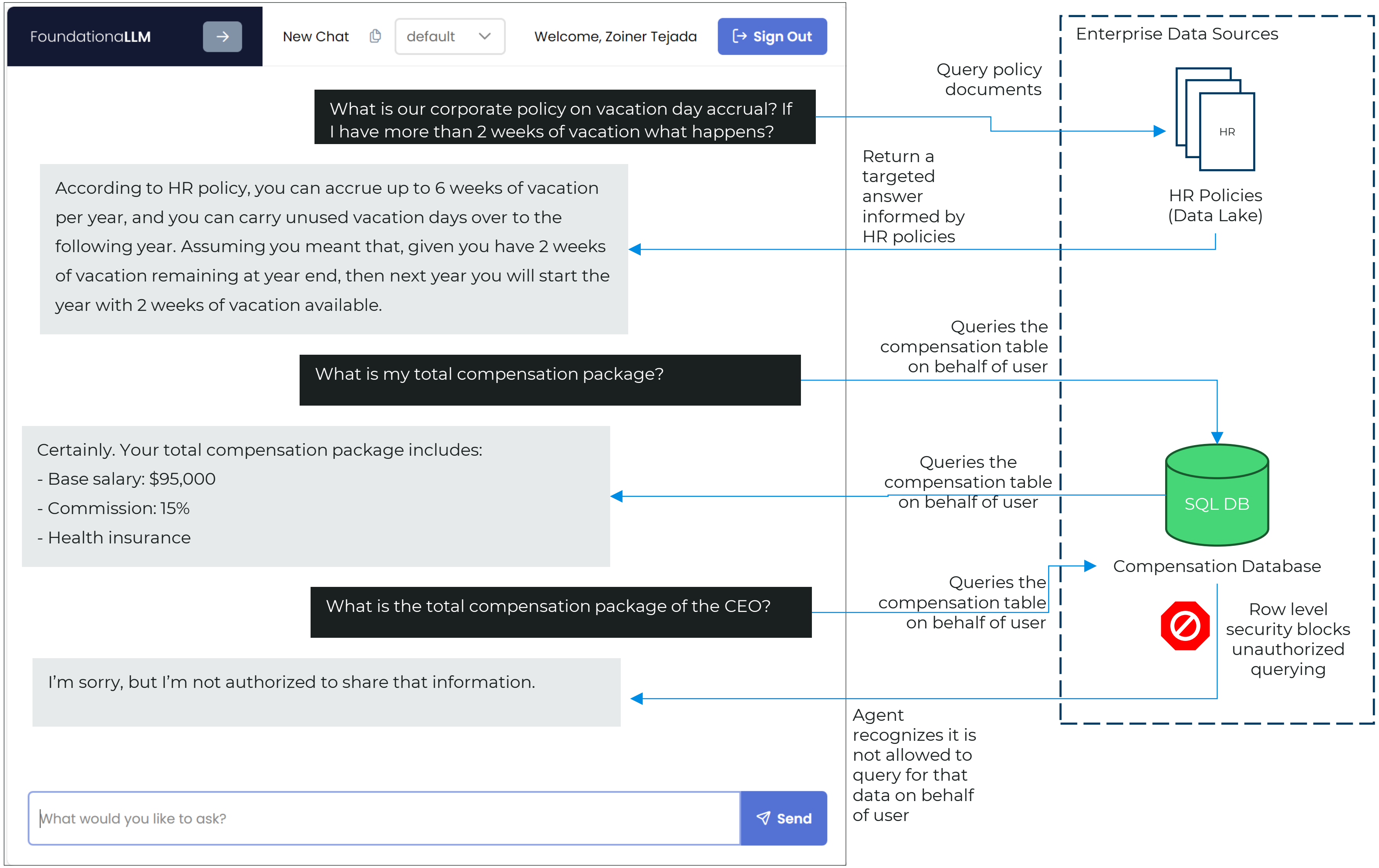
Self-Service Analytics



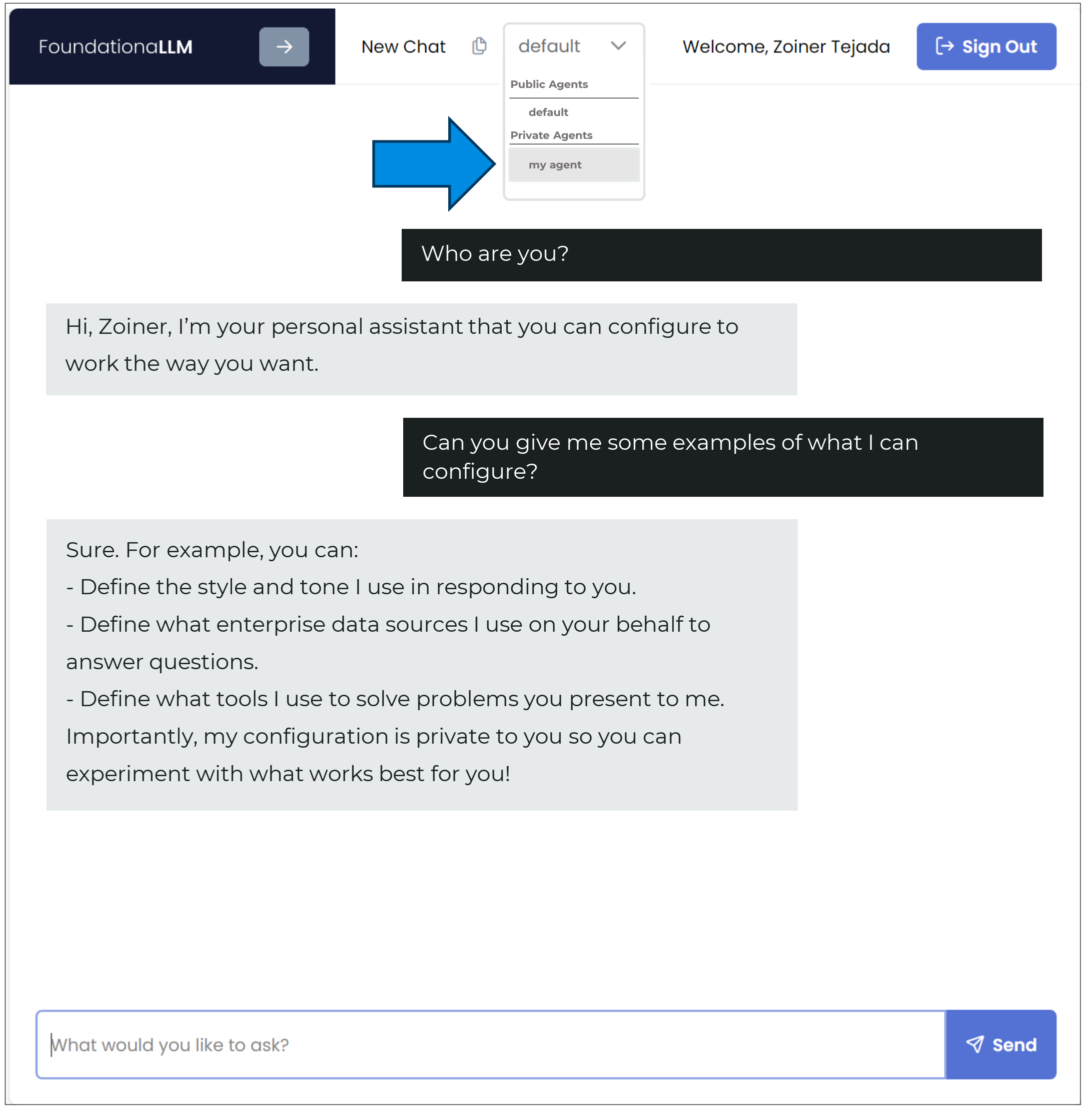
Anomaly Detection for Structured Data



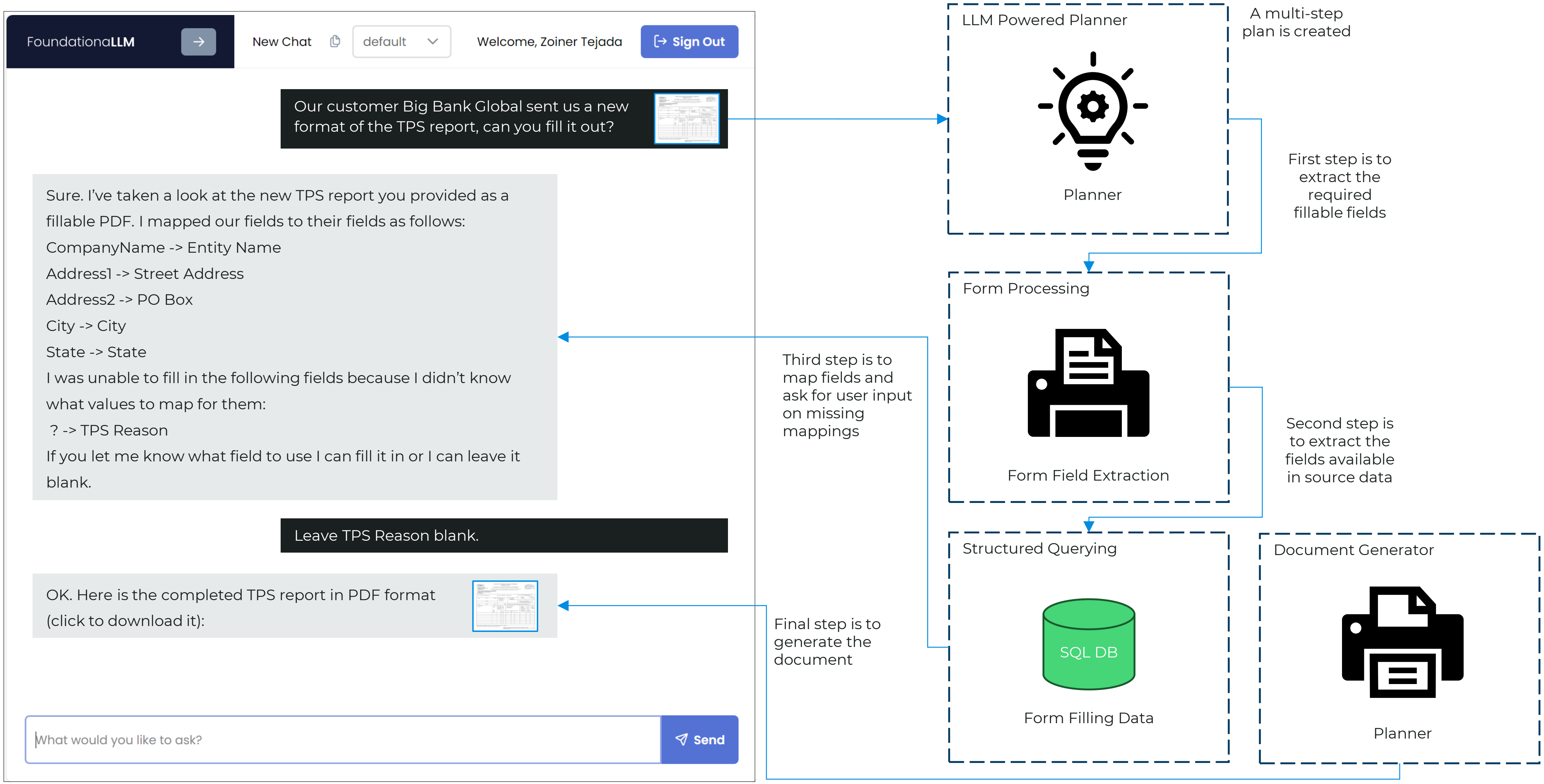
Security aware agents



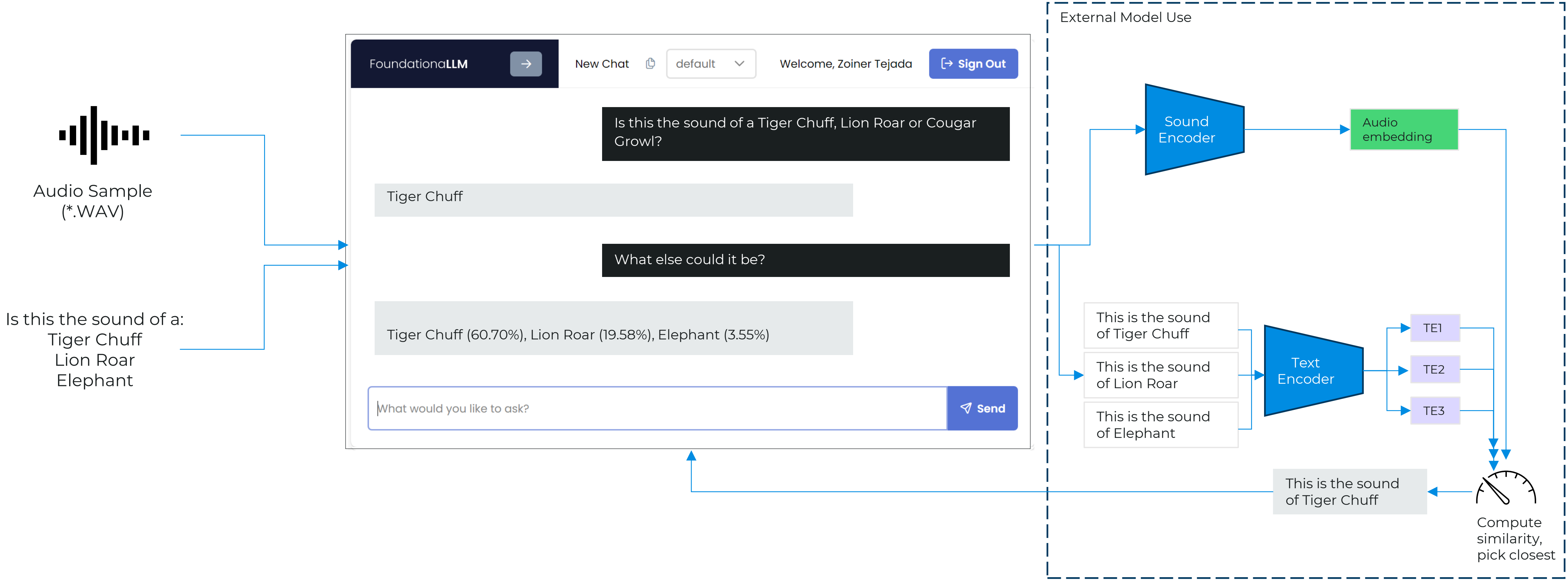
Private agents



Generative form filling & document creation



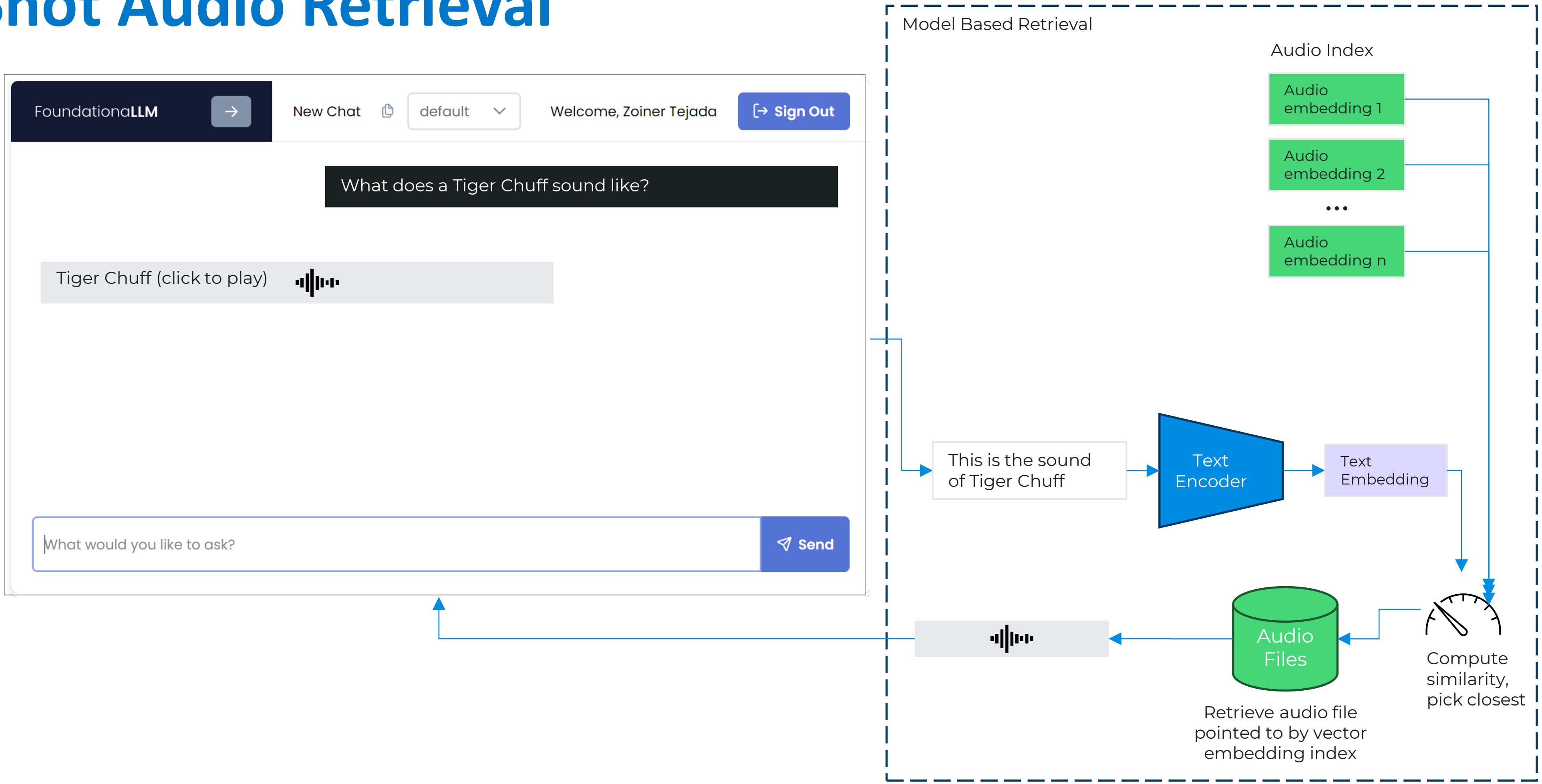
Zero Shot Audio Classification



Sources to learn more

- CLAP: Learning Audio Concepts From Natural Language Supervision(aka CLAP paper)
- Natural Language Supervision for General-Purpose Audio Representations (aka CLAP v2)

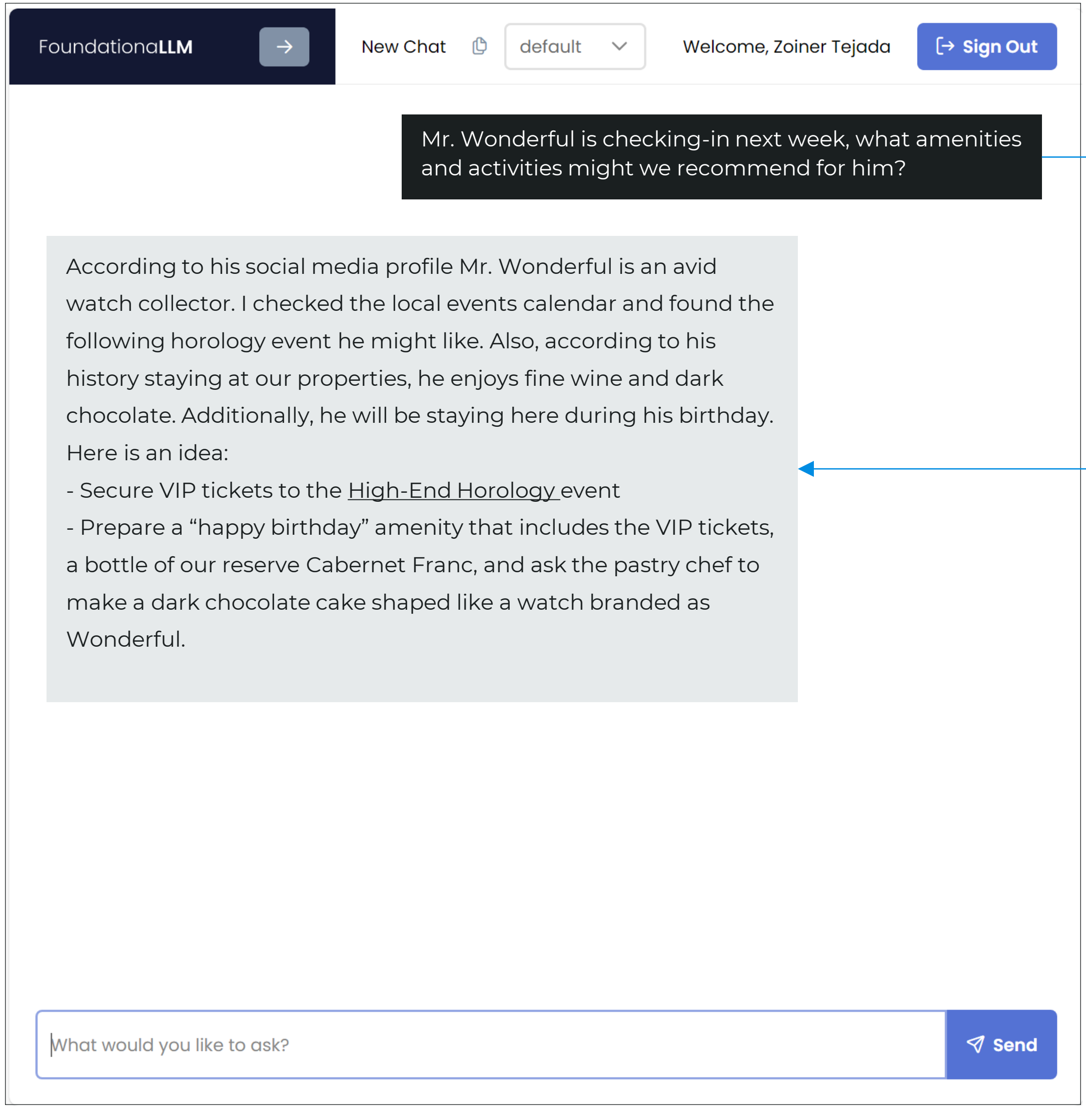
Zero Shot Audio Retrieval

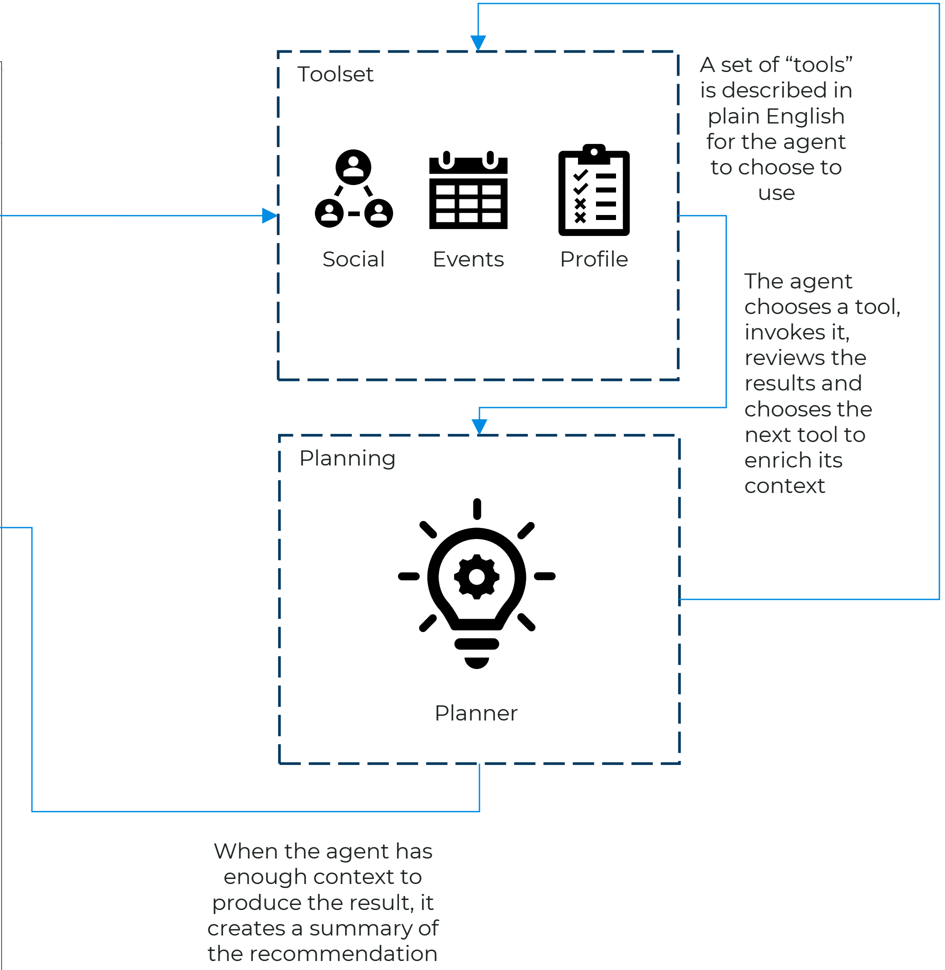


Sources to learn more

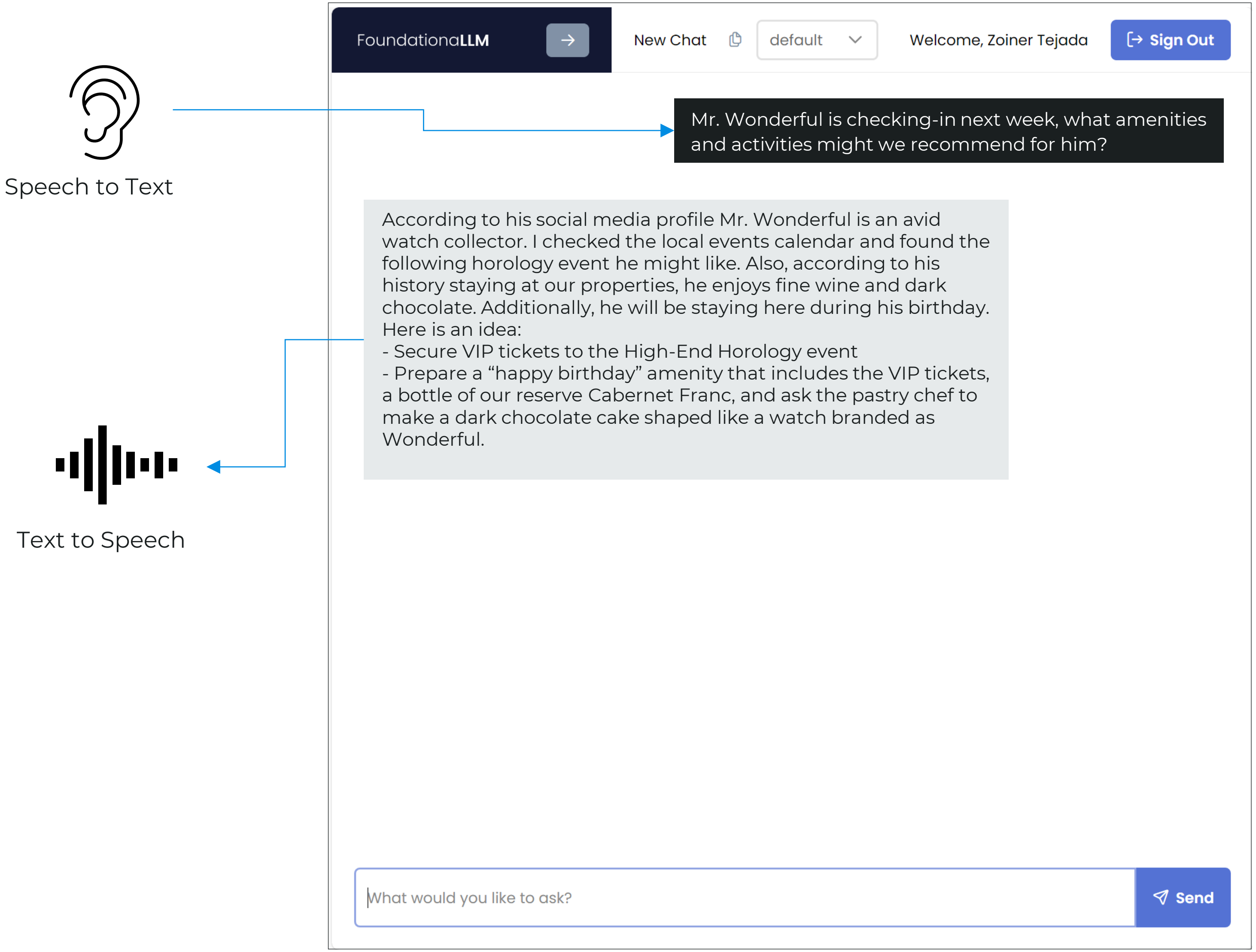
- CLAP: Learning Audio Concepts From Natural Language Supervision(aka CLAP paper)
- Natural Language Supervision for General-Purpose Audio Representations (aka CLAP v2)

Dynamic Problem Solving with Tools

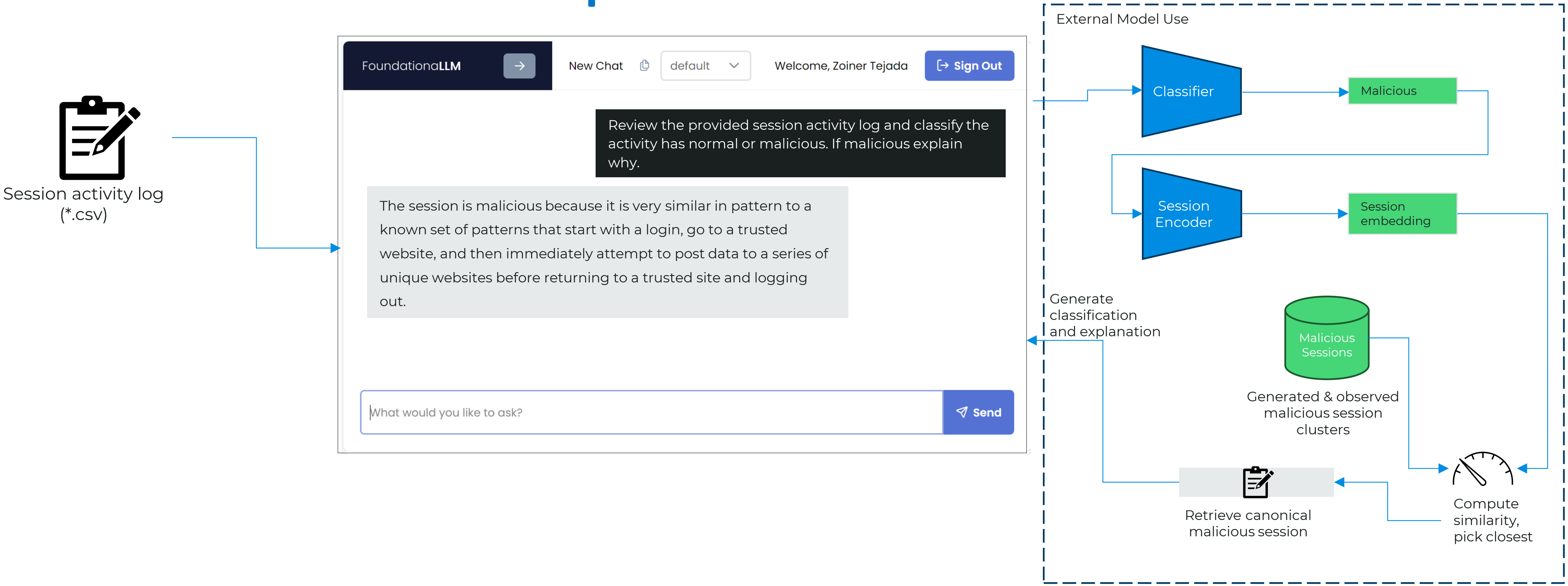




Speech powered agents



Fraud detection and explanation



Sources to learn more

- Robust Fraud Detection from Supervised Contrastive Learning(aka ConRo paper)