第2章 OpenAI API × Google Colabで環境構築から公開データセットを使った実力検証まで

言語モデルを使った異常検出AIの準備

佐藤 聖

ステップ1…OpenAIのAPIサービスの中から最適なVLMモデルを選ぶ

第1部では，VLMを使った異常検出AIを体験するために，Google ColabとOpenAIのAPIサービスを利用します．OpenAIのAPIサービスとは，OpenAIが開発した高性能なAIモデル（GPT-4o，DALL・E，Whisperなど）を，外部のアプリケーションやサービスに組み込むためのインターフェースです．

● OpenAI APIを使う理由

本格的なVLM実験には，環境構築が不要で手軽に始められるOpenAI APIがお勧めです．理由を次に示します．

▶ローカル環境の場合

Hugging Faceからモデルをダウンロードできますが，業務レベルの性能には高性能なワークステーションが必要です．一般的なGPUではメモリ不足になりがちで，モデル量子化などの工夫が求められます．

▶クラウド環境の場合

AWSなどのクラウドを利用すると，複数GPU構成が必要となり，コストがかさむ傾向があります．

▶APIサービスの場合

ローカル環境の制約を受けずに高性能なVLMを利用できるOpenAI APIが，効率的で手軽な選択肢となります．

● Google Colabでモデルを実行することも可能だが…

今回は実験環境にGoogle Colabを利用し，モデルの実行にはOpenAIのAPIサービスを使用します．OpenAIのAPIサービスを利用すると，応答速度はColab上のGPTモデル実行に比べてかなり高速で，より低コストで利用できます．

Google Colabには，実験時の負担や非効率な要因があります．例えば，コンシューマ向けGPUを上回るデータ・センタ向けGPUを利用できますが，大規模モデルの処理にはメモリが不足し，利用できるGPUはときどきで変動するため，推論速度が遅くなりがちです．

さらに，Google Colabでのモデルのダウンロードには時間がかかり，ストレージやメモリの制限により大容量モデル・ファイルの扱いが困難です．ランタイムが停止するとモデルや画像ファイルが消失するため，その都度，再ダウンロードやアップロード，またはGoogleドライブへの退避が必要になる点も大きな負担です．

表1(2) **候補モデルの比較**（2025年6月22日時点）

カテゴリ	推論モデル	フラッグシップ・チャット・モデル	コスト最適化モデル
概要	複雑な文脈や因果関係を理解し，高度な推論に対応する	高精度かつ汎用性の高い主力モデルで，あらゆる用途に対応する	軽量／高速／低コストで限定的な用途に向く
参考モデル	GPT-4o	GPT-4.1，GPT-4o	GPT-4.1 mini，GPT-4o mini
推論／知能（5段階評価，主観）	3	4	3
コンテキスト・ウィンドウ	128,000	1,047,576	1,047,576
最大出力トークン	16,384	32,768	32,768
入力コスト（100万トークン）	2.50ドル	2.00ドル	0.15ドル
出力コスト（100万トークン）	10.00ドル	8.00ドル	0.60ドル