管理ポータル操作ガイド（利用状況計測編）
はじめに

本書は、本サービスのシステム管理者が利用する管理ポータルを利用方法について記載したガイドです。

本書の対象読者は以下を想定しています。

- 本サービスを利用する、お客様のシステム管理者
- 本サービスを利用する、システムの運用構築を支援するSI担当者

本ページの記載内容

本ページでは以下の機能について説明します。

- テナント単位でのAPI利用状況の確認
- テナント単位でのトークン数(入力/出力)利用状況の確認
- テナント単位でのインデックス使用量の確認

概要

本書は、利用状況画面に表示しているAPI利用状況の集計対象サービス、集計スケジュールについて説明するガイドです。

各画面の利用方法については利用状況確認編を参照ください。

集計対象

API利用状況の集計対象は以下になります。

- APIコール数
- 入力トークン数
- 出力トークン数

APIコール数

APIコール数は以下を実施することでカウントされます。

サービス	対象URL	カウント単位
一般対話API(OpenAI API)の利用	**<Generative AI APIベースURL>**genai-oai-api/v1	API実行毎に1回
履歴付き対話APIの利用	**<Generative AI APIベースURL>**genai-api/v1/chat	API実行毎に1回

検索対話APIの利用	**<Generative AI APIベースURL>**genai-api/v1/searchchat	API実行毎に1回
テンプレート対話APIの利用	**<Generative AI APIベースURL>**genai-api/v1/templatechat	API実行毎に1回
チャットUIの利用(拡張対話APIの利用)	**<Generative AI APIベースURL>**genai-ui-api/v1/exchat	1メッセージ送信毎に１回

入力/出力トークン数

APIの入力/出力トークン数は以下を実施することでカウントされます。

サービス	**対象URL**	**カウント単位**
一般対話API(OpenAI API)の利用	**<Generative AI APIベースURL>**genai-oai-api/v1	API実行毎に1回集計 入力トークン数：リクエストのトークンのサイズ 出力トークン数：LLMからのレスポンスのトークンのサイズ
履歴付き対話APIの利用	**<Generative AI APIベースURL>**genai-api/v1/chat	
検索対話APIの利用	**<Generative AI APIベースURL>**genai-api/v1/searchchat	
テンプレート対話APIの利用	**<Generative AI APIベースURL>**genai-api/v1/templatechat	
チャットUIの利用(拡張対話APIの利用)	**<Generative AI APIベースURL>**genai-ui-api/v1/exchat	

集計スケジュール

API利用状況はリアルタイムで集計されません。
集計処理は定期実行され、前回の集計処理時以降のAPI利用状況が加算されます。

また、毎月1日にAPI利用状況をリセットします。

API管理ポータル・利用状況画面の**今月のAPI使用状況**をリセットし、**過去実績のAPI使用状況**に移動します。

請求額の算出方法

請求額は、利用したAPIの入力/出力トークン数とモデル毎に定められた度数から算出します。

モデル毎に定められた度数ついては、利用可能なLLM をご参照ください。