Purpose

We need to be able to set proper HTTP headers when calling AzureOpenAI APIs to make it possible for AzureOpenAI team for usage tracking.

Goal

Follow the guidelines from AzureOpenAI team on HTTP header definitions.

Should be able to break down detailed PromptFlow usage scenarios, i.e.:

Authoring/test/evaluation runs

Batch inference (prompt flow converted to component running in a pipeline)

Real-time inference (prompt flow deployed as online endpoint)

How are the AzureOpenAI APIs called

The Azure OpenAI APIs are called by the OpenAI Python SDK. See here for details.

OpenAI Python SDK will set some default headers when calling Azure OpenAI APIs. In addition, it also allows the callers (PromptFlow sdk for this case) define some custom HTTP headers.

The call path looks like:

PromptFlow SDK --(call with custom headers)--> OpenAI Python SDK --(call with custom headers + OpenAI populated headers)--> Azure OpenAI API

What custom headers to add

x-ms-useragent

The value should look like promptflow-runtime/{version} promptflow-sdk/{version}.

The promptflow-sdk part is a fixed value, which indicates that the request was from PromptFlow.

The {version} part is PromptFlow defined value. It might be either a semver like 0.0.1 or a git commit hash. This field is for PromptFlow internal usage only and the format might change in the future.

As mentioned above, we call Azure OpenAI's endpoint via the OpenAI's python library, which is not owned by our team. We chose x-ms-useragent rather than the default User-Agent due to "This header is used when the platform does not support changing the User-Agent (i.e. in a browser)."

ms-azure-ai-promptflow

This is a customized header for PromptFlow.

According to the design doc , all the headers that start from ms-azure-ai- will be logged. We chose this field to log some PromptFlow-specific information.

Header Explanation Value

ms-azure-ai-promptflow-run-mode promptflow run mode Flow, SingleNode, FromNode, BulkTest, Eval

ms-azure-ai-promptflow-called-from AOAI call source aoai-tool - call from built-in aoai tool

ms-azure-ai-promptflow-subscription-id AML workspace subscription id 96aede12-2f73-41cb-b983-6d11a904839b

ms-azure-ai-promptflow-resource-group AML workspace resource group promptflow

ms-azure-ai-promptflow-workspace-name AML workspace name promptflow-canary-dev

ms-azure-ai-promptflow-workspace-id AML workspace id

ms-azure-ai-promptflow-edition promptflow runtime edition enterprise, community

ms-azure-ai-promptflow-compute-type promptflow runtime compute type managed\_online\_deployment, compute\_instance, local

ms-azure-ai-promptflow-runtime-mode promptflow runtime mode compute - authoring run, serving - scoring run

ms-azure-ai-promptflow-flow-id run flow id b6736c3d-5446-4b21-9fd1-c7f1246e67c3

ms-azure-ai-promptflow-root-run-id root run id 63ea2263-90ab-4473-b9dc-0e47a55eef73

ms-azure-ai-promptflow-index index 0

ms-azure-ai-promptflow-run-id run id 63ea2263-90ab-4473-b9dc-0e47a55eef73\_0

ms-azure-ai-promptflow-variant-id variant id variant\_0

References

"HTTP Headers for Telemetry" from Cognitive Services wiki page

The design doc listed in the wiki page