
Chat with your enterprise data using LLM

This sample demonstrates a few approaches for creating ChatGPT-like experiences over your own data. It uses Azure OpenAI Service to access the ChatGPT model (gpt-35-turbo and gpt3), and vector store (Pinecone, Redis and others) or Azure cognitive search for data indexing and retrieval.

The repo provides a way to upload your own data so it's ready to try end to end.

Updates

- 3/30/2024 - Refactored to keep on Chat, Chat Stream, QnA, Upload and Admin functionality. All others will be moved to it's own repo.
- 3/10/2024 - Move the Prompt Flow version to entaoaipf
- 3/9/2024 - Initial version of advanced RAG techniques and Multi-modal RAG pattern
- 2/28/2024 - Removed SEC analysis features and it's moved into it's own repo at sec
- 1/28/2024 - Remove PitchBook features as they are moved into it's own repo at pib
- 1/19/2024 - Updated the python package & OpenAI > 1.0. Changes made to all Python API for breaking changes introduced in OpenAI and langchain.
- 10/12/2023 - Initial version of Autonomous PromptFlow. For now supporting the Pinecone indexes, but support for Cognitive Search and Redis will be updated soon.
- 9/29/2023 - Added Evaluate PromptFlow. Prompt Flow once created in Azure ML, can be attached to your existing run to evaluate against the following evaluation process :
 - Groundness - The Q&A Groundedness evaluation flow will evaluate the Q&A Retrieval Augmented Generation systems by leveraging the state-of-the-art Large Language Models (LLM) to measure the quality and safety of your responses. Utilizing GPT-3.5 as the Language Model to assist with measurements aims to achieve a high agreement with human evaluations compared to traditional mathematical measurements. `gpt_groundedness` (against context): Measures how grounded the model's predicted answers are against the context. Even if LLM's responses are true, if not verifiable against context, then such responses are considered ungrounded.
 - Ada Similarity - The Q&A `ada_similarity` evaluation flow will evaluate the Q&A Retrieval Augmented Generation systems by leveraging the state-of-the-art Large Language Models (LLM) to measure the quality and safety of your responses. Utilizing GPT-3.5 as the Language Model to assist with measurements aims to achieve a high agreement with human evaluations compared to traditional mathematical measurements. The Ada Similarity evaluation flow allows you to assess and evaluate your model with the LLM-assisted `ada_similarity` metri `ada_similarity`: Measures the cosine similarity of `ada` embeddings of the model prediction and the ground truth. `ada_similarity` is a value in the range [0, 1].

-
- Coherence - The Q&A Coherence evaluation flow will evaluate the Q&A Retrieval Augmented Generation systems by leveraging the state-of-the-art Large Language Models (LLM) to measure the quality and safety of your responses. Utilizing GPT-3.5 as the Language Model to assist with measurements aims to achieve a high agreement with human evaluations compared to traditional mathematical measurements. The Coherence evaluation flow allows you to assess and evaluate your model with the LLM-assisted Coherence metric. `gpt_coherence`: Measures the quality of all sentences in a model's predicted answer and how they fit together naturally. Coherence is scored on a scale of 1 to 5, with 1 being the worst and 5 being the best.
 - Similarity - The Q&A Similarity evaluation flow will evaluate the Q&A Retrieval Augmented Generation systems by leveraging the state-of-the-art Large Language Models (LLM) to measure the quality and safety of your responses. Utilizing GPT-3.5 as the Language Model to assist with measurements aims to achieve a high agreement with human evaluations compared to traditional mathematical measurements. The Similarity evaluation flow allows you to assess and evaluate your model with the LLM-assisted Similarity metric. `gpt_similarity`: Measures similarity between user-provided ground truth answers and the model predicted answer. Similarity is scored on a scale of 1 to 5, with 1 being the worst and 5 being the best.
 - F1 Score - The Q&A f1-score evaluation flow will evaluate the Q&A Retrieval Augmented Generation systems using f1-score based on the word counts in predicted answer and ground truth. The f1-score evaluation flow allows you to determine the f1-score metric using number of common tokens between the normalized version of the ground truth and the predicted answer. F1-score: Compute the f1-Score based on the tokens in the predicted answer and the ground truth. F1-score is a value in the range [0, 1]. Groundedness metric is scored on a scale of 1 to 5, with 1 being the worst and 5 being the best.
- 9/22/2023 - Added PromptFlow for SqlAsk. Ensure `PFSQLASK_URL` and `PFSQLASK_KEY` configuration values are added to deployed endpoint to enable the feature. Also make sure `SynapseName`, `SynapsePool`, `SynapseUser` and `SynapsePassword` configuration values are added to `entaoui` PromptFlow connection. Moved deleting the Session Capability for ChatGpt to Admin Page.
 - 9/20/2023 - Added configuration to allow end user to change the Search Type for Cognitive Search Vector Store index (Hybrid, Similarity/Vector and Hybrid Re-rank), based on the Best Practices we shared. QnA, Chat and Prompt Flow are modified. QnA and Chat are implementing the customized Vector store implementation of Langchain and Prompt Flow using the helper functions. Fixed the issue with QnA/Chat/PromptFlow not generating followup-questions.
 - 9/18/2023 - Refactored SQL NLP to not use Langchain Database Agent/Chain and instead use custom Prompts.
-

-
- 9/15/2023 - Modified the azure search package to 11.4.0b9 and langchain to latest version. Added capability to perform evaluation on PromptFlow for both QnA and Chat. Bert PDF and Evaluation Data can be used to perform Batch and Evaluation in Prompt Flow. Sample Notebook showcasing the flow and E2E process is available. Bert Chat folder allows you to test E2E Prompt Flow, Batch Run and Evaluation in form of Notebook.
 - 9/3/2023 - Added API for Chat using the Prompt Flow. Allow end-user to select between Azure Functions as API ([ApiType](#) Configuration in Web App) or using Prompt Flow Managed endpoint.
 - 9/2/2023 - Added API for Question Answering using the Prompt Flow. Allow end-user to select between Azure Functions as API ([ApiType](#) Configuration in Web App) or using Prompt Flow Managed endpoint.
 - 8/31/2023 - Added example for LLMops using Prompt Flow. The repo will be adding the flexibility to use the Prompt Flow Deployed Model as an alternative to current Azure Functions.
 - 8/20/2023 - Added support for the Markdown files (as zip file) and removed the chunk_size=1 from Azure OpenAiEmbedding
 - 8/11/2023 - Fixed the issue with Streaming Chat feature.
 - 8/10/2023 - **Breaking Changes** - Refactored all code to use [OpenAiEndPoint](#) configuration value instead of [OpenAiService](#). It is to support the best practices as they are outlined in Enterprise Logging via Azure API Management. Your [OpenAiEndPoint](#) if using APIM will be API Gateway URL and the [OpenAiKey](#) will be the Product/Unlimited key. If not using APIM, you don't need to change the key, but ensure [OpenAiEndPoint](#) is fully qualified URL of your AOAI deployment. [OpenAiService](#) is no longer used. Changes did impact the working on Chat on Stream feature, so it's disabled for now and will be enabled once tested and fixed.
 - 8/9/2023 - Added Function calling in the ChatGpt interface as checkbox. Sample demonstrate ability to call functions. Currently Weather API, Stock API and Bing Search is supported. Function calling is in preview and supported only from "API Version" of "2023-07-01-preview", so make sure you update existing deployment to use that version. Details on calling Functions. For existing deployment add [WeatherEndPoint](#), [WeatherHost](#), [StockEndPoint](#), [StockHost](#) and [RapidApiKey](#) configuration to Azure Function App.
 - 8/5/2023 - Added Chat Interface with "Stream" Option. This feature allows you to stream the conversation to the client. You will need to add [OpenAiChat](#), [OpenAiChat16k](#), [OpenAiEmbedding](#), [OpenAiEndPoint](#), [OpenAiKey](#), [OpenAiApiKey](#), [OpenAiService](#), [OpenAiVersion](#), [PineconeEnv](#), [PineconeIndex](#), [PineconeKey](#), [RedisAddress](#), [RedisPassword](#), [RedisPort](#) property in Azure App Service (Webapp) to enable the feature for existing deployment.
 - 7/30/2023 - Removed unused Code - SummaryAndQa and Chat
 - 7/28/2023 - Started removing the Davinci model usage. For now removed the usage from all functionality except workshop. Refactored Summarization functionality based on the feedback

-
- to allow user to specify the prompt and pre-defined Topics to summarize it on.
- 7/26/2023 - Remove OpenAI Playground from Developer Tools as advanced features of that are available in ChatGPT section.
 - 7/25/2023 - Add tab for the Chat capabilities to support ChatGpt capability directly from the model instead of “Chat on Data”. You will need to add `CHATGPT_URL` property in Azure App Service (Webapp) to enable the feature outside of deploying the new Azure Function.
 - 7/23/2023 - Added the rest of the feature for PIB UI and initial version of generating the Power-Point deck as the output. For new feature added ensure you add `FMPKEY` variable to webapp configuration.
 - 7/20/2023 - Added feature to talk to Pib Data (Sec Filings & Earning Call Transcript). Because new Azure function is deployed, ensure `PIBCHAT_URL` property is added to Azure WebApp with the URL for your deployed Azure Functions
 - 7/18/2023 - Refactored the PIB code to solve some of the performance issue and bug fixes.
 - 7/17/2023 - Removed GPT3 chat interface with retirement of “Davinci” models.
 - 7/16/2023 - Initial version of Pib UI (currently supporting 5 Steps - Company Profile, Call Transcripts, Press Releases, Sec Filings and Ratings/Recommendations). You will need access to Paid subscription (FMP or modify based on what your enterprise have access to). To use with FMP you will need to add `FmpKey` in Azure Functions. Because of circular dependency you need to manually add `SecDocPersistUrl` and `SecExtractionUrl` manually in Azure Functions.
 - 7/14/2023 - Add support for GPT3.5 16K model and ability to chunk document > 4000 tokens with > 500 overlap. For the ChunkSize > 4000, it will default to 16K token for both QnA and Chat functionality. Added identity provider to the application and authentication for QnA and Chat interface. For GPT3.5 16k model, you will need to add `OpenAiChat16k` property in Azure Function app.
 - 7/13/2023 - Allow end user to select ChunkSize and ChunkOverlap Configuration. Initial version of overriding prompt template.
 - 7/11/2023 - Functional PIB CoPilot in the form of the notebook.
 - 7/8/2023 - Added the feature to Rename the session for ChatGPT. Also added the UI for the Evaluator Tool. This feature focuses on performing the LLM based evaluation on your document. It auto-generates the test dataset (with Question and Answers) and perform the grading on that document using different parameters and generates the evaluation results. It is built on Azure Durable Functions and is implemented using the Function Chaining pattern. You will need to add `BLOB_EVALUATOR_CONTAINER_NAME` (ensure the same container name is created in storage account) and `RUNEVALUATION_URL` (URL of the Durable function deployment) configuration in Azure Web App for existing deployment and if you want to use the Evaluator feature. In the Azure function deployment add `AzureWebJobsFeatureFlags` (value `EnableWorkerIndexing`) and `OpenAiEvaluatorContainer` settings.
-

-
- 7/5/2023 - Added the feature to Delete the session. That feature requires the feature that is in preview and you will need to enable that on the CosmosDB account on your subscription. Added simple try/catch block in case if you have not enabled/deployed the CosmosDB to continue chatGPT implementation.
 - 7/4/2023 - Initial version of storing “Sessions” for GPT3.5/ChatGpt interface. Session and messages are stored/retrieved from CosmosDb. Make sure you have CosmosDb service provisioned or create a new one (for existing deployment). You will need to add [CosmosEndpoint](#), [CosmosKey](#), [CosmosDatabase](#) and [CosmosContainer](#) settings in both Azure Functions App and Web App.
 - 6/25/2023 - Notebook showcasing the evaluation of the answer quality in systematic way (auto generating questions and evaluation chain), supporting LLM QA settings (chunk size, overlap, embedding technique). Refer to Evaluator notebook for more information.
 - 6/18/2023 - Add the admin page supporting Knowledge base management.
 - 6/17/2023 - Added “Question List” button for Ask a question feature to display the list of all the questions that are in the Knowledge base. Following three properties [SEARCHSERVICE](#), [SEARCHKEY](#) and [KBINDEXNAME](#) (default value of aoaikb) needs to be added to Azure App Service to enable “Question List” button feature.
 - 6/16/2023 - Add the feature to use Azure Cognitive Search as Vector store for storing the cached Knowledge base. The questions that are not in KB are sent to LLM model to find the answer via OAI, or else it is responded back from the Cached Datastore. New Property [KbIndexName](#) needs to be added to Azure Function app. Added the Notebook to test out the feature as part of the workshop. TODO : Add the feature to add the question to KB from the chat interface (and make it session based). A feature further to “regenerate” answer from LLM (instead of cached answer) will be added soon.
 - 6/7/2023 - Add OpenAI Playground in Developer Tools and initial version of building the CoPilot (for now with Notebook, but eventually will be moved as CoPilot feature). Add the script, recording and example for Real-time Speech analytics use-case. More to be added soon.
 - 5/27/2023 - Add Workshop content in the form of the notebooks that can be leveraged to learn/execute the scenarios. You can find the notebooks in the Workshop folder. Details about workshop content is available [here](#).
 - 5/26/2023 - Add Summarization feature to summarize the document either using stuff, mapreduce or refine summarization. To use this feature (on existing deployment) ensure you add the [OpenAiSummaryContainer](#) configuration to Function app and [BLOB_SUMMARY_CONTAINER_NAME](#) configuration to Azure App Service (Ensure that the value you enter is the same as the container name in Azure storage and that you have created the container). You also need to add [PROCESSSUMMARY_URL](#) configuration to Azure

App Service (Ensure that the value you enter is the same as the Azure Function URL).

- 5/24/2023 - Add feature to upload CSV files and CSV Agent to answer/chat questions on the tabular data. Smart Agent also supports answering questions on CSV data.
- 5/22/2023 - Initial version of “Smart Agent” that gives you flexibility to talk to all documents uploaded in the solution. It also allow you to talk to SQL Database Scenario. As more features are added, agent will keep on building upon that (for instance talk to CSV/Excel or Tabular data)
- 5/21/2023 - Add Developer Tools section - Experimental code conversion and Prompt guru.
- 5/17/2023 - Change the edgar source to Cognitive search vector store instead of Redis.
- 5/15/2023 - Add the option to use “Cognitive Search” as Vector store for storing the index. Azure Cognitive Search offers pure vector search and hybrid retrieval – as well as a sophisticated re-ranking system powered by Bing in a single integrated solution. Sign-up. Support uploading WORD documents.
- 5/10/2023 - Add the options on how document should be chunked. If you want to use the Form Recognizer, ensure the Form recognizer resource is created and the appropriate application settings [FormRecognizerKey](#) and [FormRecognizerEndPoint](#) are configured.
- 5/07/2023 - Option available to select either Azure OpenAI or OpenAI. For OpenAI ensure you have [OpenAiApiKey](#) in Azure Functions settings. For Azure OpenAI you will need [OpenAiKey](#) , [OpenAiService](#) and [OpenAiEndPoint](#) Endpoint settings. You can also select that option for Chat/Question/SQL Nlp/Speech Analytics and other features (from developer settings page).
- 5/03/2023 - Password required for Upload and introduced Admin page starting with Index Management
- 4/30/2023 - Initial version of Task Agent Feature added. Autonomous Agents are agents that designed to be more long running. You give them one or multiple long term goals, and they independently execute towards those goals. The applications combine tool usage and long term memory. Initial feature implements Baby AGI with execution tools
- 4/29/2023 - AWS S3 Process Integration using S3, AWS Lambda Function and Azure Data Factory (automated deployment not available yet, scripts are available in /Deployment/aws folder)
- 4/28/2023 - Fix Bugs, Citations & Follow-up questions across QA & Chat. Prompt bit more restrictive to limit responding from the document.
- 4/25/2023 - Initial version of Power Virtual Agent
- 4/21/2023 - Add SQL Query & SQL Data tab to SQL NLP and fix Citations & Follow-up questions for Chat & Ask features
- 4/17/2023 - Real-time Speech Analytics and Speech to Text and Text to Speech for Chat & Ask Features. (You can configure Text to Speech feature from the Developer settings. You will need Azure Speech Services)
- 4/13/2023 - Add new feature to support asking questions on multiple document using Vector QA Agent
- 4/8/2023 - Ask your SQL - Using SQL Database Agent or Using SQL Database Chain

-
- 3/29/2023 - Automated Deployment script
 - 3/23/2023 - Add Cognitive Search as option to store documents
 - 3/19/2023 - Add GPT3 Chat Implementation
 - 3/18/2023 - API to generate summary on documents & Sample QA
 - 3/17/2023
 - Support uploading Multiple documents
 - Bug fix - Redis Vectorstore Implementation
 - 3/16/2023 - Initial Release, Ask your Data and Chat with your Data

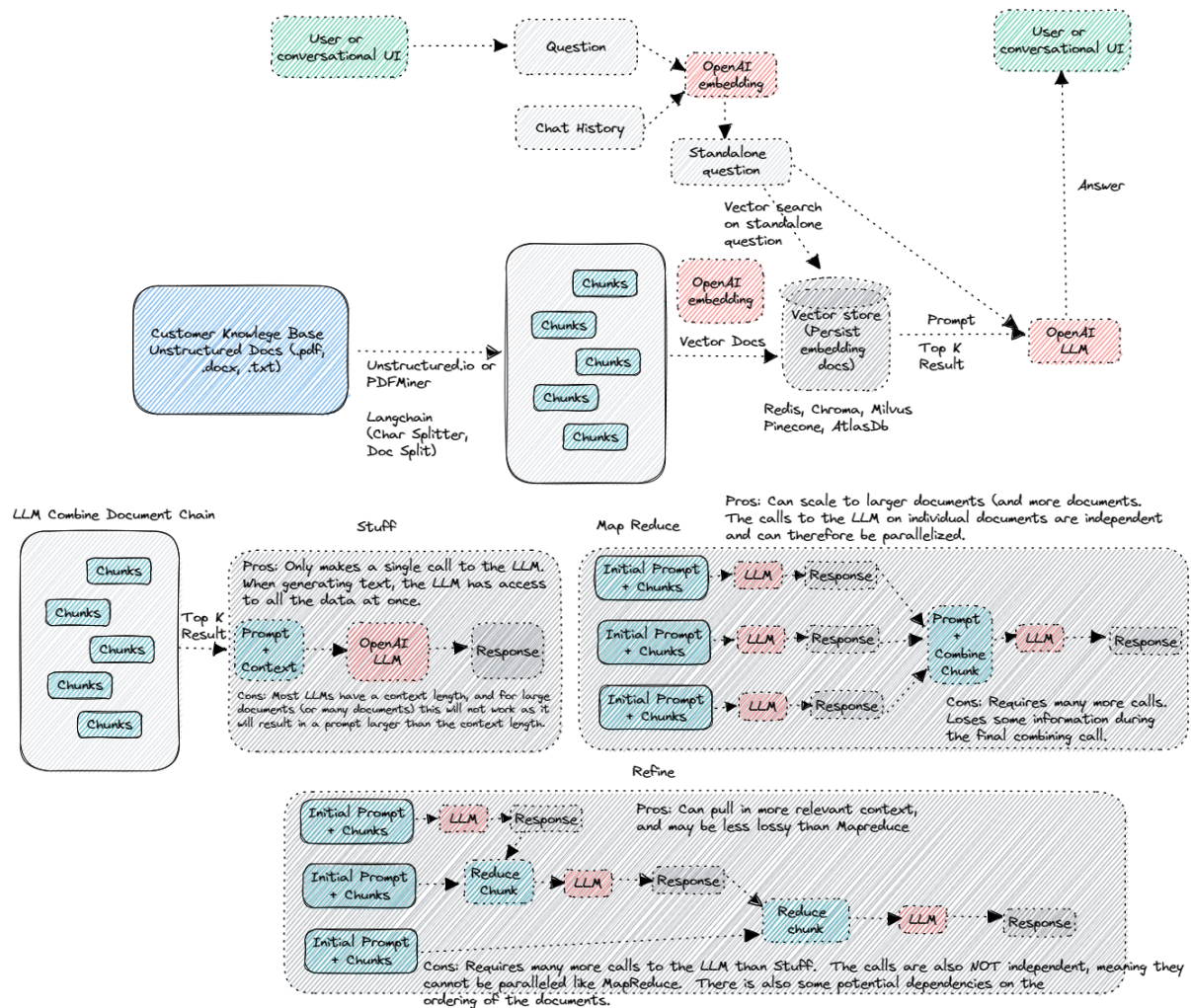
Test Website

Chat and Ask over your data

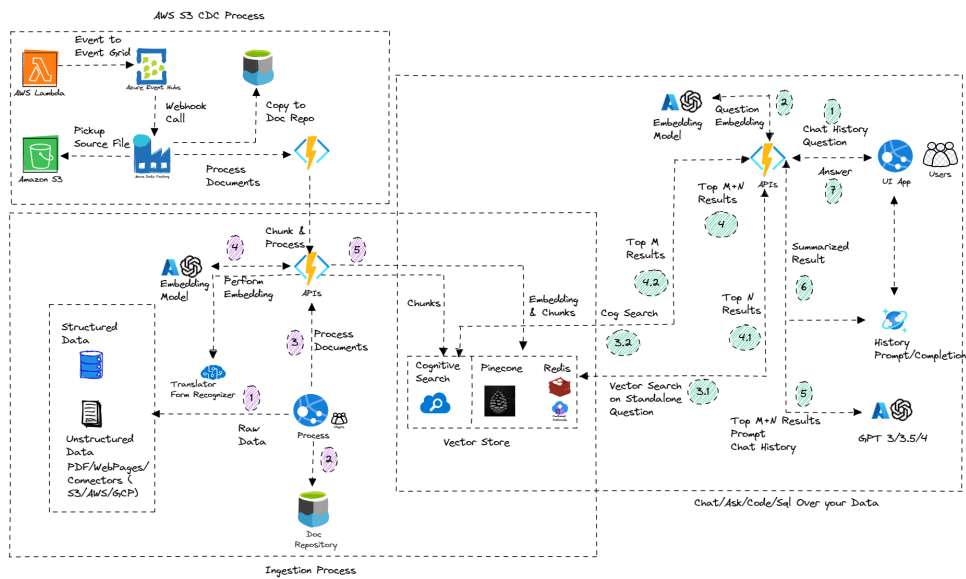
Features

List of Features

Architecture

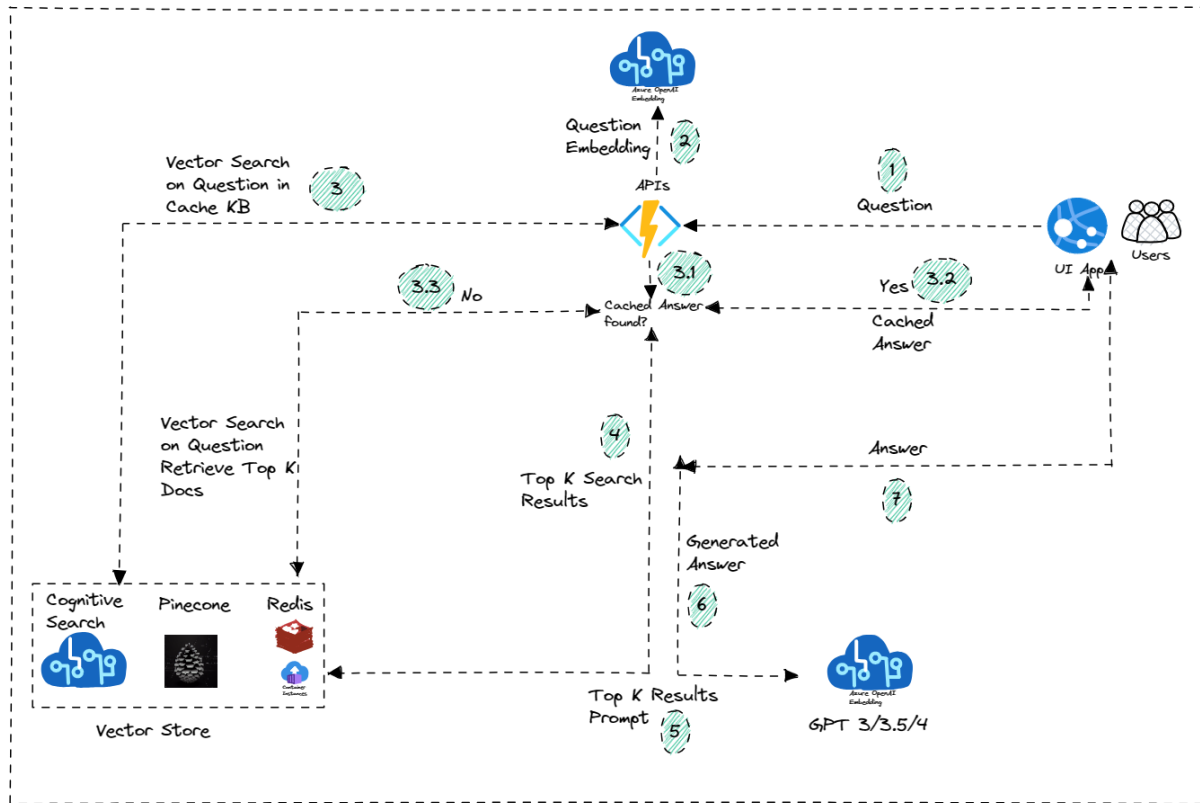


Azure Architecture



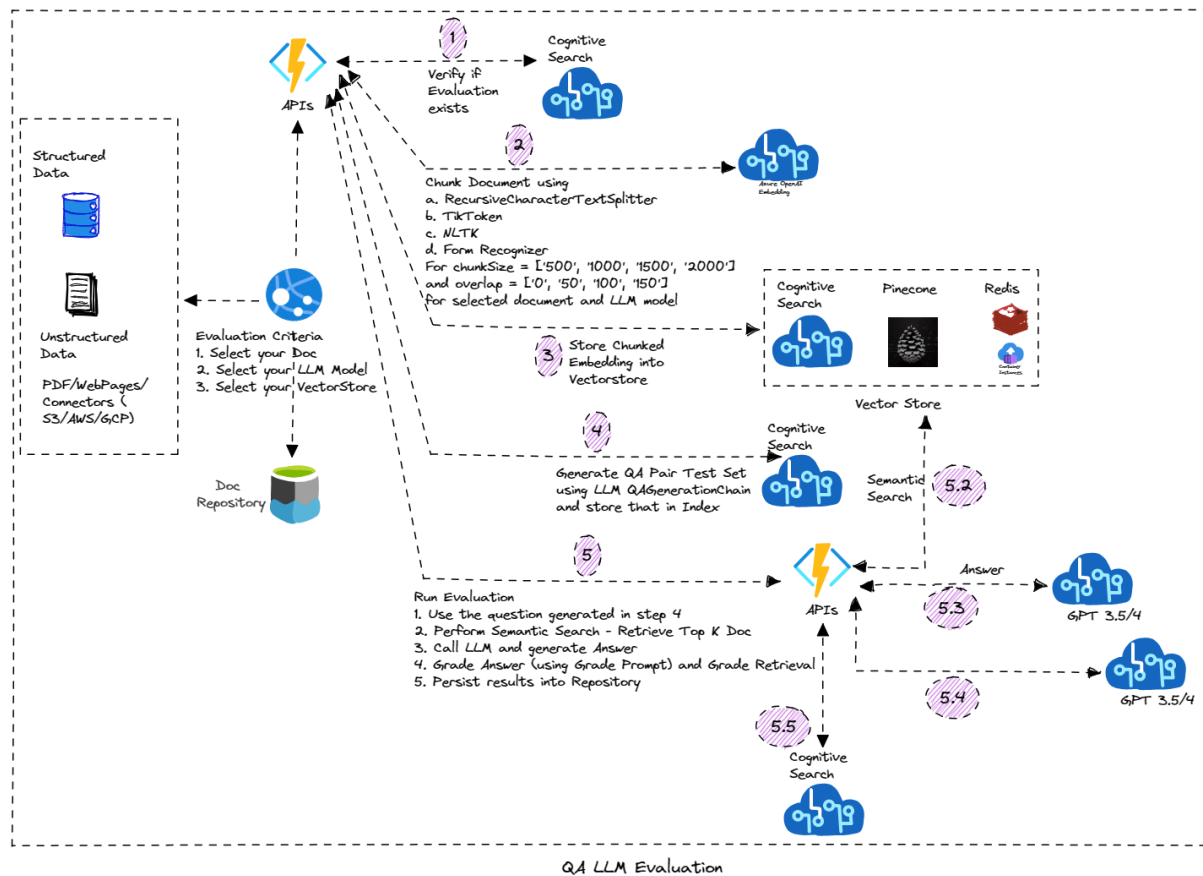
- Azure ADLS Gen2 - Store the documents with metadata
- Azure Cognitive Search as Vector store to store/search embedding data
- Azure App Service & App Service Plan to host the front end UI
- Azure OpenAI - Embedding, GPT 3.5, Turbo, 4, Instruct Models
- Azure CosmosDB - Store Session History, Log Prompts & Completion
- Azure Functions to host the backend API and Orchestration
- Azure Cognitive Services - Translator, Speech, Form Recognizer

QA over your data with Cache



QA over your data with Cache

QA LLM Evaluation



Getting Started

Get Started

Configuration

Application and Function App Configuration

Resources

- Revolutionize your Enterprise Data with ChatGPT: Next-gen Apps w/ Azure OpenAI and Cognitive Search
- Azure Cognitive Search
- Azure OpenAI Service

-
- Redis Search
 - Pinecone
 - Cognitive Search Vector Store

Contributions

We are open to contributions, whether it is in the form of new feature, update existing functionality or better documentation. Please create a pull request and we will review and merge it.

Note

Adapted from the repo at OpenAI-CogSearch, Call Center Analytics, Auto Evaluator and Edgar Crawler