2024 年 7 月 23 日
株式会社ヘッドウォータース
(コード番号：4011 東証グロース)

「Phi-3」「Llama-3」「GPT-4o mini」などの小規模言語モデルを使用して生成 AI の回答精度を向上させる「SLM ファインチューニング」カスタムサービスを開始

AI ソリューション事業を手掛ける株式会社ヘッドウォータース（本社：東京都新宿区、代表取締役：篠田 庸介、以下「ヘッドウォータース」)は、生成 AI の業務活用を推進する企業向けに「SLM ファインチューニング」カスタムサービスの提供を開始しました。

同サービスは、マイクロソフト株式会社が提供する「Azure AI モデルカタログ」から選べるオープンソース AI 基盤モデル「Phi-3」、「Llama-3」、並びに「GPT-4o mini」を中心とした小規模言語モデルを使用して、生成 AI の回答精度を向上させるサービスとなり、生成 AI が作成する文章の精度では業務利用が厳しいと考えられている企業に有用なサービスとなっております。

ヘッドウォータースでは、Azure OpenAI Service による企業向け GPT サービスラインナップの拡充を行い、企業向けの生成 AI ならびに LLM（Large Language Model：大規模言語モデル）、ヘッドウォータースの技術力を活かした RAG(Retrieval Augmented Generation)システム、SLM（Small Language Model：小規模言語モデル）を使ったエッジ AI など多くのソリューションを開発してまいりました。

生成 AI の業務活用において多くのお客様から、「専門用語、業界用語、社内用語に対応させたい」「特定のキーワードが出た場合にサジェストやレコメンドを行いたい」「回答精度を向上させたい」と言う共通の課題であり、ヘッドウォータースでもこれまで課題に対するソリューションを模索してまいりました。

このような声に応えるため、ヘッドウォータースでは、マイクロソフトの SLM「Phi-3」と、Meta 社の「Llama-3」、OpenAI 社の「GPT-4o mini」を中心とした「SLM ファインチューニング」カスタムサービスを開始しました。

■RAG の課題

通常、LLM を各社が業務活用するためには、各社の業務データや独自プロンプトなどによって LLM をカスタムする必要があります。カスタム手法としては RAG とファインチューニングの 2 種類に分けられますが、ファインチューニングは難易度が高く効果的なデータが必要であり、コストパフォーマンス面でバランスの良い RAG から着手するケースが主流となっています。

一方、RAG の課題感として、「多くのデータを参照させすぎることによるハルシネーション（誤回答）」や、「生成 AI で正答率を高めたいケースへの対応」、「一般人が知っている用語ではない社内用語、業界用語、専門用語への対応」など特定のタスクにおいて不十分なシーンも見受けられます。

そこでヘッドウォータースは、「社内用語、業界用語、専門用語への対応」「正答率を高めるための対応」として SLM の活用によって課題解決を行います。

■SLM の特徴

SLM の主な特徴は、「LLM の軽量化」にありますが、その他の特徴として「扱うデータ量が少ない」事にあります。

SLM の学習データとして「業界固有の用語やニュアンス」や「間違ってはいけない回答」など「他のナレッジよりも優先するべきナレッジ」を SLM に用意することで、不正確さや無関係な情報を生成するリスクを最小限に抑える事ができます。

さらに SLM は LLM と比較してコンピューティングリソースが抑えられる為、運用コスト効率が高く、応答時間の短縮や消費エネルギー削減と言ったメリットがあります。

SLM には、マイクロソフトが提供する「Phi-3」を活用することで Microsoft Azure や Copilot+ PC との親和性を考慮、さらに SLM の弱点とされている日本語対応を解決するために、Meta 社の「Llama-3」をベースに開発された日本語学習済みモデルを利用し、運用コストと速度における課題解決のために OpenAI 社の「GPT-4o mini」を活用します。

本来高コストになりやすいファインチューニングも「扱うデータ量が少ない SLM を利用する」事で、LLM のファインチューニングよりも安価に提供が可能となりました。

ファインチューニングには、本来データ サイエンスの専門知識が必要となりますが、長年培ってきた機械学習の知見と複数人在籍する Kaggle メダリストの知見を掛け合わせることで、LLM と社内用語や業界用語を明示的に分ける手法において複数の導入実績があり、精度の向上も確認されています。

ヘッドウォータースでは、SLM ファインチューニングと Microsoft Fabric を活用した Advanced RAG サービスや、Microsoft Azure で構成された生成 AI 基盤「SyncLect Generative AI」と SLM ファインチューニングを組み合わせて提供することで、更なるコストパフォーマンスの向上に努めています。比較的高い正答率を求められる製造業や金融業、放送業、ヘルスケア業などのエンタープライズ企業で生成 AI の業務活用や、生成 AI を活用したお客様のサービスプラットフォーム支援を行ってまいります。

■今後について

今後は、SLM サービスラインナップを拡充することで次のようなソリューション展開を図ってまいります。

- ✓ マルチモーダル SLM「GPT-4o mini」「Phi-3 Vision」や「Florence-2」を活用したマルチタスクエッジ映像解析
- ✓ 個人情報をクラウドに持ち出さない生成 AI×オンプレミス
- ✓ オフライン環境に対応するローカル SLM
- ✓ Copilot+ PC 上で稼働する Windows AI アプリケーション
- ✓ モバイルデバイス上で稼働するオンデバイス SLM アプリケーション ...etc

ヘッドウォータースでは、アライアンス戦略を中長期戦略の柱として掲げており、顧客企業ともビジネスパートナーになることで共に生成 AI 経済圏を拡大する取り組みを行っております。顧客ビジネスに生成 AI を組み込み、相互送客することで生成 AI がより身近に、当たり前に利用される世界へと近づけてまいります。

なお、本件による当社の当期業績に与える影響は軽微であります。今後開示すべき事項が発生した場合には速やかにお知らせいたします。

以上

■SLM（小規模言語モデル）とは

SLM（小規模言語モデル）は、LLM（大規模言語モデル）よりもサイズが小さく軽量化された言語モデルです。高速なトレーニングと推論が可能で、リソース効率も高まり、コストパフォーマンスに優れています。また、リソースに制約のあるデバイスやエッジコンピューティングに適しており、セキュアで機密性が高いと言った様々な特徴があります。より小型となる言語モデルの可能性が生成 AI カテゴリーで注目されており、小規模言語モデルの採用が増加しております。

■ファインチューニング(Fine-Tuning：微調整)とは

ファインチューニングとは、既に学習済みのモデルに新たな層を追加し、モデル全体を再学習する手法です。 モデルを再利用するため、一から学習するよりも短時間で少ないデータでモデルの構築が可能です。

■Phi-3 とは

マイクロソフトが提供するオープンソースの小規模言語モデル (SLM) です。さまざまな言語、推論、コーディング、数学ベンチマークで同等のサイズと次段階のサイズのモデルを上回る、最高レベルの能力とコスト効率を発揮します。

https://azure.microsoft.com/ja-jp/products/phi-3

■GPT-4o mini とは

OpenAI 社が提供するマルチモーダル言語モデルの「GPT-4o」の小型モデルです。開発者の利用コストが「GPT3.5」よりも 60%以上安価なモデルであり、精度の向上に加えて速度も大幅に速くなっています。

■Azure AI モデルカタログとは

すぐに使えるようにパッケージ化されたトップクラスの基盤モデルで、OpenAI、Meta、Mistral AI、Stability AI、Hugging Face など主要なオープンソース生成 AI モデルの開発を加速します。

https://azure.microsoft.com/ja-jp/products/ai-model-catalog

■RAG(Retrieval Augmented Generation)とは

Retrieval Augmented Generation（RAG）は、大規模言語モデル（LLM）と外部のデータベースや情報源を結びつけるための新しい技術です。外部の知識ソースを検索し、より強化した文章生成を行います。

■Copilot+ PC とは

Copilot+ PC は、リアルタイム翻訳や画像生成などの AI を多用するプロセス専用のコンピューター チップであり、40 兆回以上の操作を秒速で実行 (TOPS) できる超高速ニューラル プロセッシング ユニット (NPU) を搭載した、新しいクラスの Windows 11 PC です。

https://www.microsoft.com/ja-jp/windows/copilot-plus-pcs

■参考

Microsoft Fabric をデータプラットフォームとした 「Advanced RAG」サービス開始

https://www.headwaters.co.jp/news/gen_ai_microsoft_fabric_advanced_rag.html

産業用エッジ生成 AI ソリューション「LLaVA Edge Vision」を開発

https://www.headwaters.co.jp/news/nvidia_genai_llava_edge_vision_ai_expo.html

生成 AI×エッジ AI に向け小規模言語モデル SLM と画像言語モデル VLM の検証を開始

https://www.headwaters.co.jp/news/edgeai_generativeai_nvidia_slm_vlm.html

Azure OpenAI Service リファレンスアーキテクチャの Advanced パートナー認定について

https://www.headwaters.co.jp/news/azure_openai_service_advanced_partner.html

<会社情報>

会社名：株式会社ヘッドウォーターズ

所在地：〒163-1304　東京都新宿区西新宿 6-5-1 新宿アイランドタワー 4 階

代表者：代表取締役　篠田 庸介

設立　：2005 年 11 月

URL　：https://www.headwaters.co.jp

<本件のお問い合わせ>

株式会社ヘッドウォーターズ

メール：info@ml.headwaters.co.jp