ハイライトレポート

〔2026.10.8〕

MediaTek Genio 520で実装するEdge SLM

生成AIをエッジ端末上で動作させる「Edge SLM」をテーマに、MediaTek Genio 520を用いたウェビナを開催しました。本ウェビナでは、SLMとRAGを組み合わせた実機デモを通じて、NPU上でモデルを動作させるための実装構成や、TTFTを意識した応答設計、外部データを回答根拠として活用するRAGの仕組みを解説。本レポートでは、実機への実装・運用を検討するうえで押さえておきたい技術的なポイントを整理します。

Edge SLMを成立させるNPU実行セット

Edge SLMは、端末のメモリ容量や消費電力の制約下で、特定用途に必要な推論を成立させる設計です。今回の実機デモでは、Genio 520のNPUでSLMを安定して動作させるため、学習済みモデルに量子化や入力サイズの固定を行い、NPUで実行できる形式（DLA）へ事前に変換しました。生成したDLAは、Tokenizerや実行設定などと整合させた「実行セット」として管理します。動作確認済みの組み合わせ単位で更新・切り戻しを行うことで、再現性のある実行環境を維持できます。

体感速度はTTFTと出力設計で決まる

端末の使いやすさは、回答完了までの総時間だけでなく、質問してから最初のトークンが生成されるまでの時間であるTTFT（Time To First Token）や、画面に最初の反応が表示されるまでの時間にも左右されます。RAG検索、Prompt読み込み、Prefill、Generation、画面表示、音声出力に処理を分け、必要な根拠だけをPromptへ渡すことで初動の短縮につなげます。生成結果は内部タグを除去し、UIには文字単位、TTSには句読点や文末単位で渡すなど、それぞれの出力特性に合わせた設計が必要です。

RAGは検索候補を回答根拠へ変換する

SLMが保持していない商品情報、価格、売り場、FAQなどの現場知識は外部データとして分離し、RAGから必要な情報を取得します。商品情報を意味フィールドごとに分け、Embeddings DBで候補を検索した後、質問意図、対象フィールド、言語、商品IDで絞り込み、必要な根拠だけをSLMへ渡します。JSONとEmbeddings DBをあわせて更新し、根拠が取得できない場合は無理に回答を生成しないことも、誤った案内を防ぐうえで重要です。

実機デモで確認された設計上の注意点

利用者交代時には会話履歴を初期化し、JSONとEmbeddings DBを同期して検索対象と回答根拠の整合性を維持します。検索候補と質問との関連性を確認してからSLMへ渡すことや、UIとTTSで出力単位を分けることも実機品質を左右します。Edge SLMを実機で利用するには、NPU実行、RAGによる根拠管理、応答速度、画面・音声出力までを含めて検討し、対象業務や必要な知識、端末条件を明確にしたうえで段階的に検証することが重要です。

■ 本日の登壇者 ■



VIA Technologies Japan株式会社
VISVIA事業部 セールスマネージャー
小間 拓実

TECH WORDS 用語集

本レポートで使用する主な技術用語を、わかりやすく解説します。

SLM Small Language Model	小規模言語モデル。大規模なLLMに比べてモデルサイズを抑え、端末上での利用や特定用途への適用に適した生成AIモデルです。
NPU Neural Processing Unit	AIの推論処理に特化したプロセッサ。画像認識や言語モデルなどのAI処理を、端末上で高速・省電力に実行するために用いられます。
DLA Deep Learning Archive	MediaTek Genioで使用される、NPU向けのモデル実行形式。学習済みモデルをNPUで効率的に実行できるよう、事前に最適化・変換して生成するファイル形式です。
Tokenizer	文章をモデルが処理できる最小単位（トークン）に分割する仕組み。入力されたテキストを数値のトークン列に変換します。
Prefill	入力された質問や指示（プロンプト）をモデルに読み込ませ、生成の準備を行う処理。ここまでの処理時間が、TTFTに大きく影響します。
Generation	モデルが回答のトークンを1つずつ生成していく処理。生成速度（Tokens Per Second：TPS）は、この処理の速さを示す指標です。
RAG Retrieval-Augmented Generation	外部データを検索し、取得した情報を根拠として回答を生成する仕組み。モデルが持たない最新情報や特定ドメインの知識を活用できます。
Embeddings DB	文章などの意味を数値データ（ベクトル）に変換して保存するデータベース。意味の近い情報を検索し、RAGで利用します。
TTS Text-to-Speech	テキストを音声に変換する技術。生成した回答を音声で出力する際に使用します。
TTFT Time To First Token	質問してから、モデルが回答の最初のトークンを生成するまでの時間。ユーザーの体感速度を左右する指標です。

他記事、ウェビナ情報はこちら

リョーサン菱洋テクラボ
エンジニアによりそうマガジンサイト

Copyright© Ryoyo Ryosan, Inc. All rights reserved