利用可能なLLM

はじめに

本書は、本サービスで利用できる、LLM(大規模言語モデル)の状況や注意点等について記載しています。
本サービスでは、NECが提供するcotomi LLM やOpenAI社のモデルなどを利用できます。

利用可能LLM一覧

2026/2/26現在、以下のモデルが利用可能です。

モデル名	説明	オプトアウト対象	注記
cotomi-v3.0	NECが提供するLLMモデル。高速・高精度を両立した回答が期待できる。	○	
gpt-5.2-us	GPT-5.2の米国東部リージョンデプロイモデル	○	LLMへのリクエストを米国東部リージョンで処理
gpt-5.1	GPT-5.1のグローバル標準デプロイモデル	○	Azure東日本リージョンに閉じない可能性あり
gpt-5	GPT-5のグローバル標準デプロイモデル	○	同上
gpt-5-mini	GPT-5 miniのグローバル標準デプロイモデル	○	同上
gpt-5-nano	GPT-5 nanoのグローバル標準デプロイモデル	○	同上
gpt-5-nano-us	GPT-5 nanoの米国東部リージョンデプロイモデル	○	LLMへのリクエストを米国東部リージョンで処理
gpt-4.1	GPT-4.1のグローバル標準デプロイモデル	○	Azure東日本リージョンに閉じない可能性あり

gpt-4.1-mini	GPT-4.1 miniのグローバル標準デプロイモデル	○	同上
gpt-4.1-nano	GPT-4.1 nanoのグローバル標準デプロイモデル	○	同上
gpt-4o	GPT-4oのグローバル標準デプロイモデル	○	同上
gpt-4o-mini	GPT-4o-miniのグローバル標準デプロイモデル	○	同上
o4-mini	OpenAI o4-miniのグローバル標準デプロイモデル	○	Azure東日本リージョンに閉じない可能性あり
o3	OpenAI o3のグローバル標準デプロイモデル	○	同上
scan-std-model-v1-jp	図表文脈理解使用時に利用するモデル	○	国内リージョン利用限定モデル
scan-std-model-v1-apac	図表文脈理解使用時に利用するモデル	○	国内・国外リージョン自動選択モデル 以下のリージョンを選択する可能性があります。 東京、ソウル、大阪、ムンバイ、シンガポール、シドニー
scan-std-model-v2-apac	図表文脈理解使用時に利用するモデル(デフォルト)	○	国内・国外リージョン自動選択モデル 以下のリージョンを選択する可能性があります。 東京、ソウル、大阪、ムンバイ、シンガポール、シドニー

multilingual-e5-large	多言語対応のテキスト埋め込みモデル	○	新規のベクトル化では、本モデルが利用される
claude-sonnet-4.5	高精度と速さを両立したバランス型モデル（グローバル）	○	国外を含むグローバルリージョンで処理される
claude-sonnet-4.5-jp	高精度と速さを両立したバランス型モデル（日本リージョン）	○	日本リージョン(東京・大阪)で処理される
claude-haiku-4.5	シンプルなタスクに適した低コストで高速なモデル（グローバル）	○	国外を含むグローバルリージョンで処理される
claude-haiku-4.5-jp	シンプルなタスクに適した低コストで高速なモデル（日本リージョン）	○	日本リージョン(東京・大阪)で処理される
claude-sonnet-4.0	高精度と速さを両立したバランス型モデル（グローバル）	○	国外を含むグローバルリージョンで処理される

補足

- GPTモデルおよびoシリーズモデルの詳細は Azure によって直接販売される Foundry Models - Microsoft Foundry(外部サイト）をご参照ください。

モデル利用時に計上される度数

2026/2/26現在のモデルと度数の関係です。

1,000,000トークンごとに計上される度数 を記載しています。ご利用のモデルによって、同一の文章でも消費されるトークン数が異なる場合があります。詳細については、販促資料をご確認いただくか、問い合わせ窓口までご連絡ください。

モデル名	入力トークンに対する度数	出力トークンに対する度数	備考
cotomi-v3.0	40	160	
gpt-5.2-us	28	224	Azure OpenAIのページ参照

			Azure OpenAI Service – 価格 | Microsoft Azure
gpt-5.1	20	160	同上
gpt-5	20	160	同上
gpt-5-mini	4	32	同上
gpt-5-nano	0.8	6.4	同上
gpt-5-nano-us	0.8	6.4	同上
gpt-4.1	32	128	同上
gpt-4.1-mini	6.4	25.6	同上
gpt-4.1-nano	1.6	6.4	同上
gpt-4o	40	160	同上
gpt-4o-mini	2.4	9.6	同上
o4-mini	17.6	70.4	同上
o3	32	128	同上
scan-std-model-v1-jp	54	27	
scan-std-model-v1-apac	54	27	
scan-std-model-v2-apac	54	267	
multilingual-e5-large	1.6	0	
claude-sonnet-4.5	54	267	
claude-sonnet-4.5-jp	59	294	
claude-haiku-4.5	18	89	
claude-haiku-4.5-jp	20	98	

claude-sonnet-4.0	54	267	
AIガードレールモデル (guardrail-model-v1.0)	10	40	AIガードレールによる入出力チェックの度数となります

モデル利用時に計上される度数の表に掲載されていないモデルに対しては、度数を計上しません。

注意事項

- グローバル標準デプロイモデルは、東日本リージョン以外のデータ センターに動的にルーティングされる場合があります。
- 新規作成したインデックスではmultilingual-e5-largeがベクトル化に利用されます。
- Claudeモデルの利用時において、additionalModelRequestFieldsの引数を利用することができません。