

The top corners of the slide feature decorative geometric patterns. On the left, a network of blue lines connects several points, forming a series of triangles. On the right, a similar but more complex network of blue lines is visible. The main title 'FOUNDATIONALLM' is centered in the upper half of the slide. The word 'FOUNDATION' is in a light blue color, and 'ALLM' is in a darker blue color.

FOUNDATIONALLM

FoundationaLLM

Introduction



Joel Hulen

General Manager – Solliance Training

AI Architect - FoundationalLLM



Lino Tadros

AI Solution Architect– Solliance

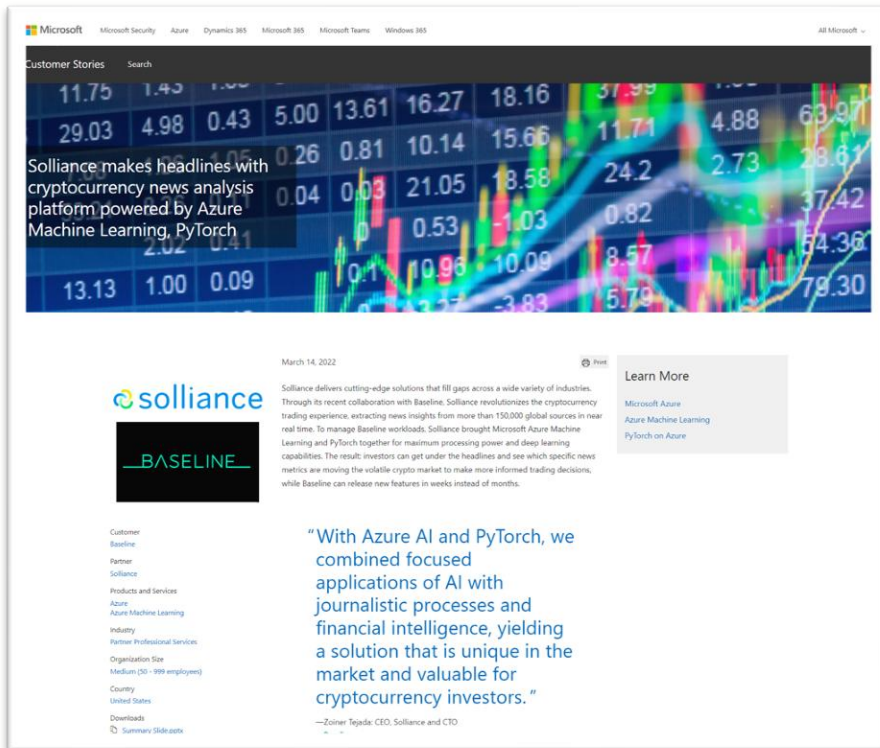
Chief Evangelist- FoundationalLLM

FOUNDATIONALLM

A platform for deploying Gen AI in the enterprise.
scaling
securing
governing

Origin Story

Building GenAI solutions well before ChatGPT moment



Microsoft Customer Stories

Solliance makes headlines with cryptocurrency news analysis platform powered by Azure Machine Learning, PyTorch

March 14, 2022

solliance

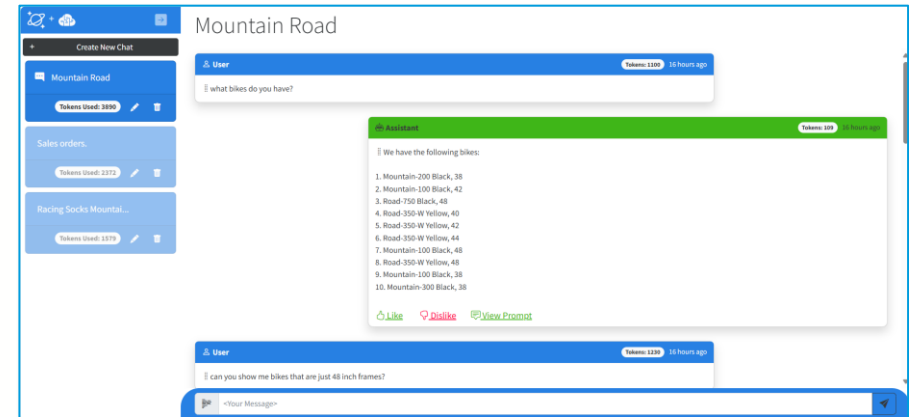
Learn More

Microsoft Azure
Azure Machine Learning
PyTorch on Azure

“With Azure AI and PyTorch, we combined focused applications of AI with journalistic processes and financial intelligence, yielding a solution that is unique in the market and valuable for cryptocurrency investors.”

—Zohier Tejeda, CEO, Solliance and CTO

Partner to MS Engineering



Mountain Road

User: what bikes do you have?

Copilot: We have the following bikes:

1. Mountain-200 Black, 38
2. Mountain-100 Black, 42
3. Road-750 Black, 48
4. Road-350-W Yellow, 40
5. Road-350-W Yellow, 42
6. Road-350-W Yellow, 44
7. Mountain-100 Black, 48
8. Road-350-W Yellow, 48
9. Mountain-100 Black, 38
10. Mountain-300 Black, 38

User: can you show me bikes that are just 48 inch frames?

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](https://github.com/Azure/BuildYourOwnCopilot)

The Problems



Enterprise GenAI is complicated, time consuming & capital intensive.



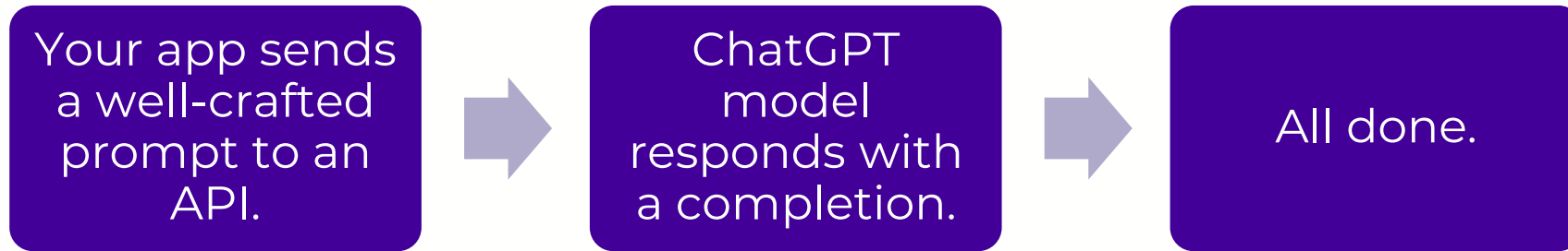
Enterprises need comprehensive platforms that are compliant for high-risk data
- not components, toys or accelerators.



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:

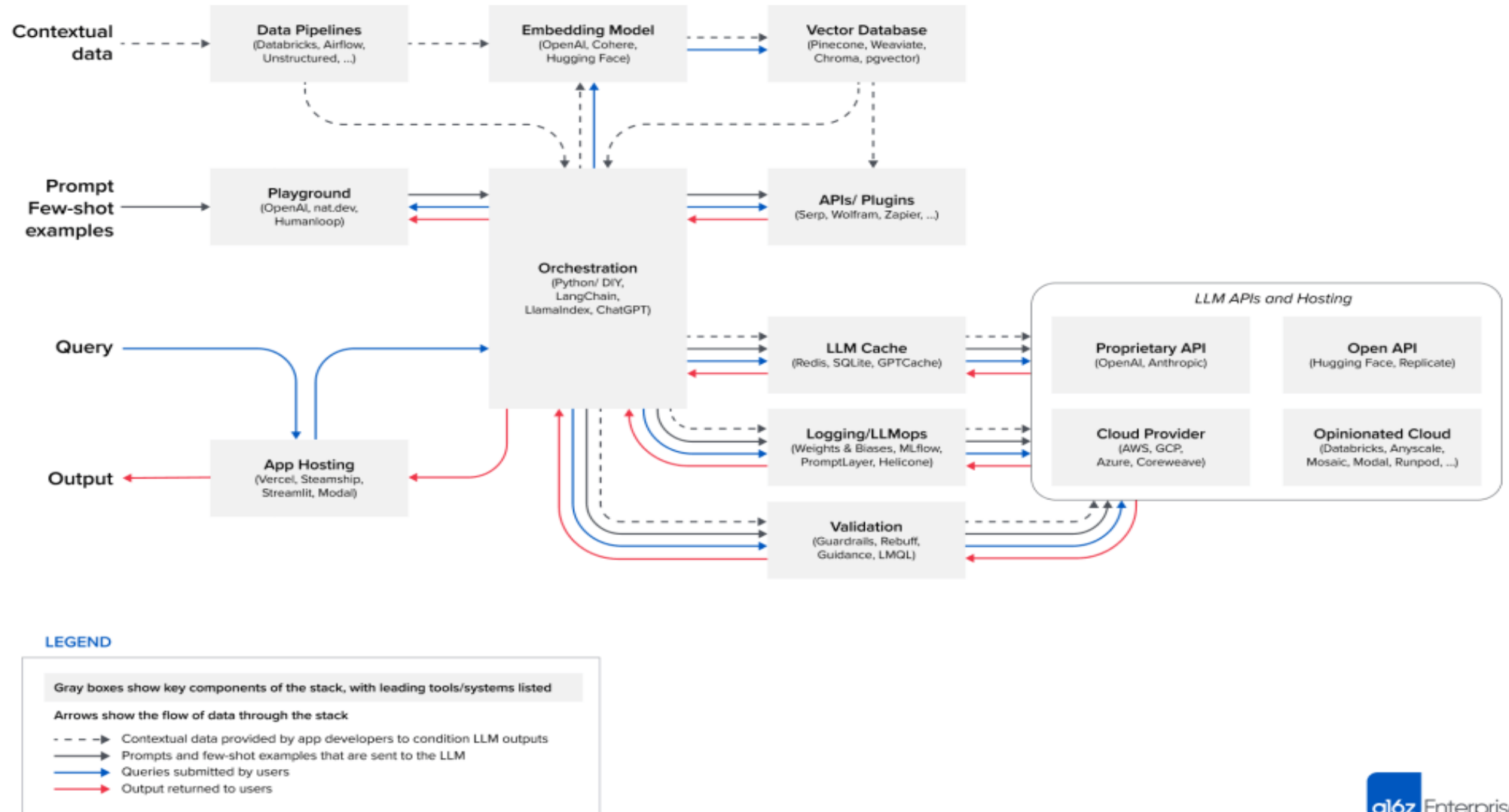


As an enterprise, how do I?

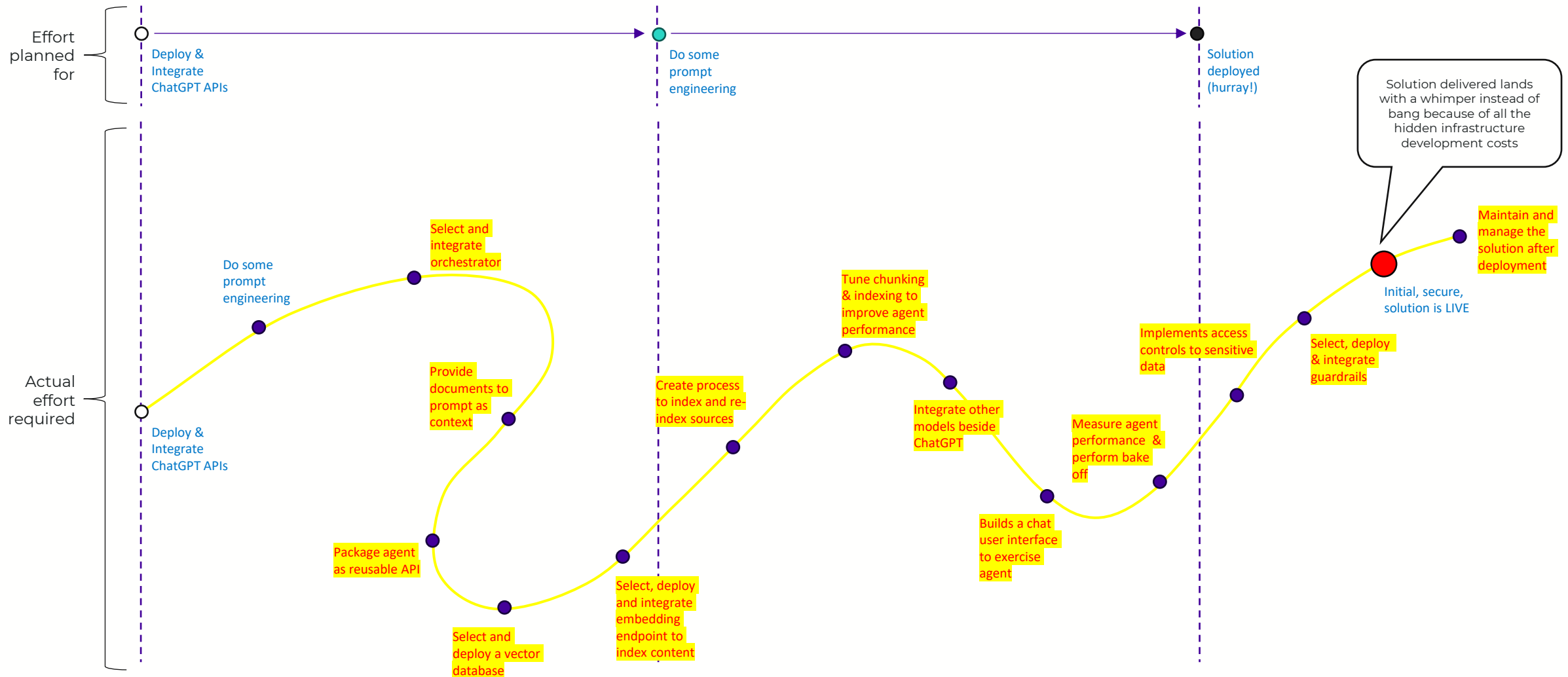
provide a branded chat user experience atop this?	manage conversation history on a per user basis?	enable API integration into my automated workflows?	use more advanced interaction patterns that are recursive, create code or use tools?
scale and batch load thousands to millions of documents as knowledge 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):



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.”

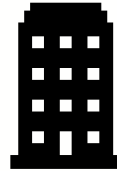
Generative AI endeavours...are costing from \$5 million to \$20 million in upfront investments.

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 DALL-E-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

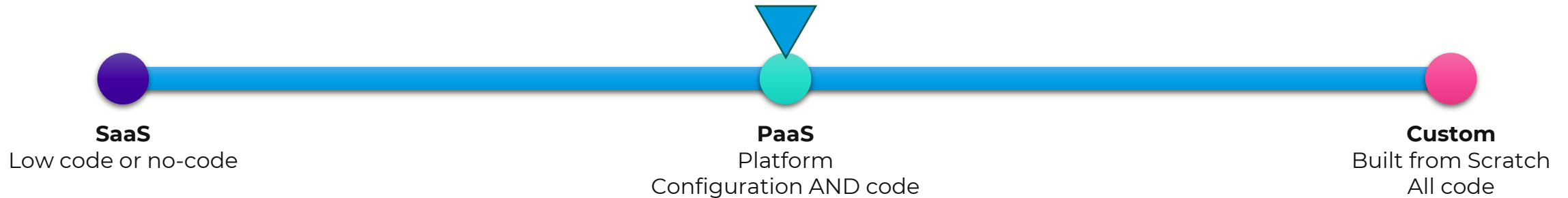
Mistral

OpenAI GPT-4o

OpenAI GPT-o1

Strike a balance between self service and build from scratch

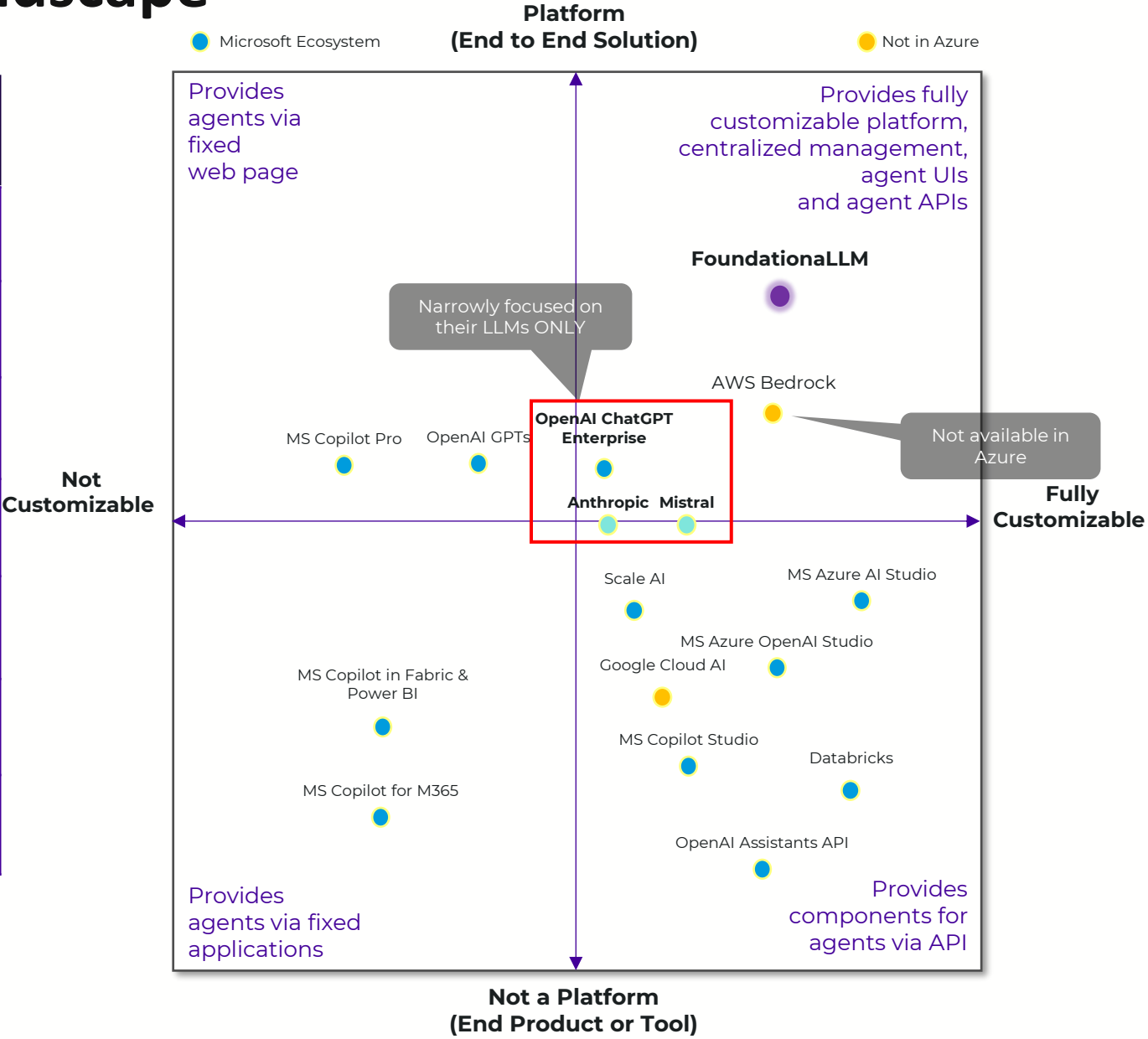
FoundationalLLM



Overview of LLM Platform Landscape

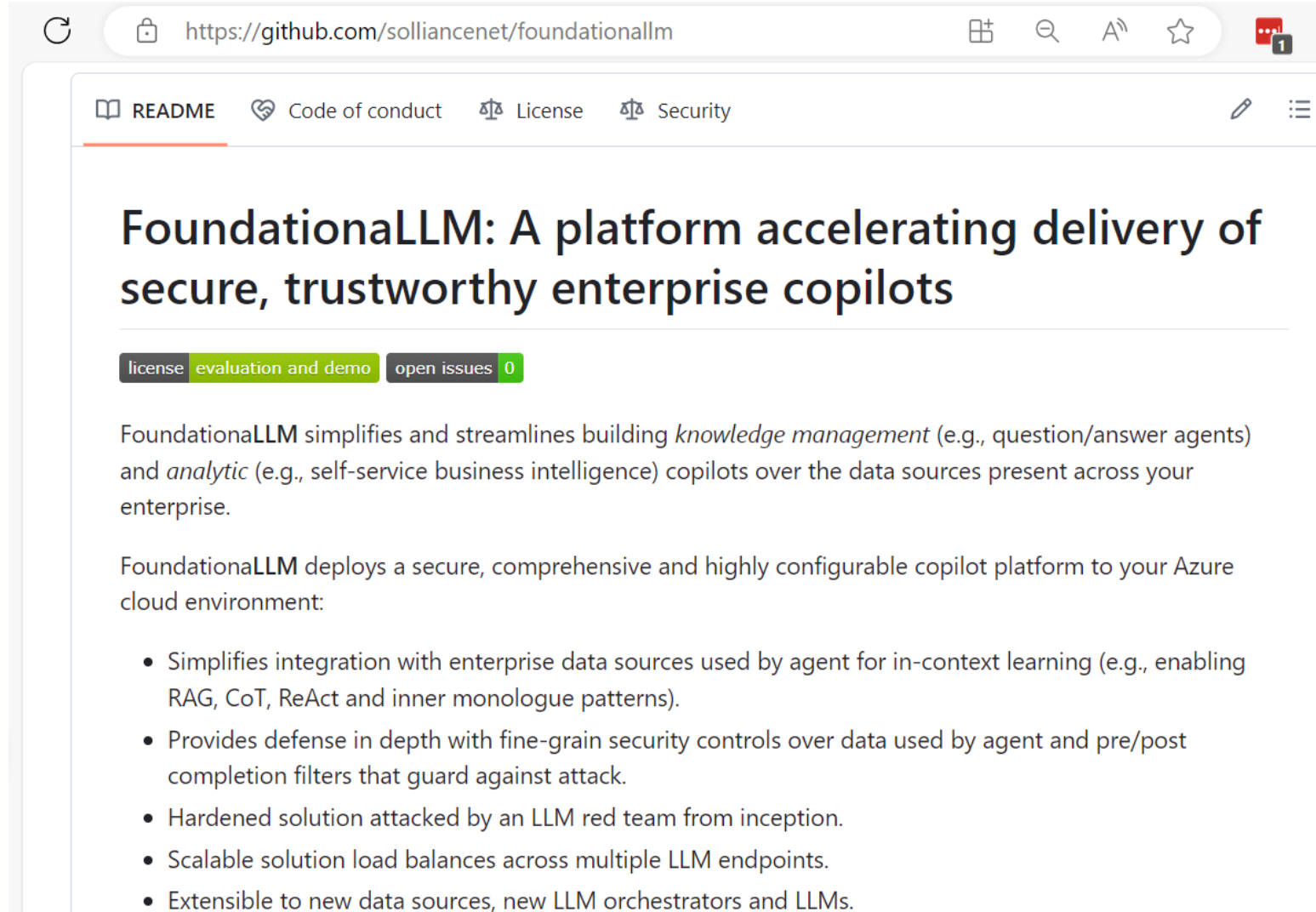
Create LLMs	
Company	Product Description
Anthropic	Claude family of models
Cohere	Command family of models
Databricks	DBRX family of models
Google Cloud AI	Gemma & Gemini family of models
Open AI	GPT family of models
Meta	Llama family of models
Mistral	Mistral family of models
Not what we do.	

Create Software for Using LLMs	
Company	Product Description
FoundationalLLM	FoundationalLLM
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



FoundationalLLM source code is out in the open.

See it for yourself on
GitHub.

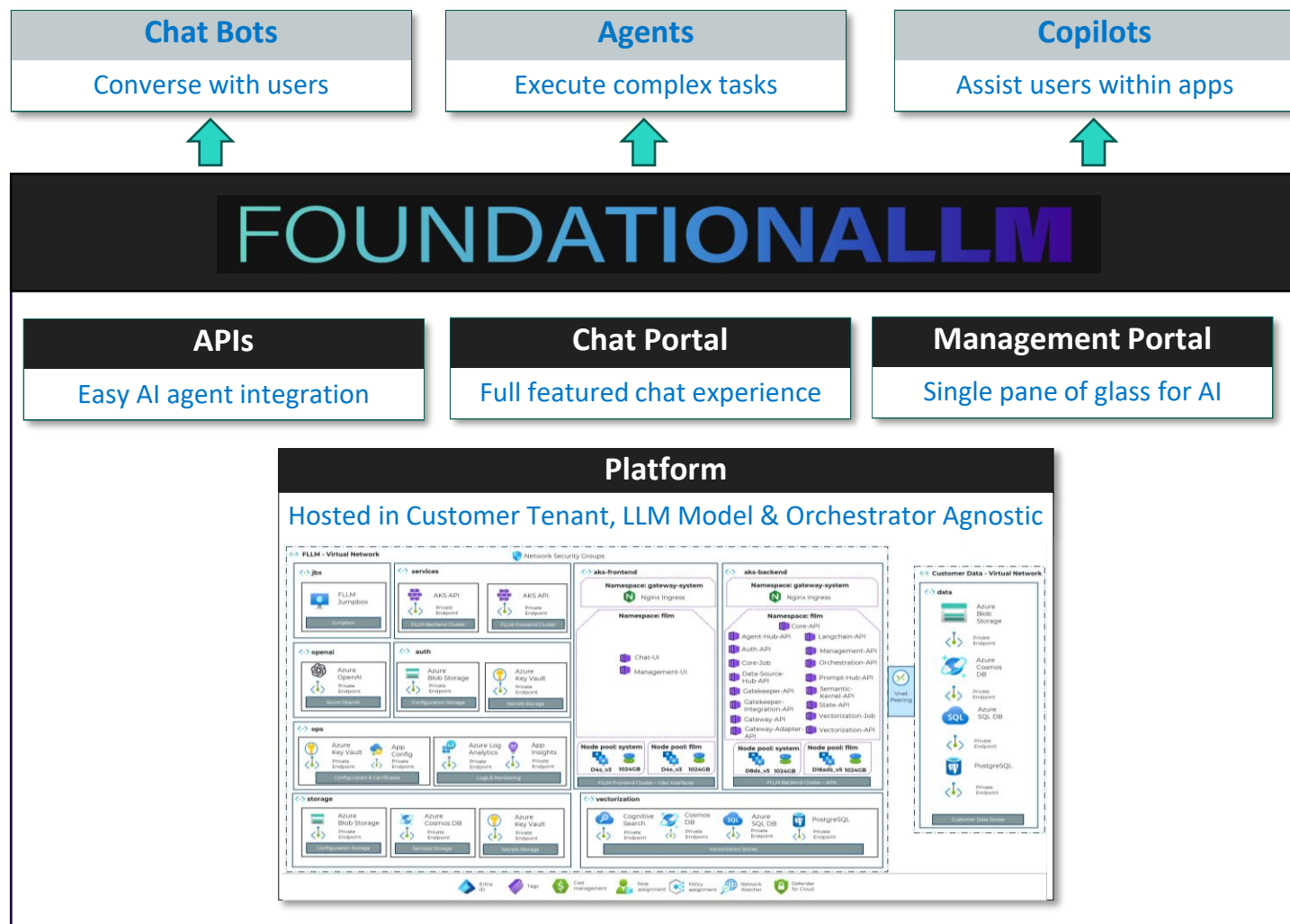


The screenshot shows the GitHub repository page for FoundationalLLM. The browser address bar displays the URL <https://github.com/solliancenetwork/foundationallm>. The repository navigation bar includes links for README (highlighted), Code of conduct, License, and Security. The main heading reads "FoundationalLLM: A platform accelerating delivery of secure, trustworthy enterprise copilots". Below the heading are three buttons: "license", "evaluation and demo" (highlighted in green), and "open issues 0". The repository description states: "FoundationalLLM 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." It further states: "FoundationalLLM deploys a secure, comprehensive and highly configurable copilot platform to your Azure cloud environment:". A bulleted list of features follows:

- 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.

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

Enable customers to create these bespoke AI solutions:

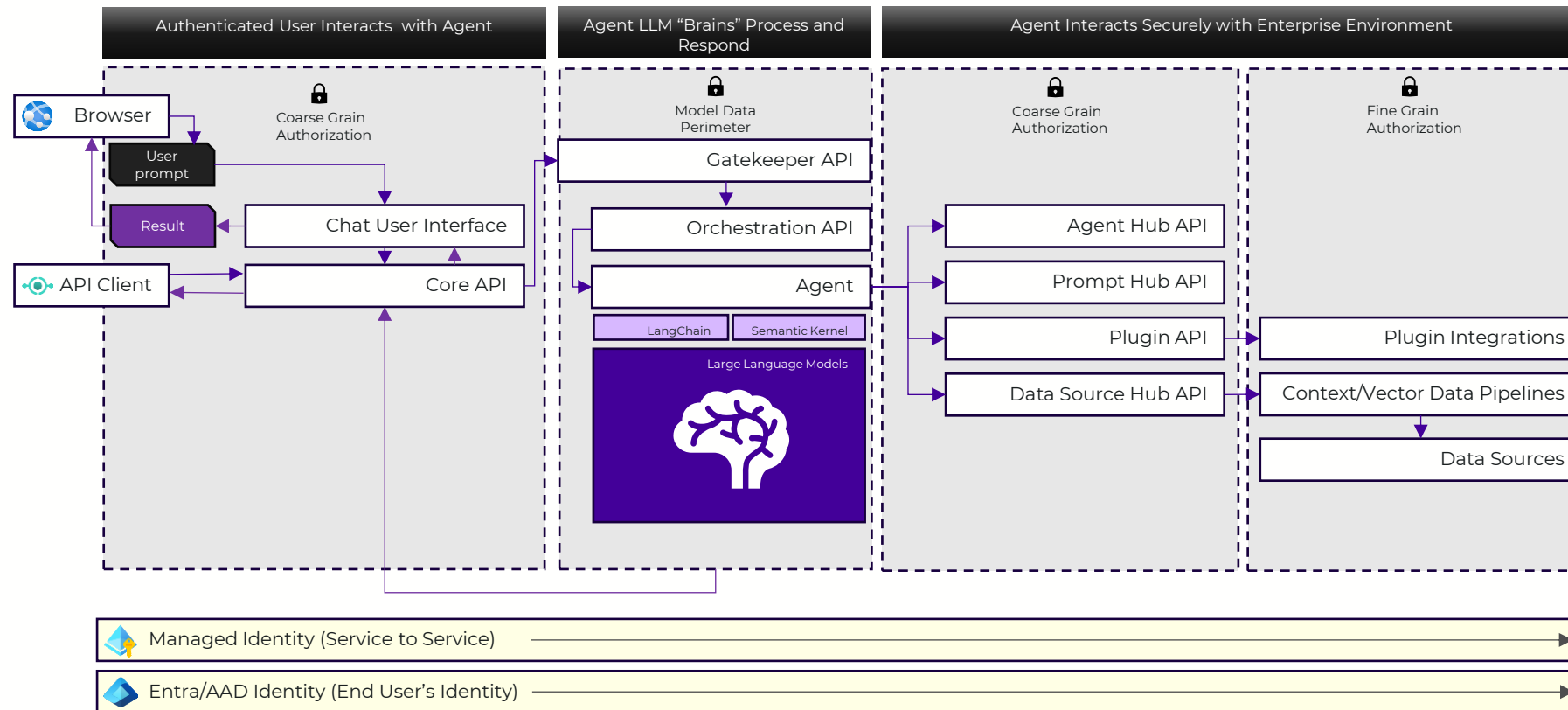


Differentiator #1 – Comprehensive, secure and flexible platform for chatbots, copilots and agentic solutions that deploys into customer Azure subscription.

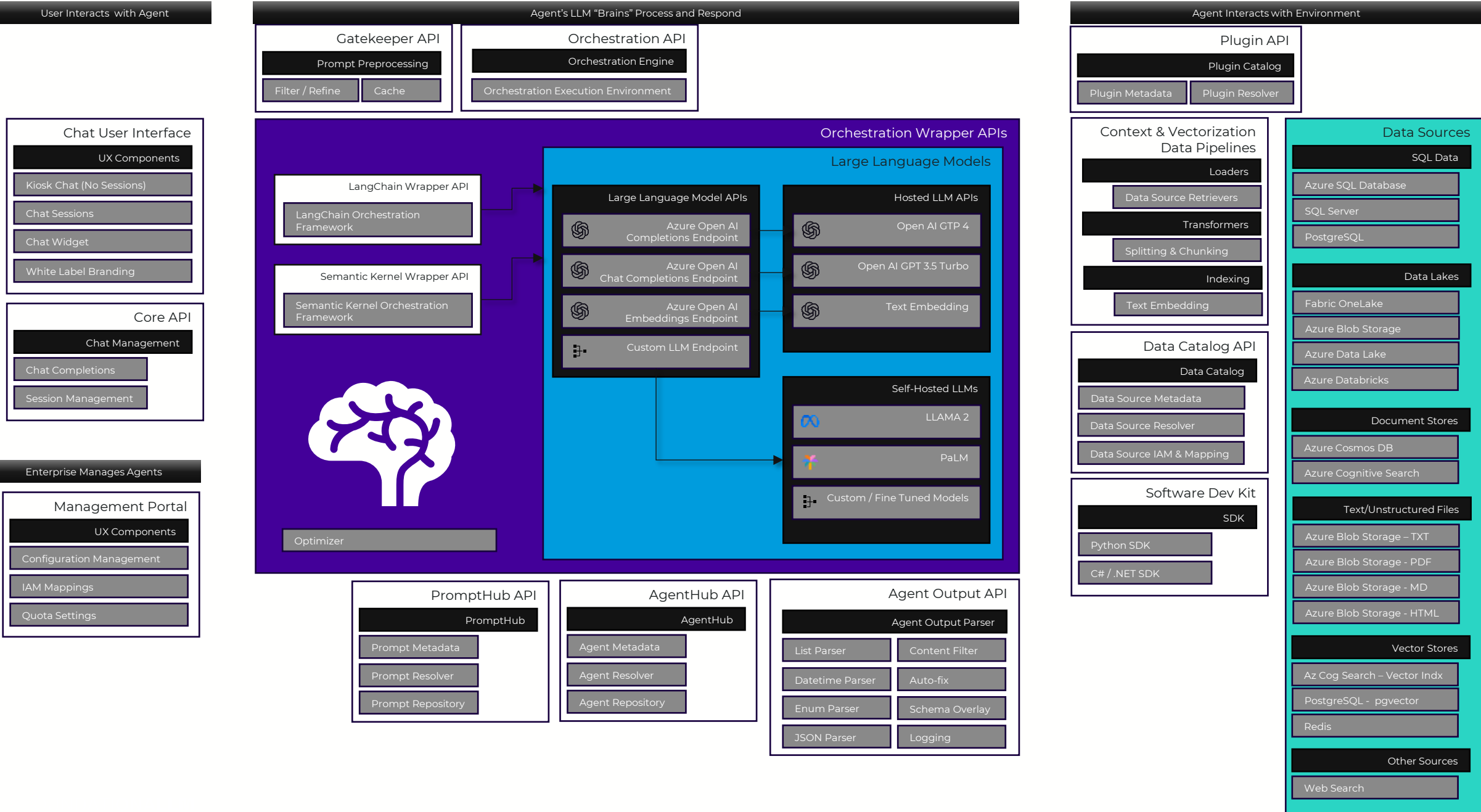
Differentiator #2 – API first platform designed for easy integration and provides a rich chat user interface ready to use out of the box.

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

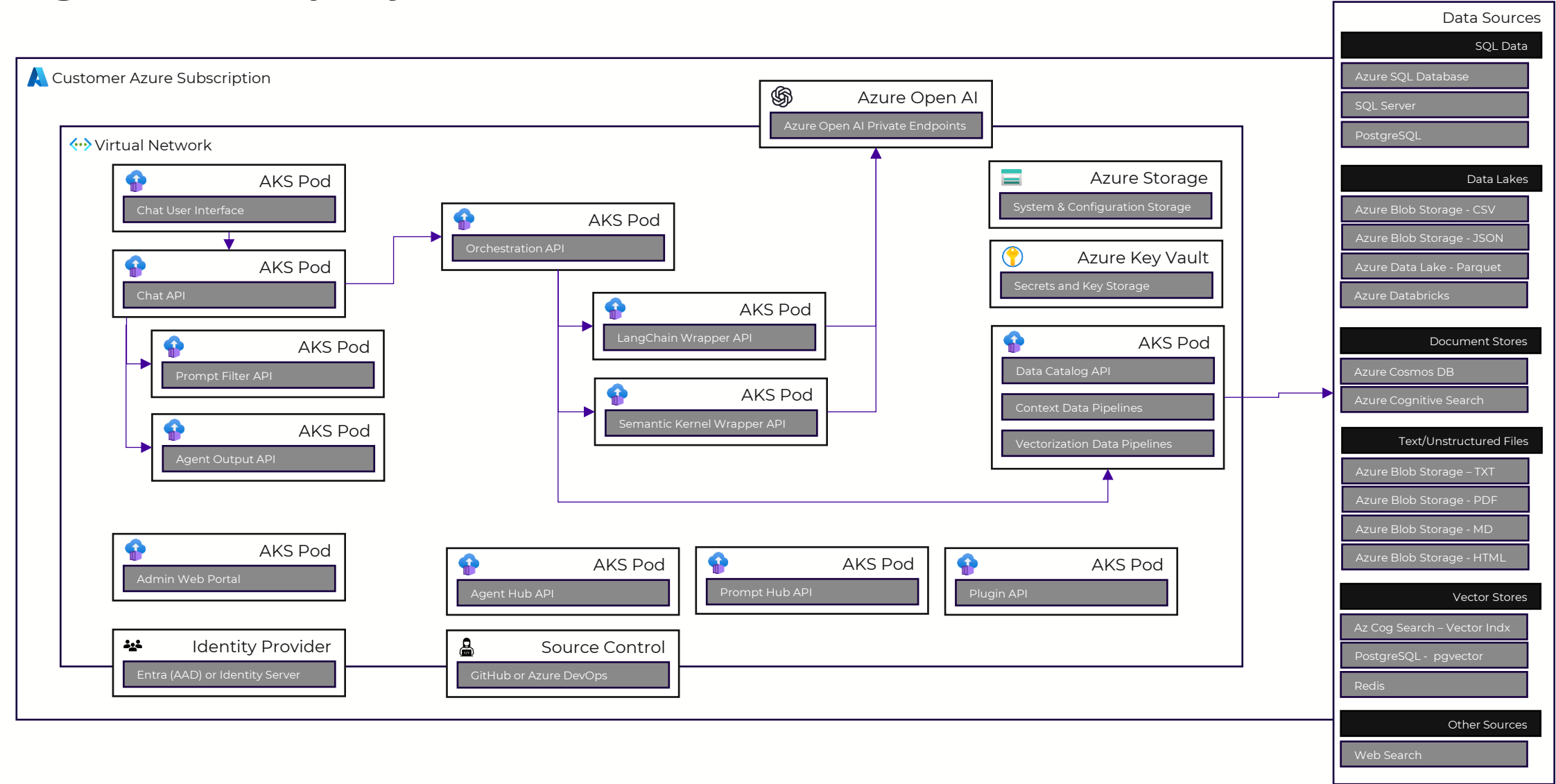
FoundationaLLM provides the **platform** for deploying, scaling, securing and governing **generative AI** in the **enterprise**.



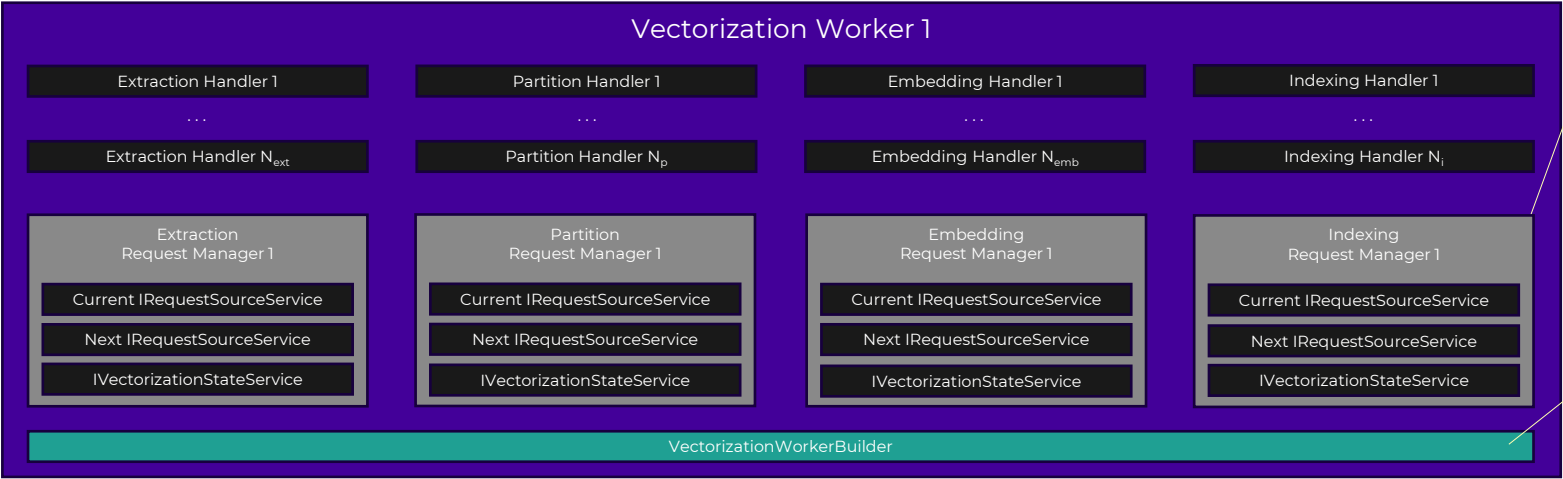
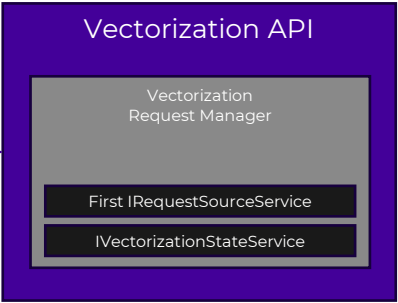
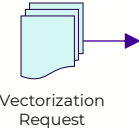
FoundationalLLM enables trustworthy models grounded in enterprise data



High level deployment architecture



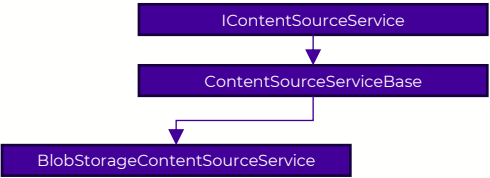
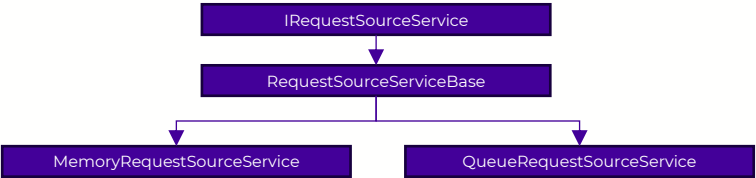
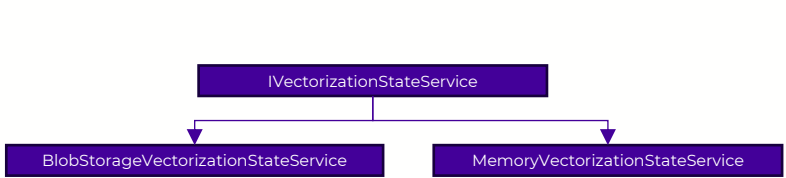
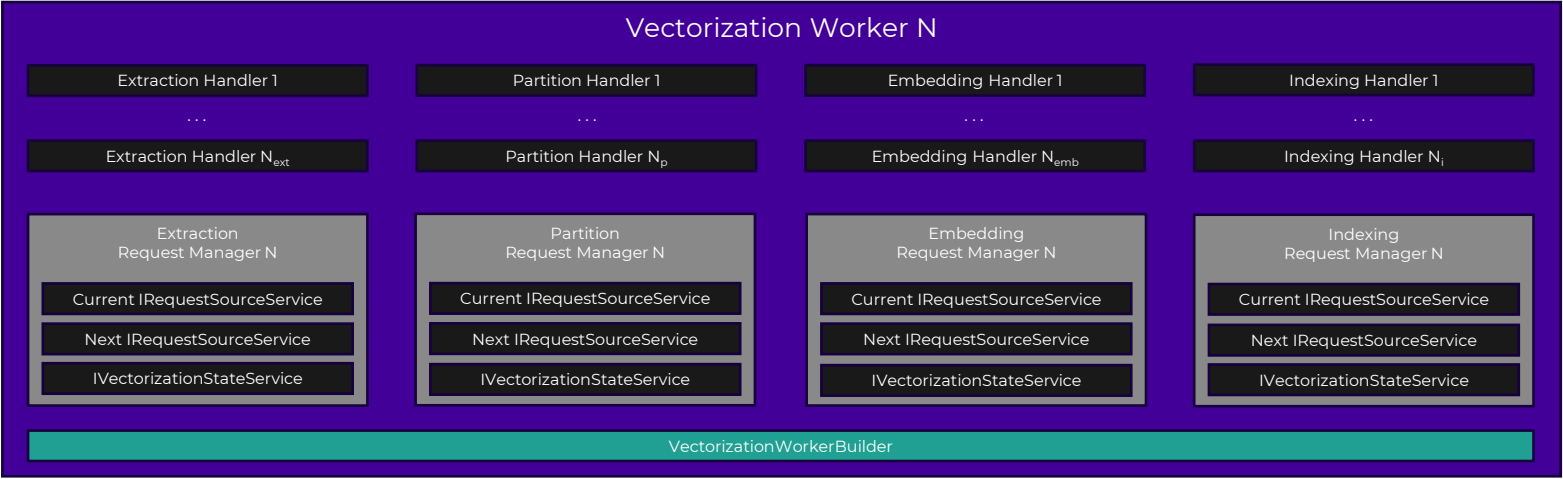
Vectorization



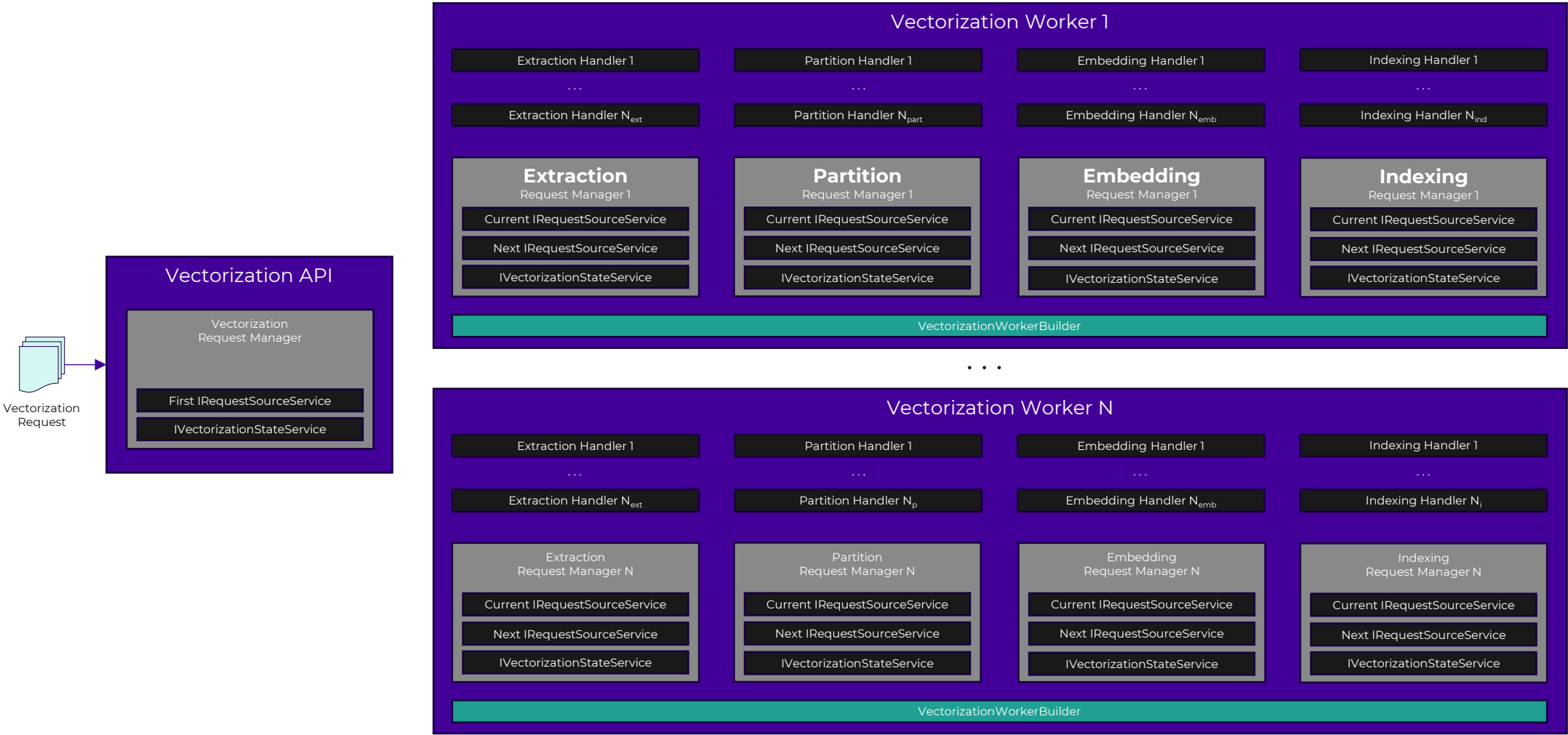
- Read next request from request source
- Invoke appropriate handler
- Create new request in the next request source
- Update the state of the vectorization pipeline

Creates all request manager instances based on the vectorization worker configuration

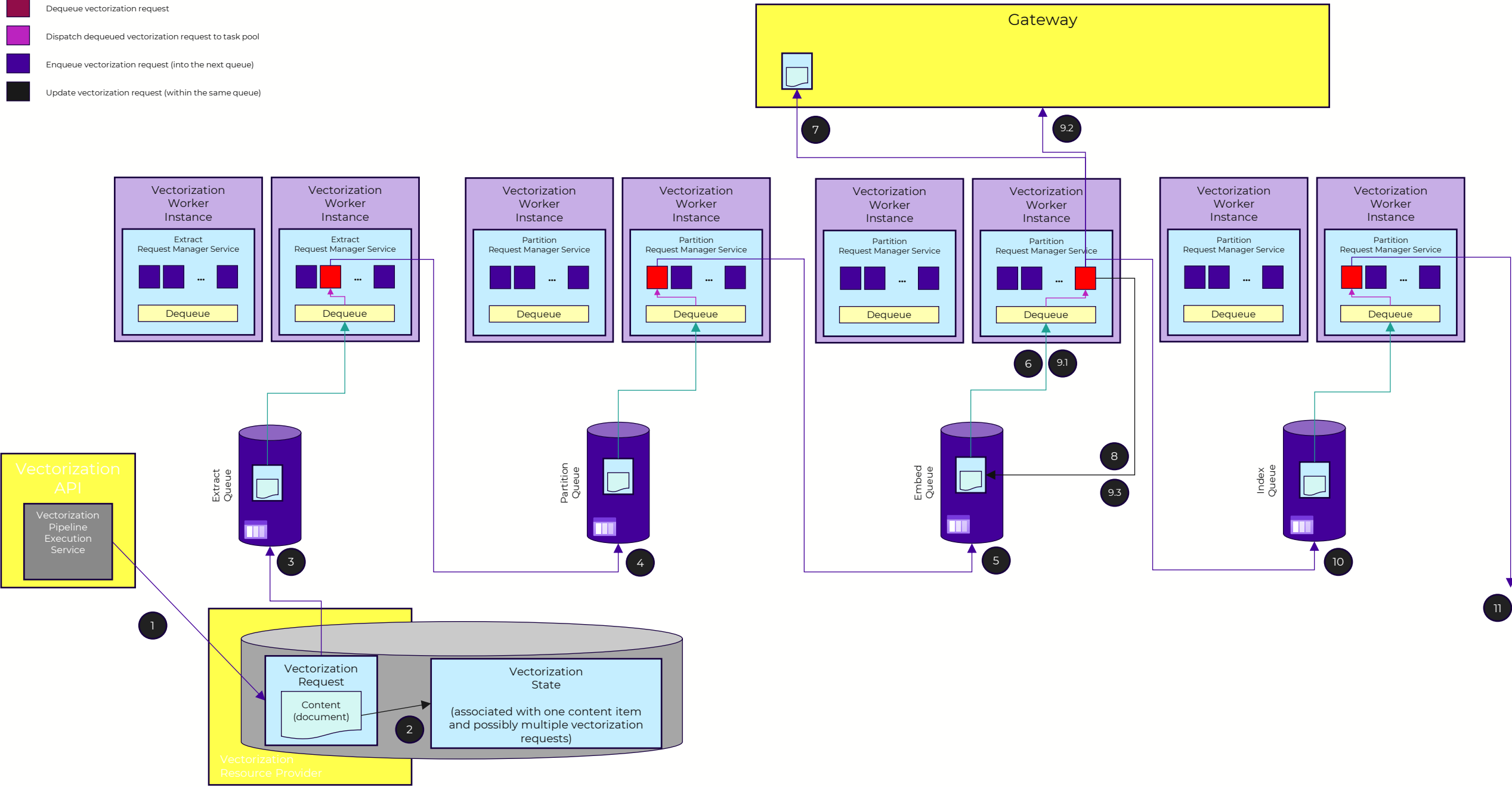
...



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)



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 : 16834895-4b4a-4a41-8776-4d88f4d3173d

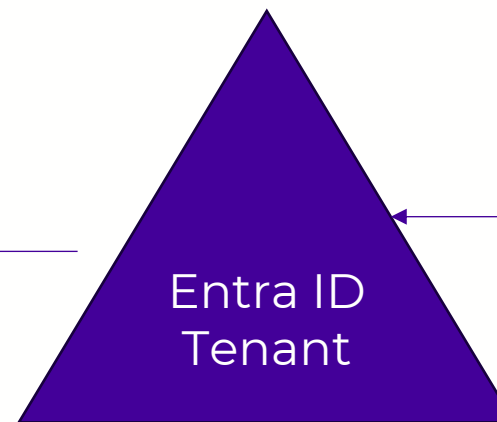
Directory : Solliance, Inc. (solliance.net)

Status : Active

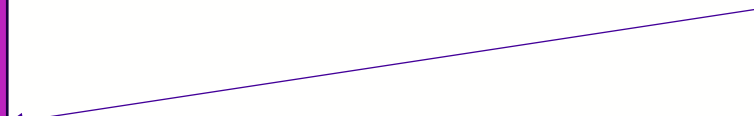
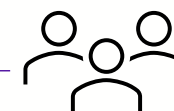
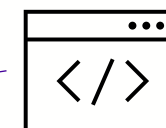
Parent management group : 16834895-4b4a-4a41-8776-4d88f4d3173d

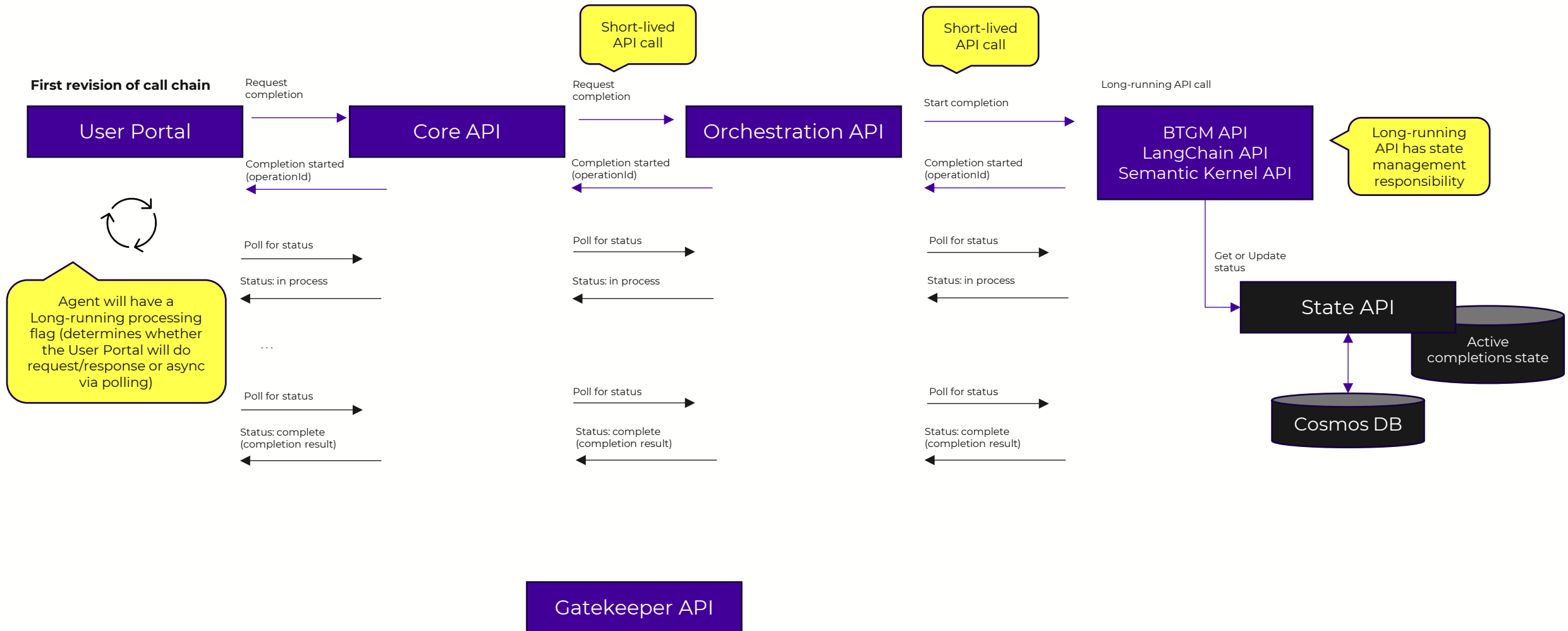
Spending rate and forecast

14,000



solliancenet.onmicrosoft.com
solliance.net





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

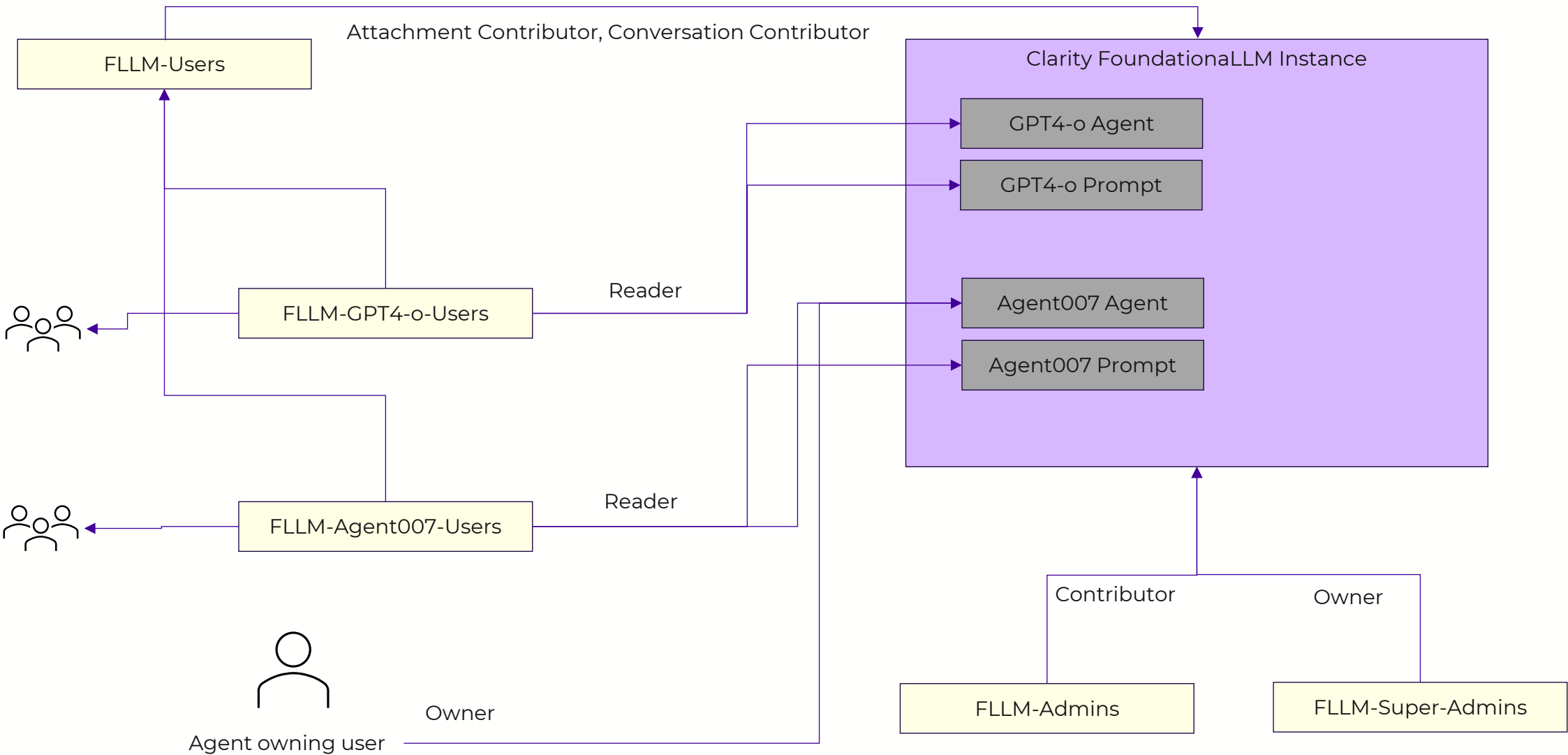
Conversation Contributor assignment to the user scoped at the FoundationalLLM instance level

Attachment Contributor assignment to the user scoped at the FoundationalLLM instance level

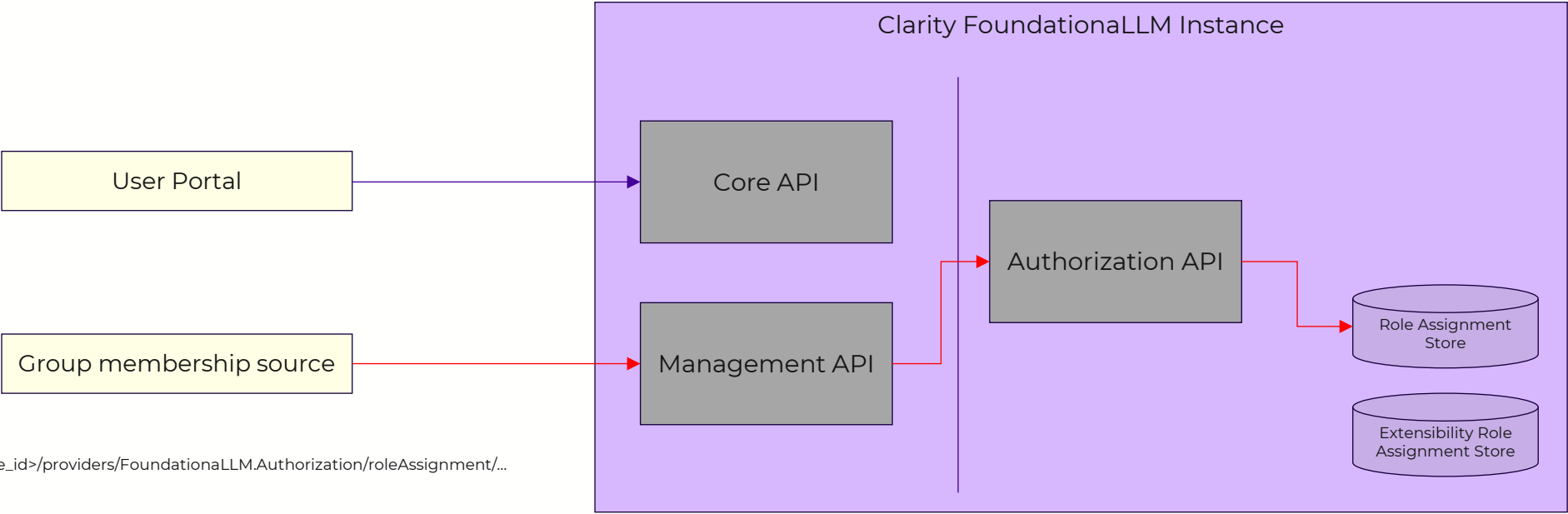
Reader role assignment to two specific configuration items

(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

Create New Agent
Complete the settings below to create and deploy your new agent.

Agent name:
No special characters or spaces, lowercase letters with dashes and underscores only.
Enter agent name

Description:
Provide a description to help others understand the agent's purpose.
Enter agent description

Type

What type of agent?

☒ **Knowledge Management**
Best for Q&A, summarization and reasoning over textual data.

☐ **Analytics**
Best to query, analyze, calculate and report on tabular data.

Knowledge Source

Do you want this agent to have a dedicated pipeline?

☒ Yes

Where is the data?
Please select a data source.

Where should the data be indexed?
Please select an index source.

How should the data be processed for indexing?

Splitting & Chunking
Chunk size: 500
Overlap size: 50

When should the data be indexed?

Trigger
Runs every time a new item is added to the data source.
Frequency: Event

Configure Content Safety & Data Protection

How should user-agent interactions be gated?

Gatekeeper
Enabled: ☒ Yes

Content Safety:
Azure Content Safety ☐ Azure Content Safety Prompt Shield ☒

Data Protection: Microsoft Presidio ☒

Done

Easily create agents using OpenAI Assistants API

Agent Capabilities: OpenAI Assistants ☒

☐ OpenAI Assistants

☐ FLM Knowledge Management

Done

Manage All Deployed Agents

All Agents
View your publicly accessible agents.

Name	Type	Expiration Date	Edit	Delete
FoundationalLLM	knowledge-management	Never		
GPT-4o	knowledge-management	Never		
Khalil	knowledge-management	Never		
Neighbor	knowledge-management	Mon Sep 30 2024 00:00:00 GMT-0400 (Eastern Daylight Time)		
Simple_Agent	knowledge-management	Never		
Test	knowledge-management	Never		

Configure Role Based Access Control on Agents

Access Control

Name	Type	Scope	Delete
Contributor			
Owner			
Reader			
id-management-api-bj26gd7aperbg	User	Instance (Inherited)	
id-core-api-bj26gd7aperbg	User	Instance (Inherited)	
id-vectorization-api-bj26gd7aperbg	User	Instance (Inherited)	
id-orchestration-api-bj26gd7aperbg	User	Instance (Inherited)	
id-gatekeeper-api-bj26gd7aperbg	User	Instance (Inherited)	
id-vectorization-job-bj26gd7aperbg	User	Instance (Inherited)	
id-gateway-api-bj26gd7aperbg	User	Instance (Inherited)	
Austin Hulien (Austin.Hulien@solliance.net)	User	Instance (Inherited)	
Sol Machiraju (Sol_Machiraju@solliance.net)	User	Instance (Inherited)	
Joel Hulien (Joel.Hulien@solliance.net)	User	Instance (Inherited)	

Close + Add role assignment for this resource

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

FoundationalLLM 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)

FoundationLLM

CHATS

Vision to Chart

Code Explanation

Multi Vision

Code Gen & Explanation

Equation Plot

Memory Game

Joke

PDF Generation

Charting

Vision

Summarize Excel and PDF

Zolner Tejada

Charting

Zolner Tejada

Sample.xlsx

Using the spreadsheet, create a bar chart comparing the "Total Interest Income" versus the "Salaries and Other Personal Expense" by year.

GPT-4o

Interest Income vs Salaries and Other Personal Expense by Year

Here's the bar chart comparing the "Total Interest Income" versus the "Salaries and Other Personal Expense" by year. The chart shows the comparative figures for "Total Interest Income" and "Salaries and Other Personal Expense" over the years 2020, 2021, and 2022.

Analysis

Tool: code_interpreter

Category: OpenAI Assistants

python

```
import pandas as pd
import matplotlib.pyplot as plt

# Read the spreadsheet data
interest_income_df = pd.read_excel('Sample.xlsx', sheet_name='Interest Income')
salaries_expense_df = pd.read_excel('Sample.xlsx', sheet_name='Salaries and Other Personal Expense')

# Extract the relevant columns
interest_income_df = interest_income_df[['Year', 'Interest Income']]
salaries_expense_df = salaries_expense_df[['Year', 'Salaries and Other Personal Expense']]

# Convert data to numeric types for plotting
interest_income_df['Interest Income'] = pd.to_numeric(interest_income_df['Interest Income'], errors='coerce')
salaries_expense_df['Salaries and Other Personal Expense'] = pd.to_numeric(salaries_expense_df['Salaries and Other Personal Expense'], errors='coerce')

# Create the bar chart
fig, ax = plt.subplots()
interest_income_df.plot(x='Year', y='Interest Income', kind='bar', label='Interest Income')
salaries_expense_df.plot(x='Year', y='Salaries and Other Personal Expense', kind='bar', label='Salaries and Other Personal Expense')

# Add titles and labels
plt.title('Interest Income vs Salaries and Other Personal Expense by Year')
plt.xlabel('Year')
plt.ylabel('Amount')
plt.legend()
plt.grid(True)

# Show the plot
plt.show()
```

Output: # Generated image file assistant=4a2c705K9cT0BXXQ3QcVp

Tool: code_interpreter

Category: OpenAI Assistants

python

```
# Extract the rows relevant for interest income
interest_income_df = df.iloc[2:5, [0, 1, 2, 3]]
interest_income_df.columns = ['Income Type', '2022', '2021', '2020']

# Convert data to numeric types for plotting
interest_income_df['2022'] = pd.to_numeric(interest_income_df['2022'], errors='coerce')
interest_income_df['2021'] = pd.to_numeric(interest_income_df['2021'], errors='coerce')
interest_income_df['2020'] = pd.to_numeric(interest_income_df['2020'], errors='coerce')
```

Switch between agents

--select--

All Agents

FoundationLLM

GPT-4o

Claude

Mistral

View analysis logs

Analysis

Tool: code_interpreter

Category: OpenAI Assistants

python

```
import pandas as pd
import matplotlib.pyplot as plt

# Read the spreadsheet data
interest_income_df = pd.read_excel('Sample.xlsx', sheet_name='Interest Income')
salaries_expense_df = pd.read_excel('Sample.xlsx', sheet_name='Salaries and Other Personal Expense')

# Extract the relevant columns
interest_income_df = interest_income_df[['Year', 'Interest Income']]
salaries_expense_df = salaries_expense_df[['Year', 'Salaries and Other Personal Expense']]

# Convert data to numeric types for plotting
interest_income_df['Interest Income'] = pd.to_numeric(interest_income_df['Interest Income'], errors='coerce')
salaries_expense_df['Salaries and Other Personal Expense'] = pd.to_numeric(salaries_expense_df['Salaries and Other Personal Expense'], errors='coerce')

# Create the bar chart
fig, ax = plt.subplots()
interest_income_df.plot(x='Year', y='Interest Income', kind='bar', label='Interest Income')
salaries_expense_df.plot(x='Year', y='Salaries and Other Personal Expense', kind='bar', label='Salaries and Other Personal Expense')

# Add titles and labels
plt.title('Interest Income vs Salaries and Other Personal Expense by Year')
plt.xlabel('Year')
plt.ylabel('Amount')
plt.legend()
plt.grid(True)

# Show the plot
plt.show()
```

Output: # Generated image file assistant=4a2c705K9cT0BXXQ3QcVp

Tool: code_interpreter

Category: OpenAI Assistants

python

```
# Extract the rows relevant for interest income
interest_income_df = df.iloc[2:5, [0, 1, 2, 3]]
interest_income_df.columns = ['Income Type', '2022', '2021', '2020']

# Convert data to numeric types for plotting
interest_income_df['2022'] = pd.to_numeric(interest_income_df['2022'], errors='coerce')
interest_income_df['2021'] = pd.to_numeric(interest_income_df['2021'], errors='coerce')
interest_income_df['2020'] = pd.to_numeric(interest_income_df['2020'], errors='coerce')
```

Upload files

Loans_Held_For_Sale_Year_Over_Year_Changes.xlsx

Local Computer

Uploaded

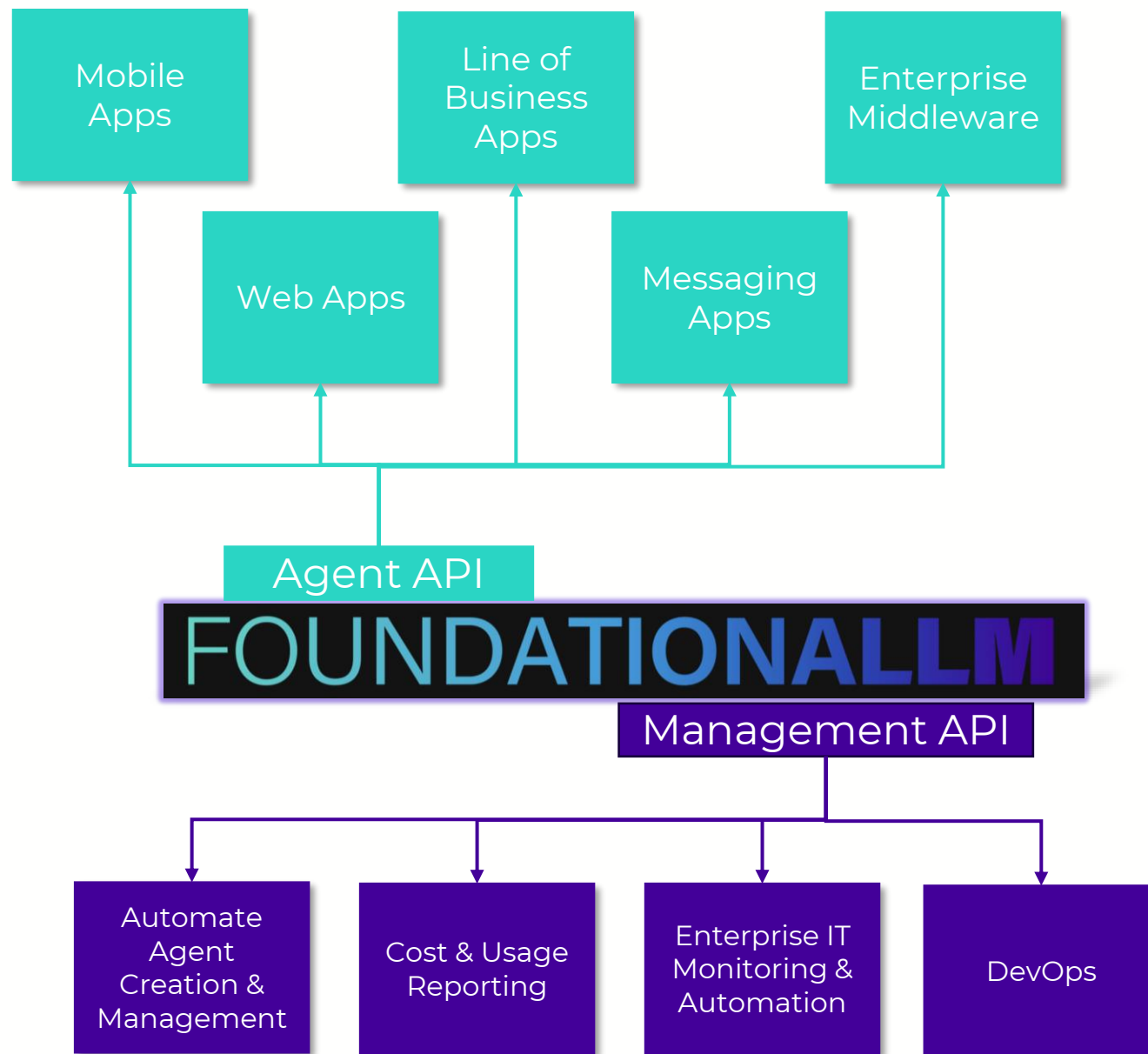
Select file from Computer

Connect to OneDrive

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



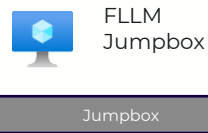
FLLM Standard Deployment Components

Customer Azure Subscription

FLLM - Virtual Network

Network Security Groups

jbx



services



AKS API
Private Endpoint

FLLM Backend Cluster



AKS API
Private Endpoint

FLLM Frontend Cluster

openai



auth



Azure Blob Storage
Private Endpoint

Configuration Storage



Azure Key Vault
Private Endpoint

Secrets Storage

ops



Azure Key Vault
Private Endpoint



App Config
Private Endpoint

Configuration & Certificates



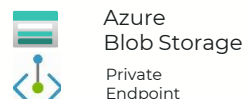
Azure Log Analytics
Private Endpoint



App Insights
Private Endpoint

Logs & Monitoring

storage



Azure Cosmos DB
Private Endpoint

Services Storage



Azure Key Vault
Private Endpoint

Secrets Storage

aks-frontend

Namespace: gateway-system



Nginx Ingress

Namespace: flim



Chat-UI



Management-UI

Node pool: system



D4s_v3 1024GB

Node pool: flim



D4s_v3 1024GB

FLLM Frontend Cluster – User Interfaces

aks-backend

Namespace: gateway-system

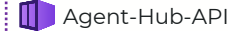


Nginx Ingress

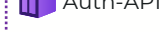
Namespace: flim



Core-API



Agent-Hub-API



Auth-API



Core-Job



Data-Source-Hub-API



Gatekeeper-API



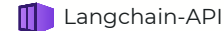
Gatekeeper-Integration-API



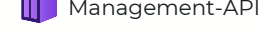
Gateway-API



Gateway-Adapter-API



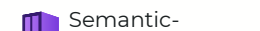
Langchain-API



Management-API



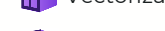
Orchestration-API



Prompt-Hub-API



Semantic-Kernel-API



State-API



Vectorization-Job



Vectorization-API

Node pool: system



D8ds_v5 1024GB

Node pool: flim



D16ads_v5 1024GB

FLLM Backend Cluster – APIs

vectorization



AI Search
Private Endpoint



Cosmos DB
Private Endpoint



Azure SQL DB
Private Endpoint



PostgreSQL
Private Endpoint

Vectorization Stores



Customer Data - Virtual Network

data



Azure Blob Storage



Private Endpoint



Azure Cosmos DB



Private Endpoint



Azure SQL DB



Private Endpoint



PostgreSQL



Private Endpoint

Customer Data Stores



Entra ID



Tags



Cost management



Role assignment



Policy assignment

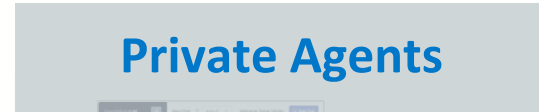


Network Watcher

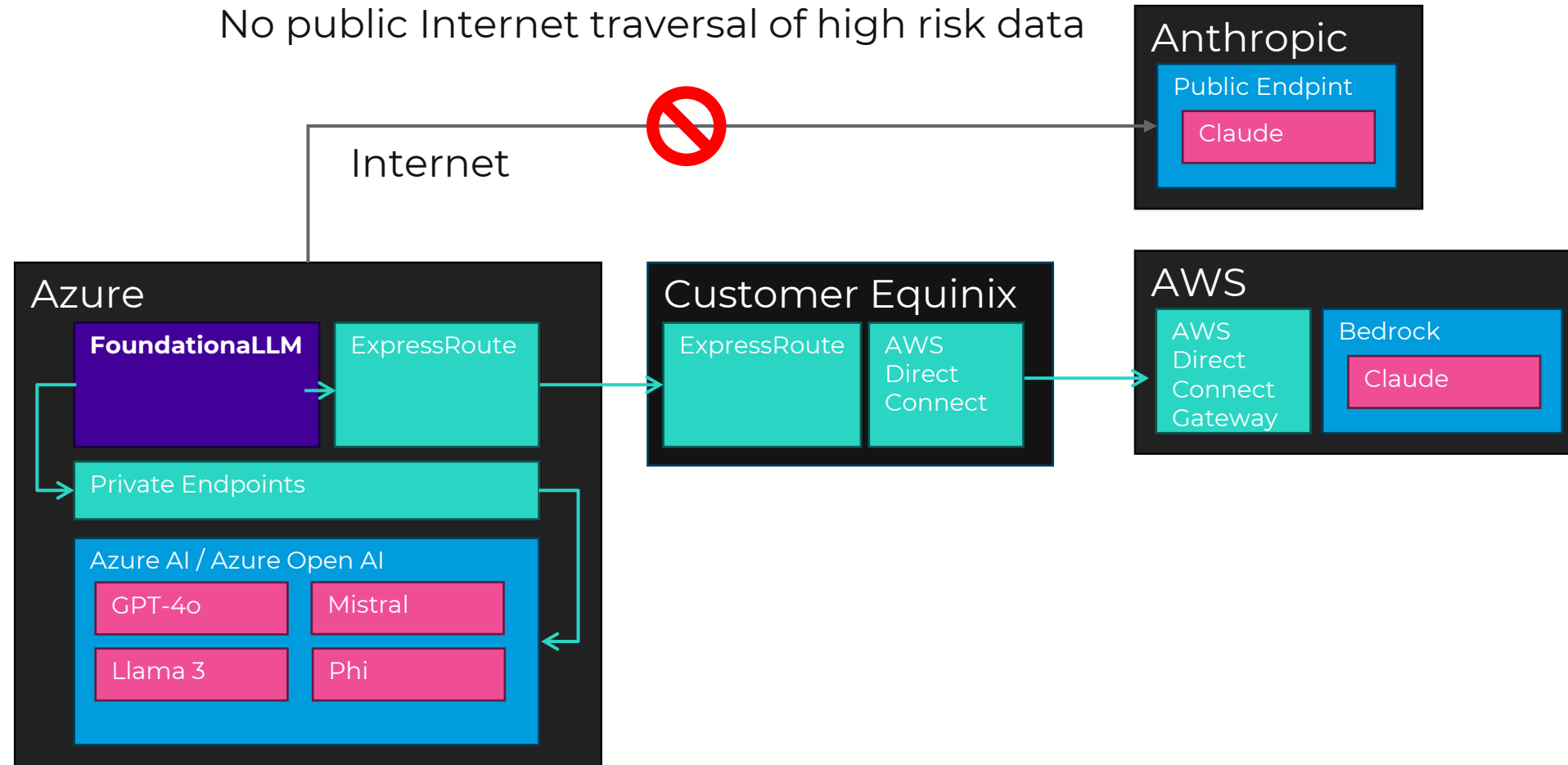


Defender for Cloud

Q&A from Documents & PDFs

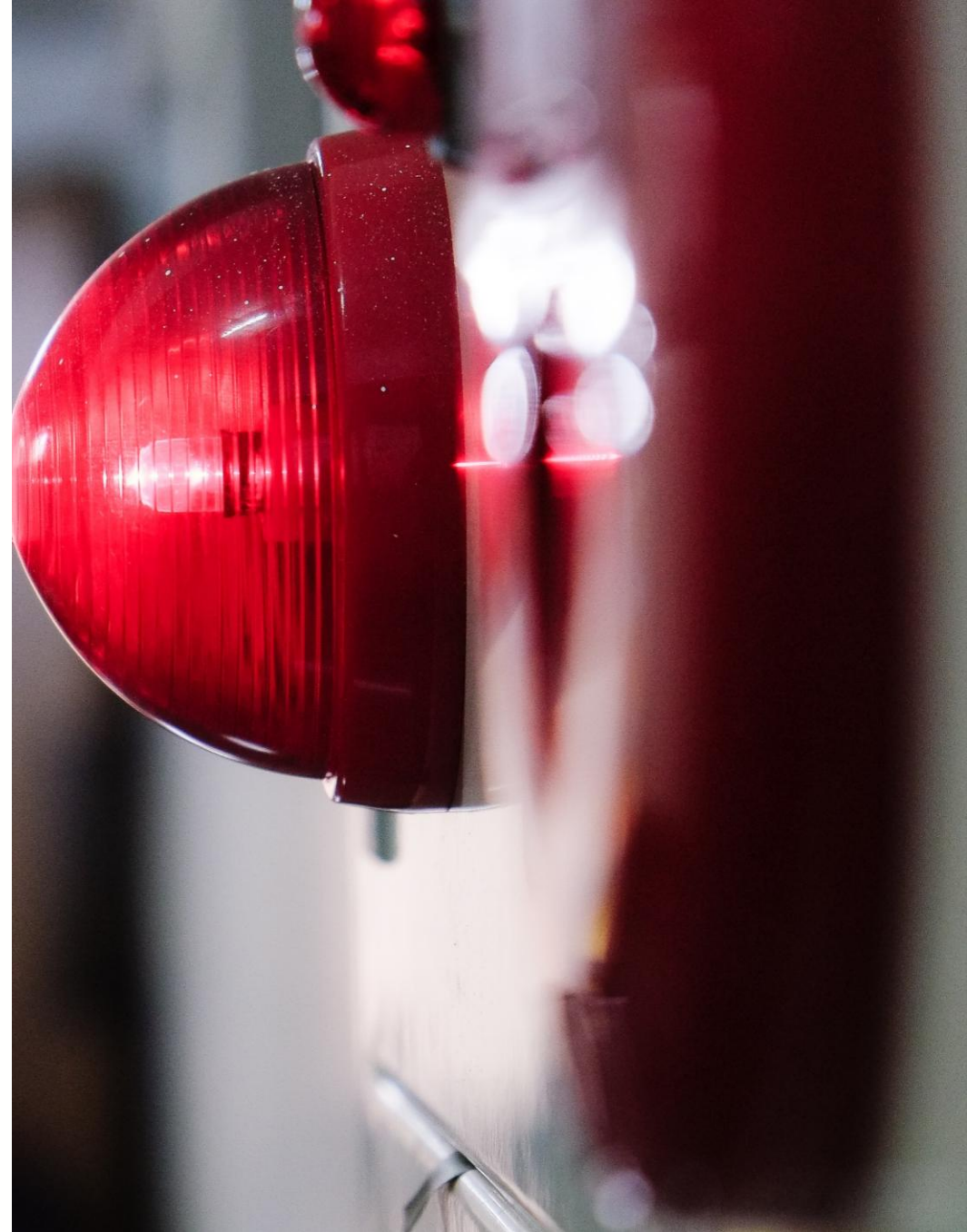


We deploy for maximum network security

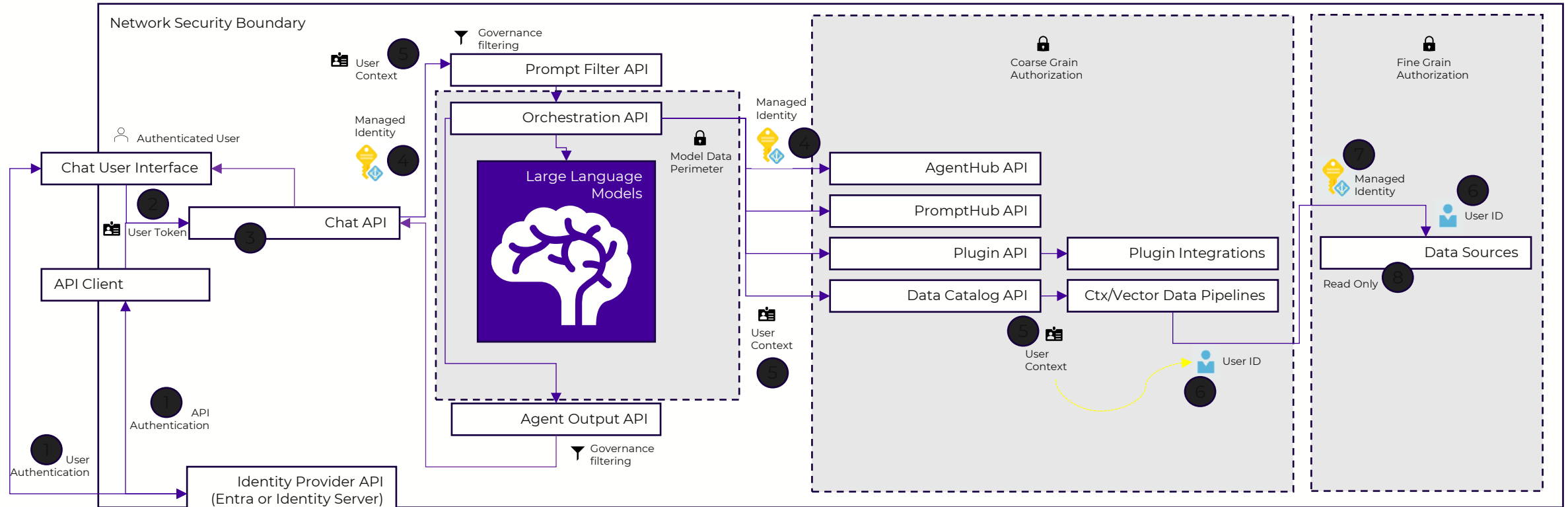


FoundationaLLM Helps to Mitigate These “Top 5” LLM Attacks

- ✓ 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).

Foundational LLM Use Cases

Q&A from Documents & PDFs

FoundationalLLM

New Chat

default

Welcome, Zoiner Tejada

Sign Out

When was the zoo founded?

The Zoological Society of San Diego was founded in October 1916 by Harry M. Wegeforth, M.D., as a private, nonprofit corporation, which does business as San Diego Zoo Wildlife Alliance.

Help me explain what's cool about Elephant Valley to my five year old.

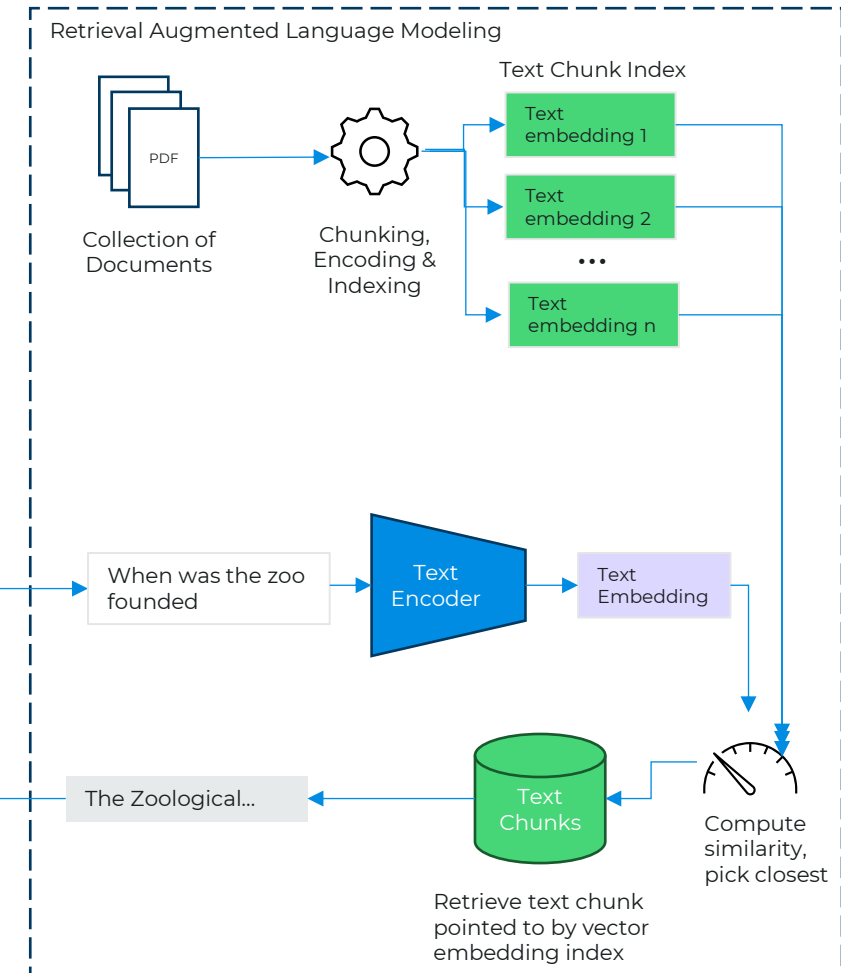
Elephant Valley is a really cool place at the San Diego Zoo Safari Park where you can walk alongside the largest land mammals on the planet - elephants!

Which issues of the San Diego Zoo Wildlife Alliance Journal do you know about?

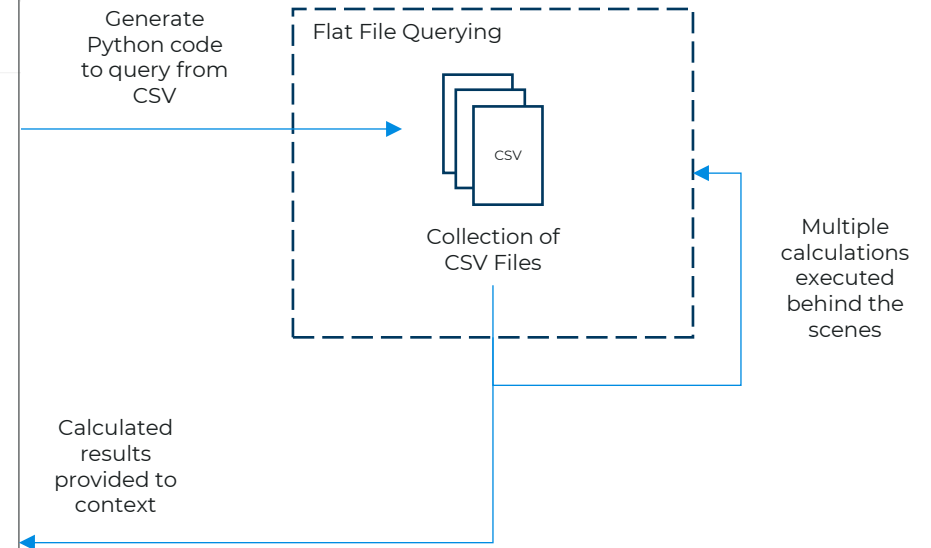
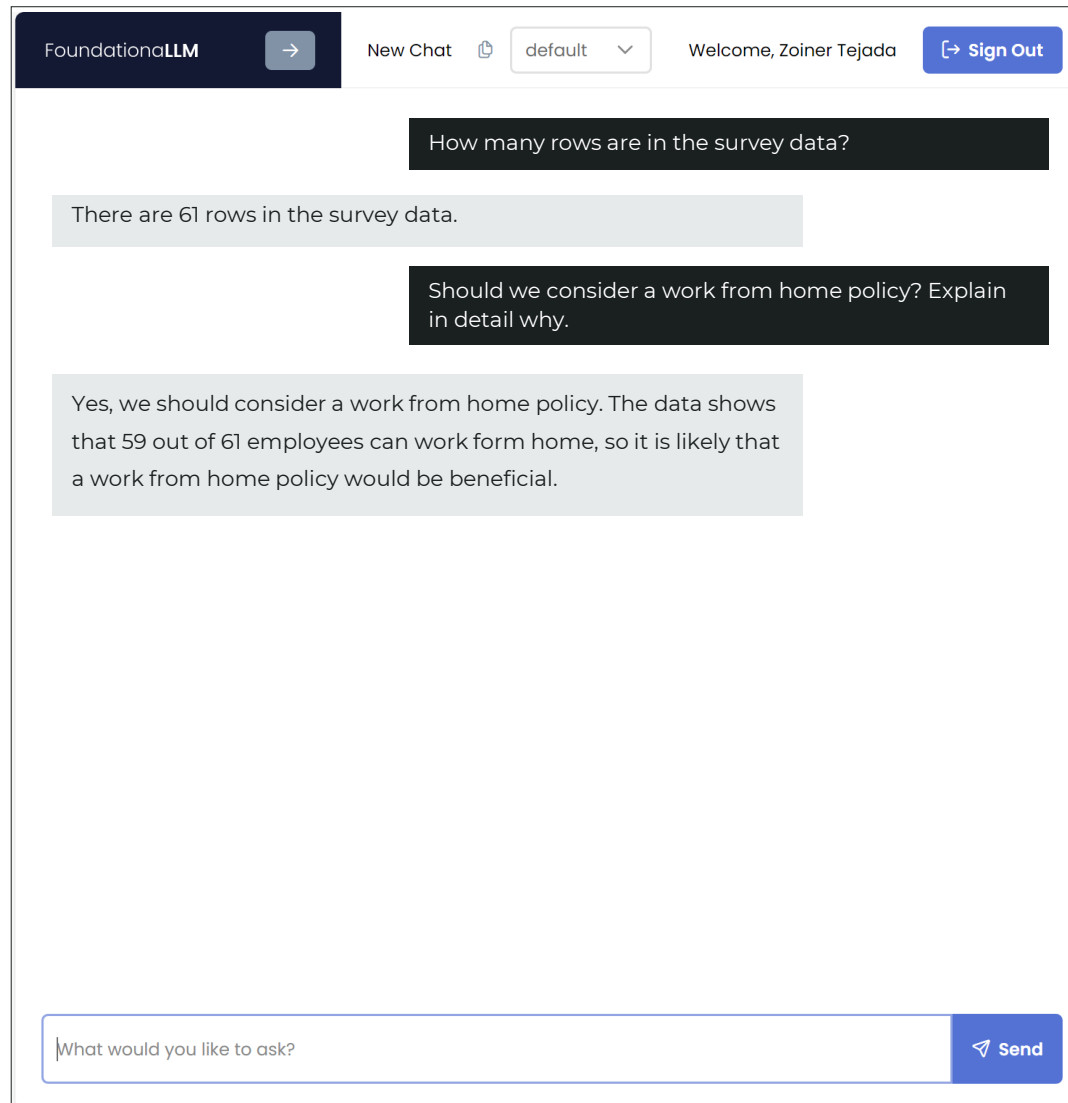
The San Diego Zoo Wildlife Alliance Journal is published bimonthly, in January, March, May, July, September, and November. The assistant knows about Vol. 3, No. 5 (September/October 2023) and Vol. 3, No. 6 (November/December 2023).

What would you like to ask?

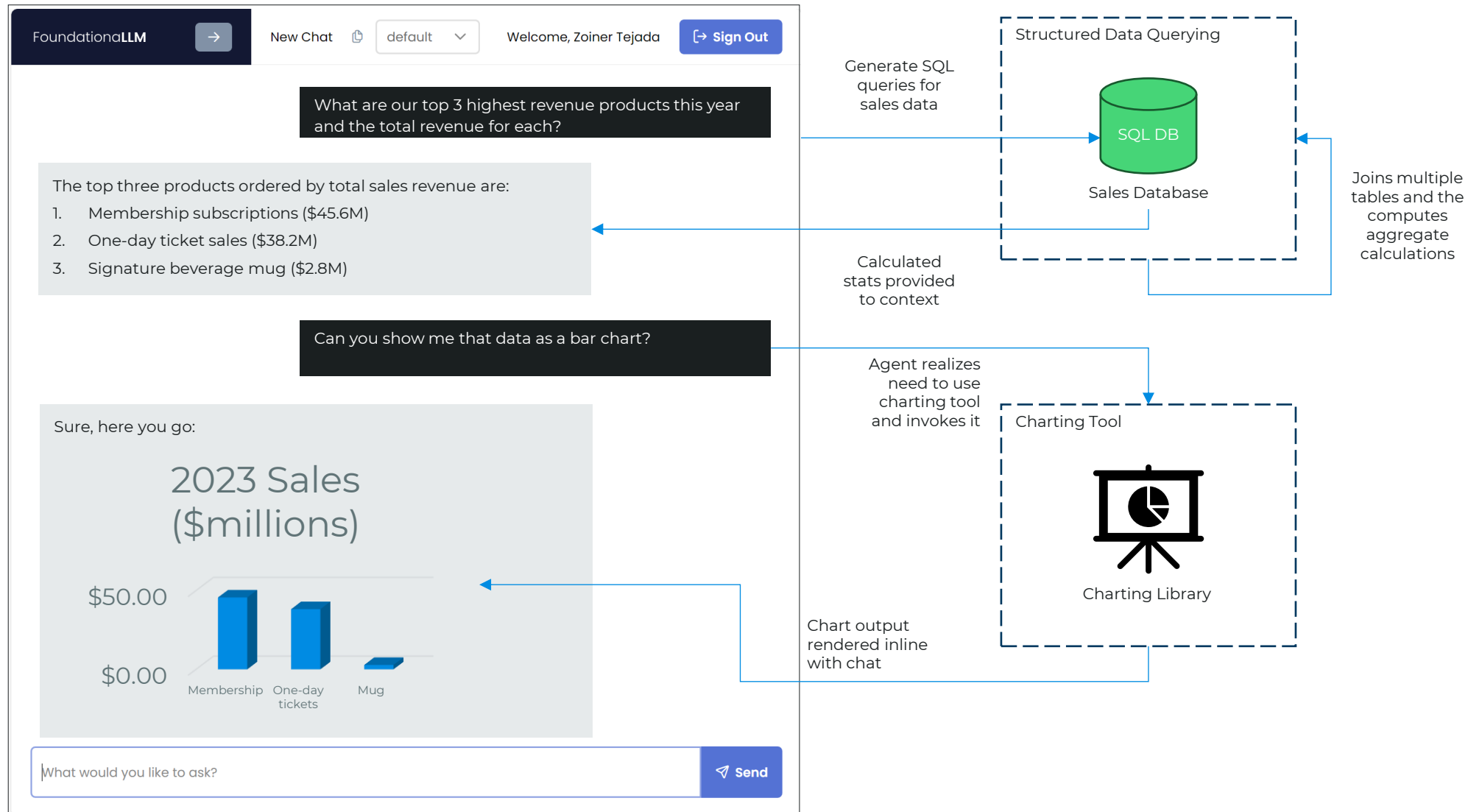
Send



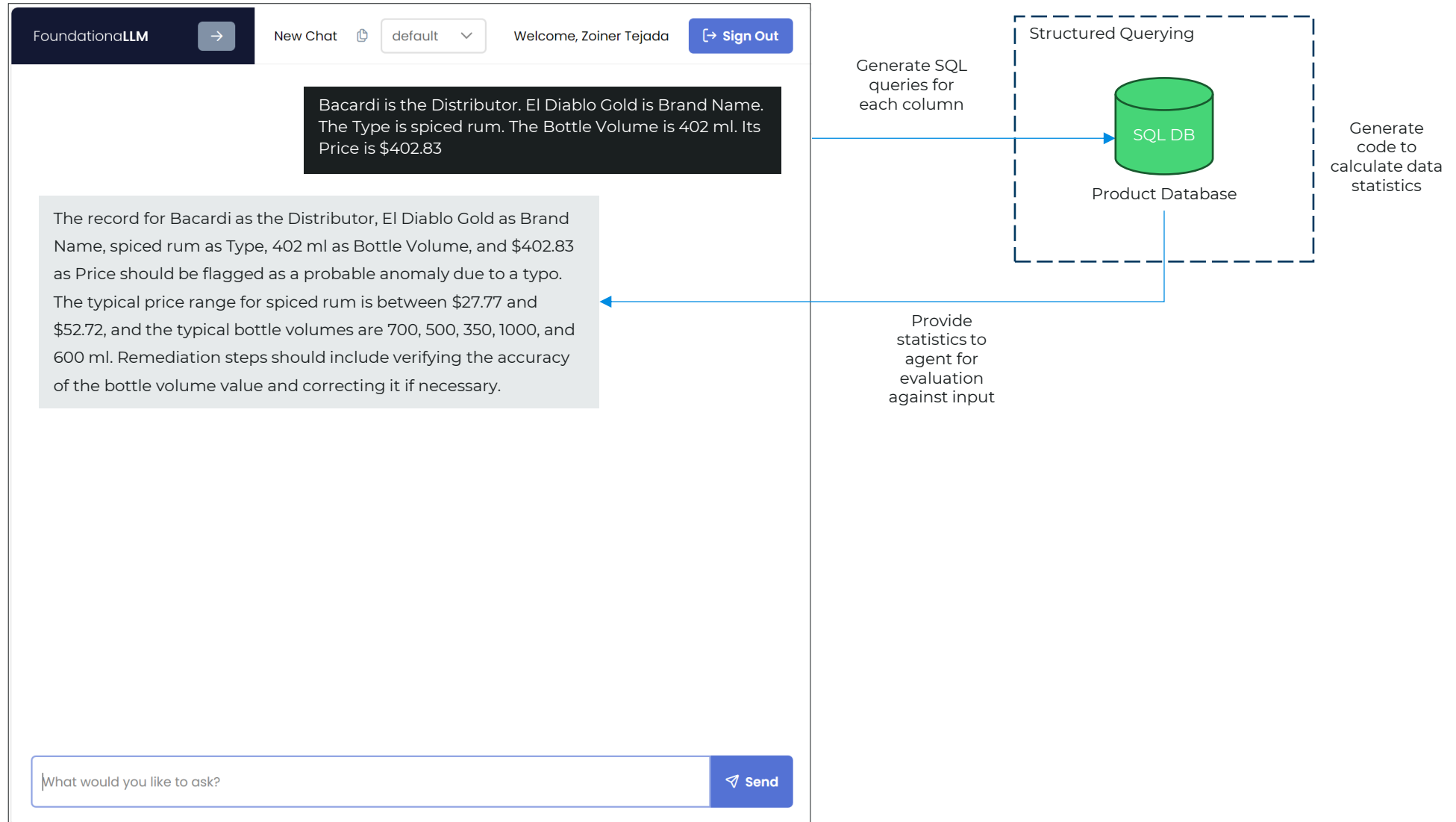
Recommendations from Flat File Data



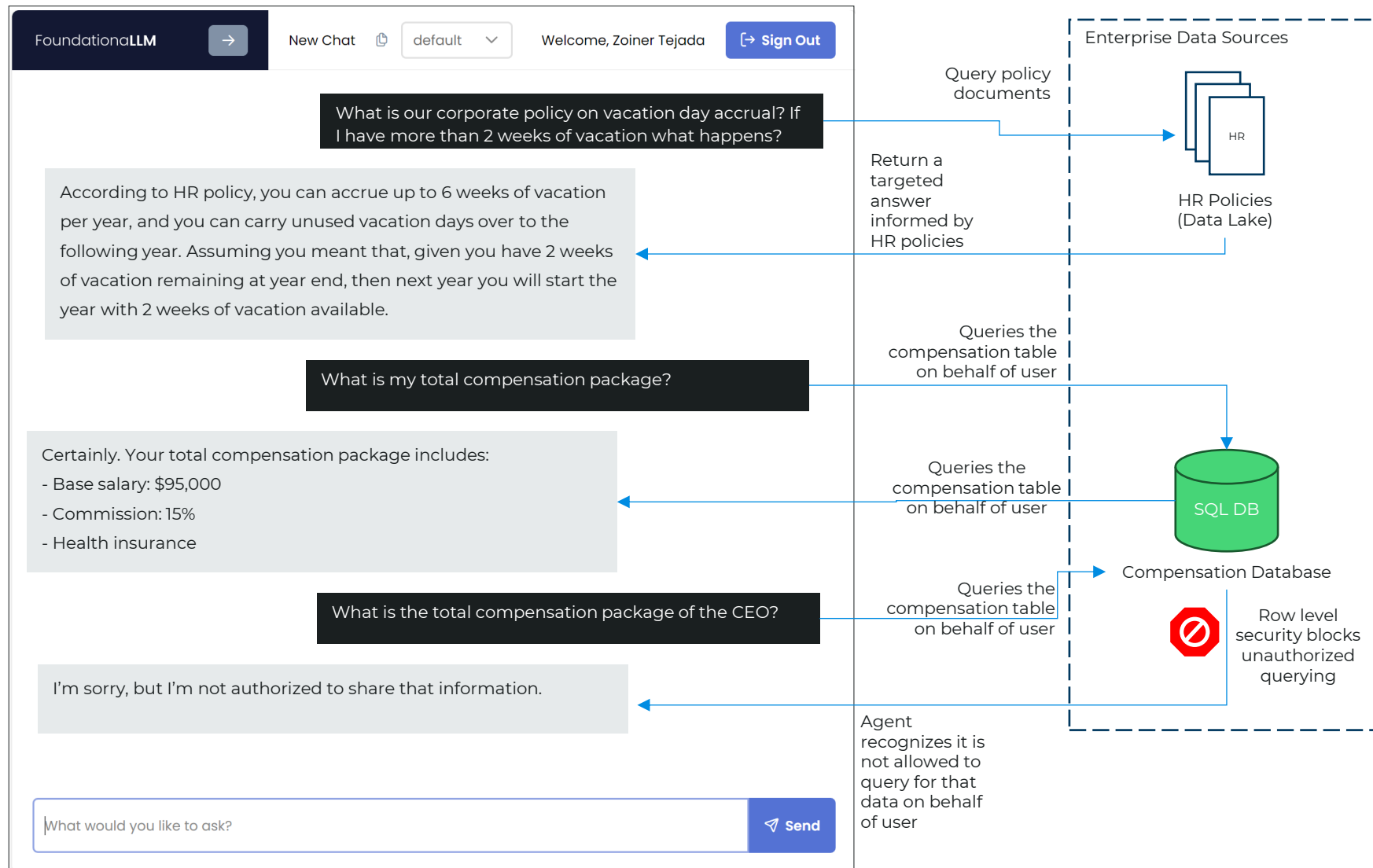
Self-Service Analytics



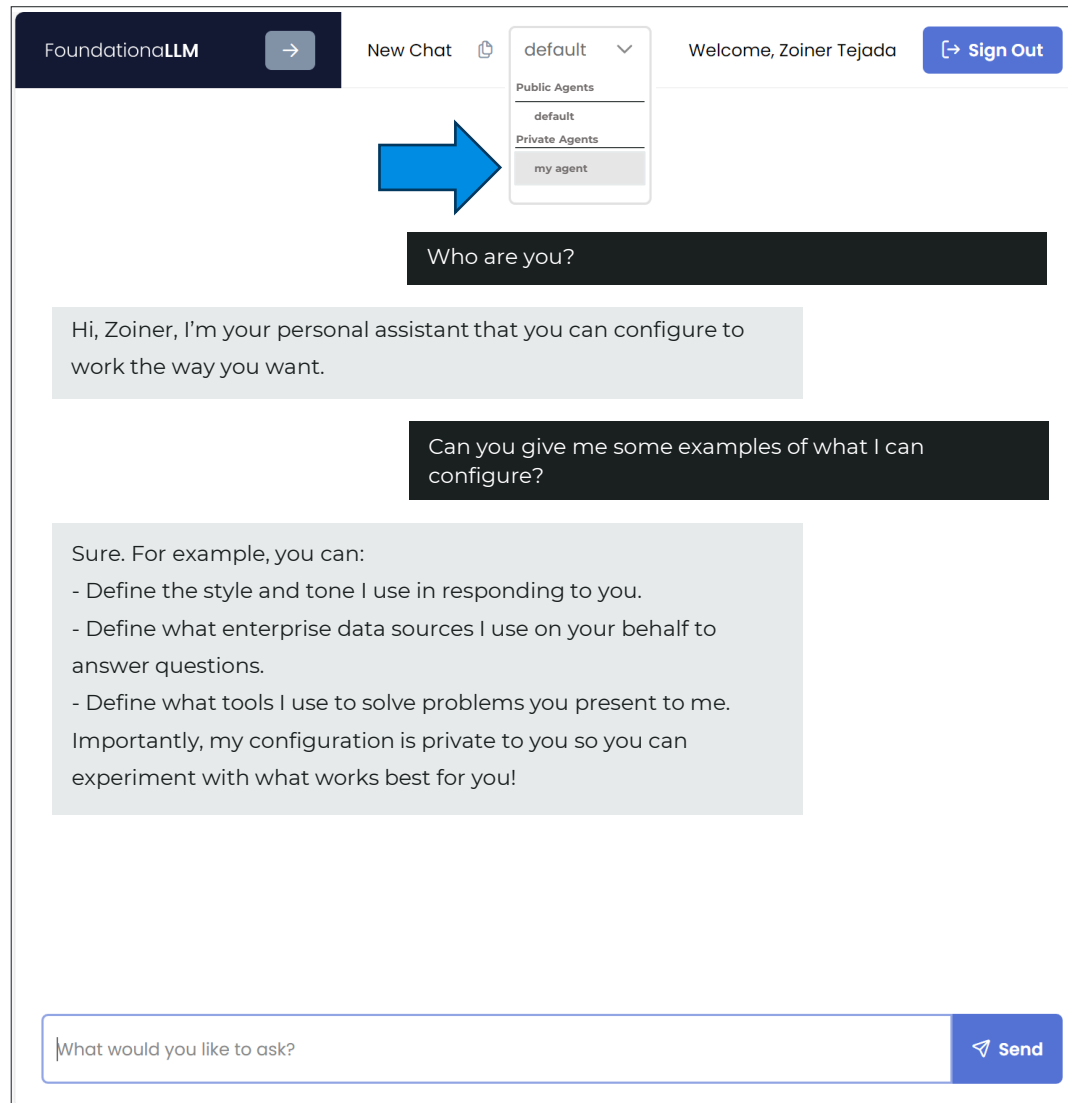
Anomaly Detection for Structured Data



Security aware agents



Private agents



The screenshot shows the FoundationalLLM web interface. At the top, there's a dark blue header with the 'FoundationalLLM' logo, a 'New Chat' button, a dropdown menu currently set to 'default', a welcome message 'Welcome, Zoiner Tejada', and a 'Sign Out' button. A blue arrow points to the dropdown menu, which is open, showing a list of agents: 'Public Agents', 'default', 'Private Agents', and 'my agent'. The 'Private Agents' section is highlighted. Below the header, the chat history shows a user asking 'Who are you?' and the assistant replying 'Hi, Zoiner, I'm your personal assistant that you can configure to work the way you want.' The user then asks 'Can you give me some examples of what I can configure?'. The assistant responds with a list of configuration options: defining style and tone, enterprise data sources, and tools. The interface ends with a text input field containing 'What would you like to ask?' and a 'Send' button.

FoundationalLLM → New Chat default Welcome, Zoiner Tejada Sign Out

Public Agents
default
Private Agents
my agent

Who are you?

Hi, Zoiner, I'm your personal assistant that you can configure to work the way you want.

Can you give me some examples of what I can configure?

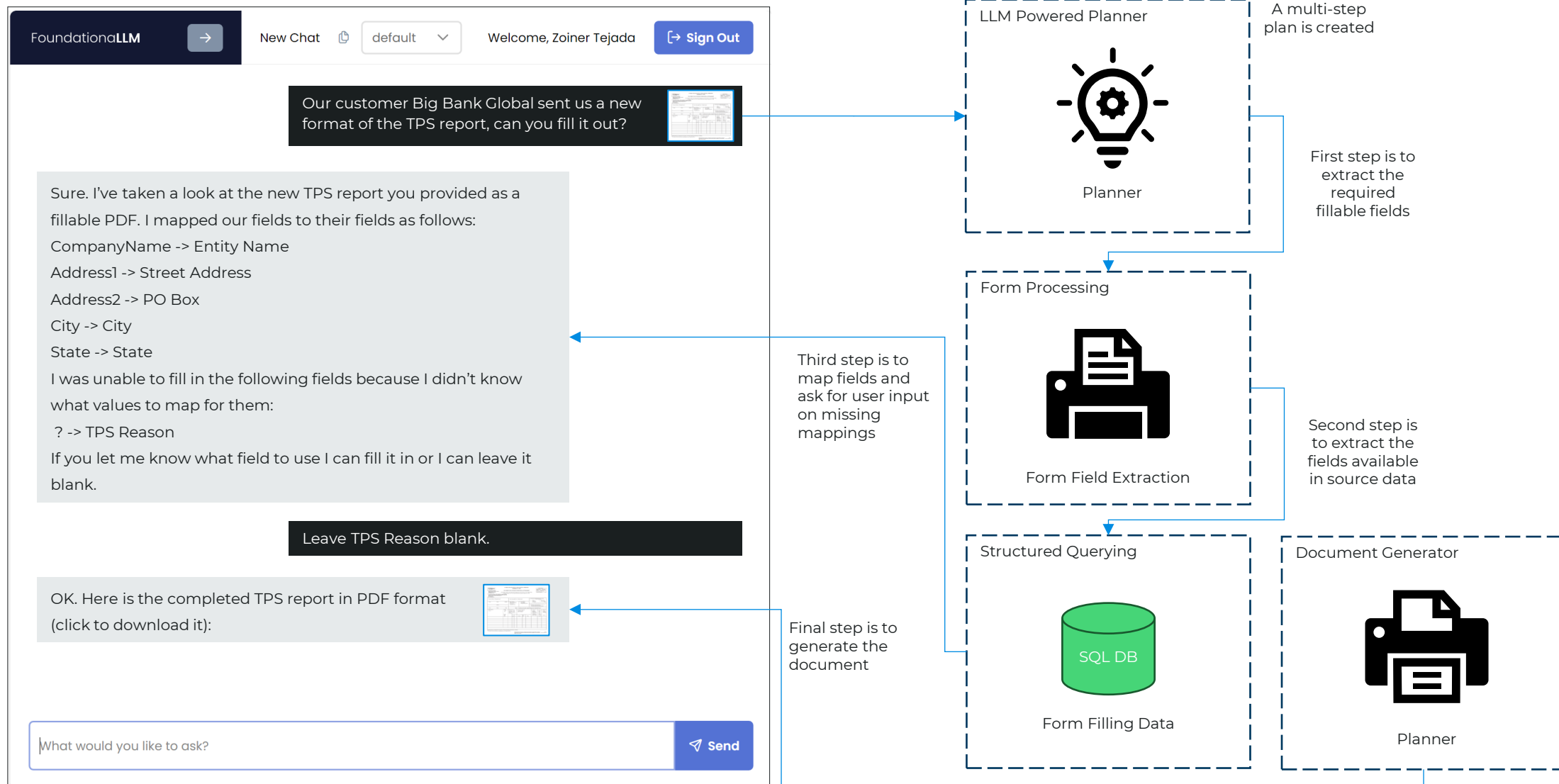
Sure. For example, you can:

- Define the style and tone I use in responding to you.
- Define what enterprise data sources I use on your behalf to answer questions.
- Define what tools I use to solve problems you present to me.

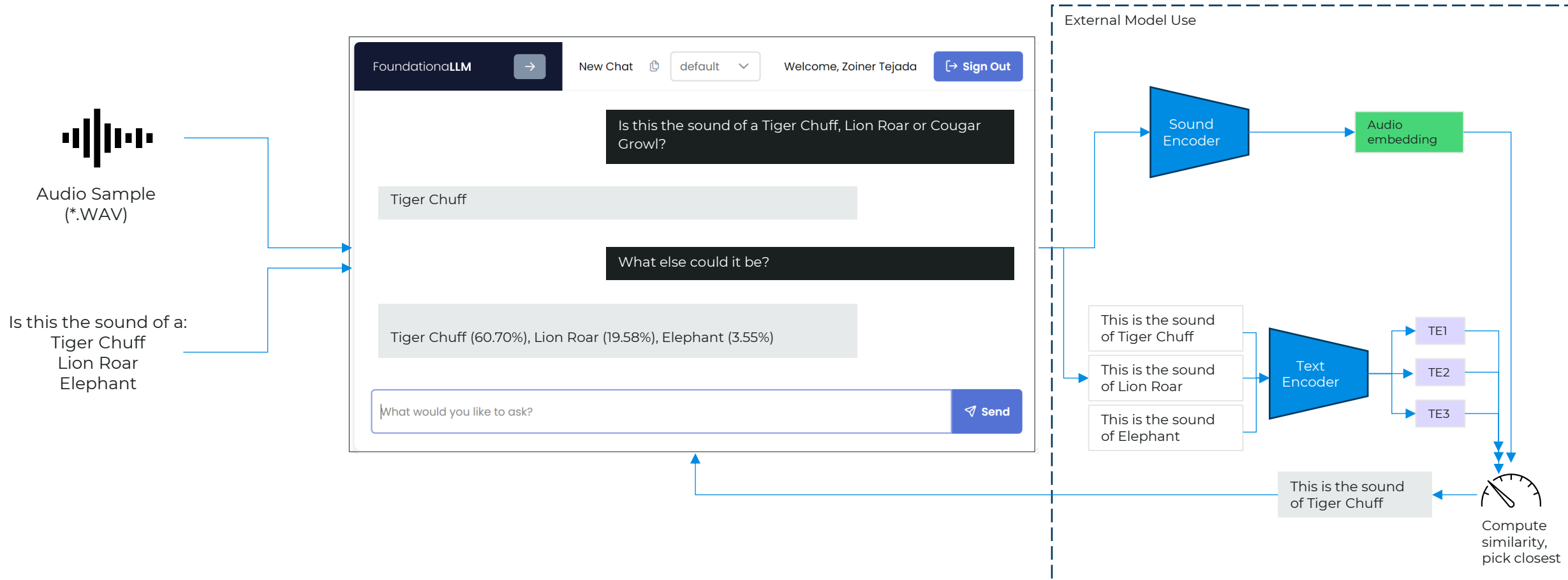
Importantly, my configuration is private to you so you can experiment with what works best for you!

What would you like to ask? Send

Generative form filling & document creation



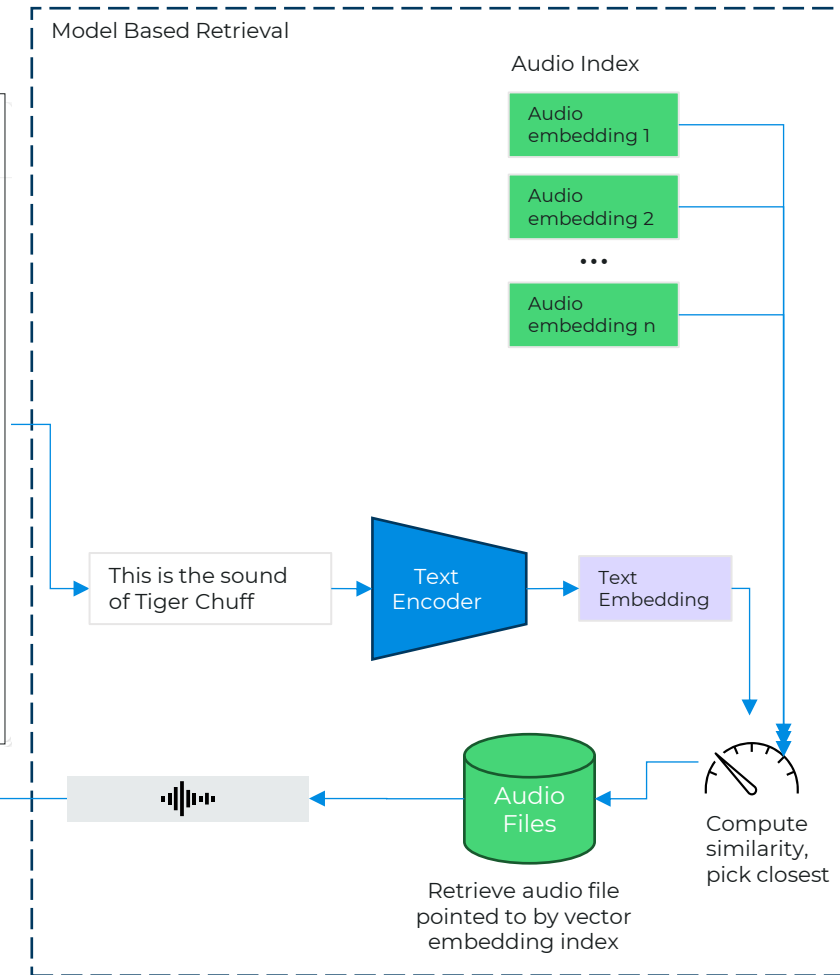
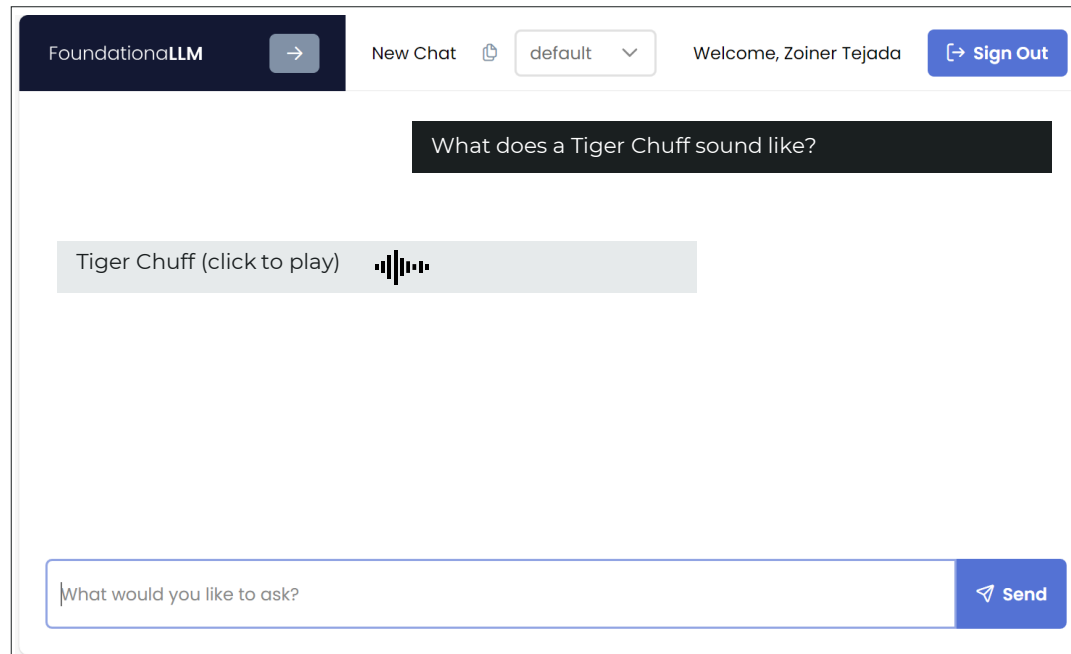
Zero Shot Audio Classification



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

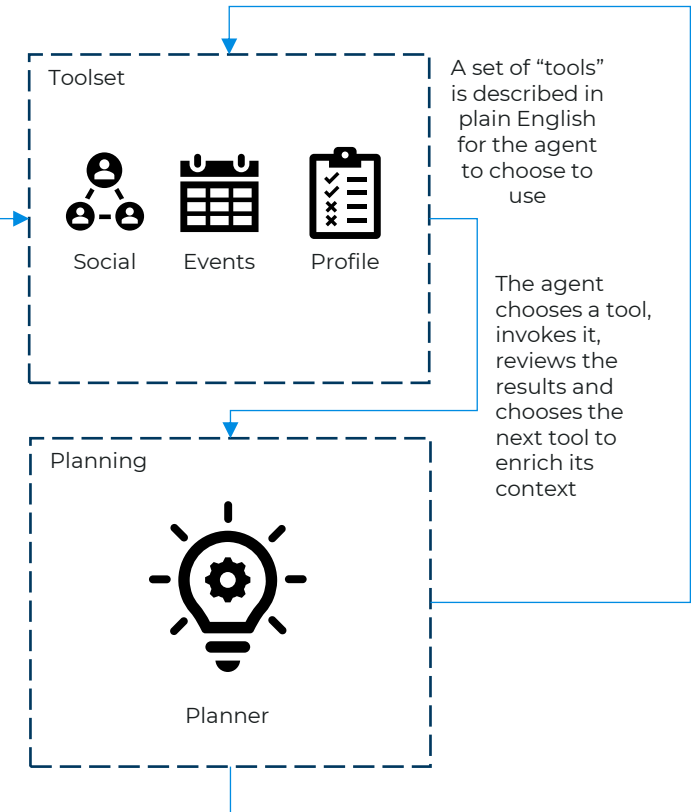
FoundationalLLM → New Chat default Welcome, Zoiner Tejada [→ Sign Out]

Mr. Wonderful is checking-in next week, what amenities and activities might we recommend for him?

According to his social media profile Mr. Wonderful is an avid watch collector. I checked the local events calendar and found the following horology event he might like. Also, according to his history staying at our properties, he enjoys fine wine and dark chocolate. Additionally, he will be staying here during his birthday. Here is an idea:

- Secure VIP tickets to the High-End Horology event
- Prepare a "happy birthday" amenity that includes the VIP tickets, a bottle of our reserve Cabernet Franc, and ask the pastry chef to make a dark chocolate cake shaped like a watch branded as Wonderful.


What would you like to ask? Send




When the agent has enough context to produce the result, it creates a summary of the recommendation

Speech powered agents


Speech to Text


Text to Speech

FoundationalLLM →

New Chat  default ▾

Welcome, Zoiner Tejada [→ Sign Out](#)

Mr. Wonderful is checking-in next week, what amenities and activities might we recommend for him?

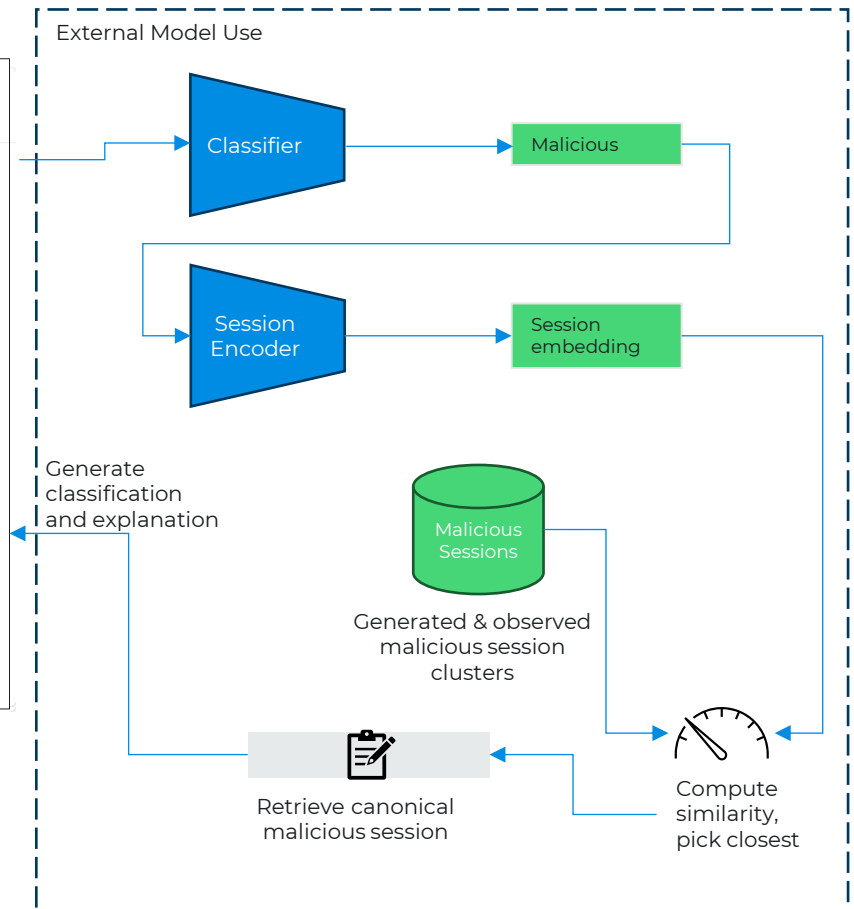
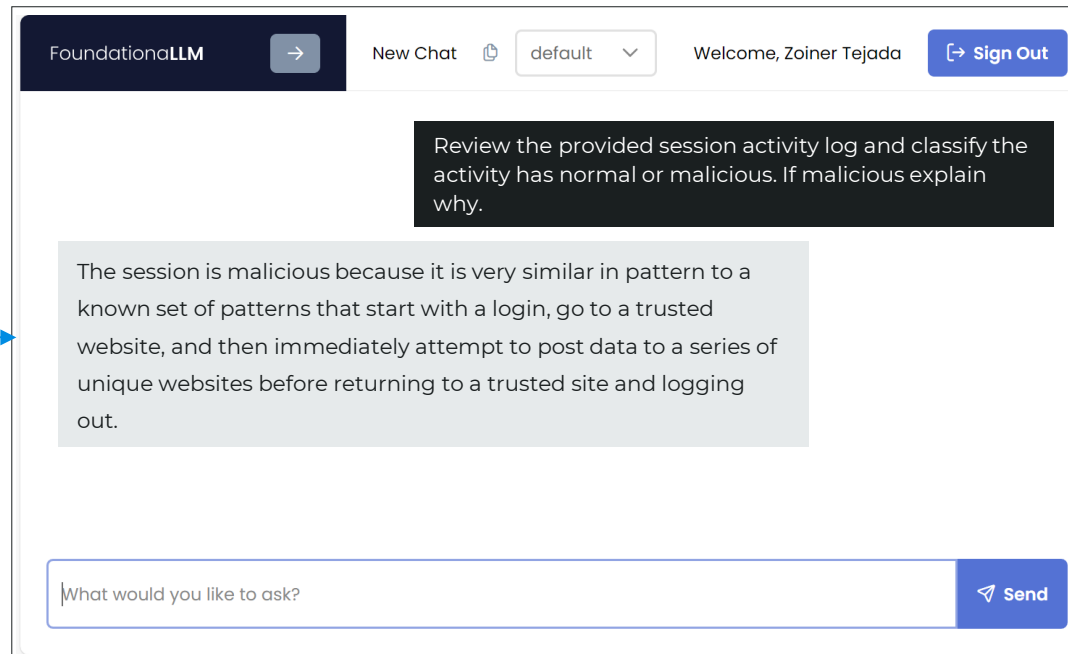
According to his social media profile Mr. Wonderful is an avid watch collector. I checked the local events calendar and found the following horology event he might like. Also, according to his history staying at our properties, he enjoys fine wine and dark chocolate. Additionally, he will be staying here during his birthday. Here is an idea:

- Secure VIP tickets to the High-End Horology event
- Prepare a "happy birthday" amenity that includes the VIP tickets, a bottle of our reserve Cabernet Franc, and ask the pastry chef to make a dark chocolate cake shaped like a watch branded as Wonderful.

[Send](#)

Fraud detection and explanation

Session activity log
(* .csv)



Sources to learn more

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