原著論文 Original Paper

パーソナルAIに向けたマルチデバイスRAGの設計

Design of a Multi-Device RAG System for Personal AI

村松 沙那恵 Muramatsu Sanae
NTT ソフトウェアイノベーションセンタ
NTT Software Innovation Center
sanae.muramatsu@ntt.com

星野 玲那 Hoshino Reina （同　上）
reina.hoshino@ntt.com

正木 晶子 Masaki-Kato Akiko （同　上）
akiko.masaki@ntt.com

江田 毅晴 Eda Takeharu （同　上）
takeharu@acm.org

keywords: MultiDevice, RAG, LLM, PersonalAI

Summary

近年，大規模言語モデル（Large Language Models: LLM）の発展に伴い，ユーザ体験の高度化が期待されている．特に，個人のプライベート情報を用いた AI のパーソナライズは，利便性を大幅に向上させる可能性がある．一方で，従来のクラウドベースの LLM においては，プライベート情報をクラウドにアップロードする必要があるため，プライバシー侵害のリスクが懸念される．そこで本研究では，プライベートな情報をクラウドに集約することなく，複数のエッジデバイスに保存し活用するための，マルチデバイス RAG（Retrieval-Augmented Generation）システムを設計および実装した．本システムは，個人が保有するスマートフォンや PC などの端末上にベクトルデータベース（Vector DB）を配置し，ベクトル化および検索処理をローカルで完結させている．また，端末間の通信には WebRCT を用いた P2P 接続を用い，必要最小限の情報のみを相互に共有する．これにより，1. 個人情報の漏洩リスクの最小化，2. ネットワーク負荷の軽減，3. ユーザ主権のデータ管理，4. サービス運用コストの削減，といった利点を実現する．これらのシステム設計を元に試作したチャットアプリを通じて，複数のデバイスに分散したプライベート情報を活用し，それらを統合した結果が得られることを確認した．

1. はじめに

近年，LLM の発展に伴い，ユーザ体験の高度化が期待されている．特に，個人のプライベート情報を用いた AI のパーソナライズは，利便性を大幅に向上させる可能性がある．一方で，既存のクラウドベースの LLM を利用する場合，プライベート情報をクラウドにアップロードする必要があるため，プライバシーの侵害や情報漏洩といったリスクや，中央集権的なデータ管理，ネットワーク帯域の負荷といった課題がある．

そこで本研究では，プライベートな情報をクラウドに集約することなく，複数のエッジデバイスに保存し活用するための，マルチデバイス RAG システムを提案する．クラウドに依存せずに，個人のプライバシーを保護しながら活用することで，パーソナル AI の実現を目指している．

2. 関連研究

2·1 オンデバイス推論

プライバシー保護の観点に加えて，ネットワーク接続が不安定または存在しない環境，さらに推論の即時性が求められるユースケースにおいて，AI モデルをクラウドではなくローカルデバイス上で実行する，オンデバイス推論の重要性が高まっている．これに応じて，[Llama3.2a], [phi3-mini]，[Gemma 3n] などをはじめとする小規模な言語モデルや小規模なマルチモーダルモデルが，オンデバイスでのリアルタイム推論に適した形でリリースされている．また，エッジ環境での AI 推論を容易にするためのプラットフォームやツールも活発に開発されている．たとえば [LLM Farm] や [Google AI] などは，LLM を使ったチャットアプリケーション構築を可能とするオンデバイス推論支援ツールとして注目されている．さらに，[llama.cppb]，[LiteRT]，[ExecuTorch] といったフレームワークは，特定ハードウェア向けの最適化を施すことで，

エネルギー効率や推論速度を向上させるプラットフォームとして提供している．

これらの取り組みは，リソース制約のあるデバイスにおいても高精度かつ低レイテンシな自然言語処理を実現するための基盤技術である．これらの発展に伴い，エッジデバイスの利活用も一層拡大してくことが予想される．

2·2 プライバシーを配慮した方法

プライバシーを考慮しながら，エッジ端末を活用して推論･学習を行う手法が提案されている．privateLoRA[Wang 23] は，データの局所性を担保することでプライバシーを配慮した推論・学習を行う．この手法では，エッジデバイスにプライベートデータで学習された LoRA アダプタを配置し，一方で凍結したベースの重みは，クラウド側に配置することで，スケーラビリティを向上させている．また，[Li 23] では，プライバシー保護をしながら prompt-tuning を実施する方法を提案している．

これらの手法は，エッジ端末とクラウド側の LLM を合わせて活用するため，連携するための専用の LLM を用意する必要に加えて，エッジ端末とクラウドとの通信が発生する．本研究では，ネットワーク制約がある中でも安定動作するよう，エッジ端末内だけでの推論を対象とした．

2·3 複数のエッジデバイス協調

計算リソースを相互に貸し合うことで，大規模な AI モデルの推論や学習を可能とする分散処理の研究が進められている．たとえば，AI モデルを複数のデバイス間で分割し，あるデバイスが AI モデルの一部を実行し，実行した中間状態を他のデバイスに送信することで，残りの計算を継続するといったアプローチがある [Xu 23][Bakhtiarnia 23].

本研究で提案するシステムは，こうした計算負荷の分散を目的としたアプローチとは異なり，複数のエッジデバイスに分散して存在するデータ自体を活用する点に主眼がある．すなわち，本システムは計算リソースの分散利用ではなく，データのローカリティを保持したまま，エッジデバイスの協調により推論を行う方式である．

2·4 マルチデバイス RAG

LLM を活用したシステムの多くは，RAG を用いた構成が採られている．RAG とは，外部知識を動的に活用することで，LLM の性能を向上させるためのアプローチである．RAG は主に 2 つのコンポーネントから構成される． 1つは，Retriever（検索器）である．Retriver は，ユーザからの入力（query）に，事前に構築された Vector DB から，関連性の高い文書や文節を取得する． 2つ目は，Generator（生成器）である．Generator では，検索された文書をもとに，LLM が自然言語で応答を生成する．

[Xu 25] では，RAG の分散型実装として，エッジデバイス間のピアツーピアネットワークを活用する手法が提案している．各エッジデバイスにはローカルの LLM および Vector DB が搭載されており，外部知識の取得が必要か否かの判断や，生成された応答が十分であるかどうかの評価を自律的に行う構成となっている．

このアプローチは，各エッジデバイスに問い合わせをするかを判断するためにキャッシュを持つことが特徴である．一方，本研究では，各エッジデバイスに問い合わせるという網羅性を重視した設計である．また，実用性を重視した観点から，実際にスマートフォンなどのエッジデバイスで動作可能な実装例を通じて，その有効性と実現可能性を示している．

3. 想定ユースケース

本システムの想定ユースケースは，スマートフォンや車載システムなどの個人デバイスと連携し，ユーザごとの文脈に応じた AI を構築することで，パーソナライズされた情報提示や支援を実現することである．以下に，その具体的に 2 点述べる．

3·1 ケ ー ス 1

久しぶりに再会した友人同士が，あらかじめ予約している宿泊地へ向かう途中で観光や食事を楽しめる場所を探すというシナリオである．それぞれのスマートフォンに蓄積されたプライベートな情報（例：過去の訪問履歴，嗜好，位置情報）を RAG を用いて抽出・検索し，複数人のコンテキストに基づいた候補地を生成する．さらに，周辺のイベント情報や混雑状況，地域のトレンドなどの外部情報も組み合わせ，リアルタイムで適応的なレコメンドを提供する．ユーザ同士の関係性や共通の関心を考慮した協調フィルタリング技術を応用することで，同行者の満足度も高めることができる．これにより，ユーザは事前の明示的な入力なしに，文脈に合った目的地候補を自然なかたちで提案される体験を得る．

3·2 ケ ー ス 2

ユーザが飲食店などのレビューを投稿することはないものの，食後に家族と共に車内で感想を交わすという状況に着目する．プライベートな車内空間におけるこのような自然発話を，音声認識と自然言語理解により処理し，明示的フィードバックでは得られないユーザの主観的評価を抽出する．抽出された意見は，スマートフォンに蓄積された検索履歴や位置履歴と組み合わせて分析され，次回以降のレコメンデーションに反映される．これにより，ユーザが意識的に情報を提供しなくても，その行動や会話から得られる暗黙的フィードバックを基に，好みに合致した提案が行える．加えて，音声データの処理はエッジデバイス上で行われ，必要な部分のみをクラウドと連

携させることで，プライバシー保護と処理効率を両立している．

4. 提案システム

4·1 概　　要

本研究が提案するシステム（図 1）は，ユーザ端末に分散して存在するデータをクラウドを介さずに処理する構成を採用している．具体的には，各エッジデバイス上に Vector DB を配置し，ベクトル化や検索処理をローカルに実行する．また，各端末は [webRTC] ベースの P2P 通信を通じて，必要に応じて他の端末から限定的に情報を取得し，必要最小限のデータもしくは結果のみのを他のデバイスへ受け渡す．この構成により，以下の利点が得られる：

- プライバシ保護：個人が保有する情報をクラウドに送信せず，エッジデバイス上での処理で完結することで，個人情報の漏洩リスクを最小化できる．
- ネットワーク負荷の軽減：ベクトル化とベクトル検索の大部分が端末内または P2P 経由で完結するため，ネットワーク負荷の軽減ができる．
- ユーザ主権のデータ管理：中央集権型アーキテクチャではないため，ユーザ主権のデータ管理ができる．
- コストメリット：サービス事業者がユーザの情報を管理する必要がなく，ユーザが保有するデバイスのストレージや計算資源を活用することで，サービス事業者側の運用コストを低減できる．

4·2 処理シーケンス

図 2 に，本研究で提案するシステムのシーケンス図を示す．本システムは，ユーザからの質問入力に基づき，複数のデバイス間で質問のベクトル化，コンテキスト検索，ならびに推論処理を協調して実行するフローで構成されている．Device 1 はユーザからの入力を直接受信するデバイスであり，Device 2 に保存された情報も参照しながら，最終的な応答の生成を行う．

シーケンスの具体的な処理は以下の通りである．まず，Device 1 はユーザの入力を受け取り，その質問をベクトル化し，Device 2 へ送信する．Device 1 および Device 2 は，それぞれが保持する Vector DB を用いて，ユーザ入力ベクトルとの類似性に基づく検索を行い，関連するコンテキスト情報を取得する．Device 2 で取得されたコンテキストは，Device 1 に転送され，Device 1 において統合・選別が行われる．取得された複数のコンテキストの選別には，ベクトル間の距離等の指標が用いられる．ユーザ入力と選別されたコンテキストを組み合わせてプロンプトを構成し，これを LLM への入力として推論処理を実行する．その結果得られた応答はユーザへ提示される．

なお，Device 1 または Device 2 においてベクトル化や LLM による推論が不可能な場合には，近傍のエッジデバイス（たとえばユーザの PC 等）を Device 3 として利用し，[Ollama] 等の API ベースのツールを介して処理を実行可能である．

本図では 2 台のデバイス構成を前提としているが，3 台以上の構成においても，本シーケンス図と同様の流れにより拡張的に対応可能である．

4·3 実装で必要となる技術

ここでは実装例として必要となる技術を紹介する．

§ 1 Flutter

[Flutter] は Google によって開発されたオープンソースの UI フレームワークであり，単一のコードベースで iOS，Android，Web，Windows，macOS，Linux といった複数のプラットフォームに対応したアプリケーションを構築できる点が特徴である．Flutter は一つの Dart 言語で記述されたコードベースを用いて，複数の OS 向けアプリを構築できるため，開発と保守のコスト削減が可能である．

§ 2 WebRTC

WebRTC は，W3C（World Wide Web Consortium）と IETF（Internet Engineering Task Force）によって標準化が進められているオープンソースプロジェクトであり，リアルタイムの音声通話・ビデオ通信・データ通信を可能とする技術である．エッジ間の p2p 通信を実現するために，webRTC を採用した．

5. 実　　験

提案された設計に基づいて複数のデバイスが連携可能かを検証するために実装を行い，その連携が実際に可能であるかを確認した．

5·1 実 験 環 境

本システムは，Flutter（v3.29.2）および Dart（v3.7.2）を用いて開発した．スマートフォンは，Xperia を用いた．ネットワーク構成図は，図 3 に示した．

ollama を使って，LLM は API 呼び出しをしている．

5·2 実 験 手 法

本実験では，OpenAI API を用いて日記データを生成した．一方のデバイスには 1 月分の日記データを保存し，他方のデバイスには 2 月分の日記データをそれぞれ保存した．その後，両デバイス間の接続を行い，それぞれのデバイスにのみ存在するデータを抽出可能であるかを検証した．

5·3 実 験 結 果

図 4 に示すように，1 月および 2 月の日記の内容を前提とする質問に対しても，適切に回答できることを確認した．

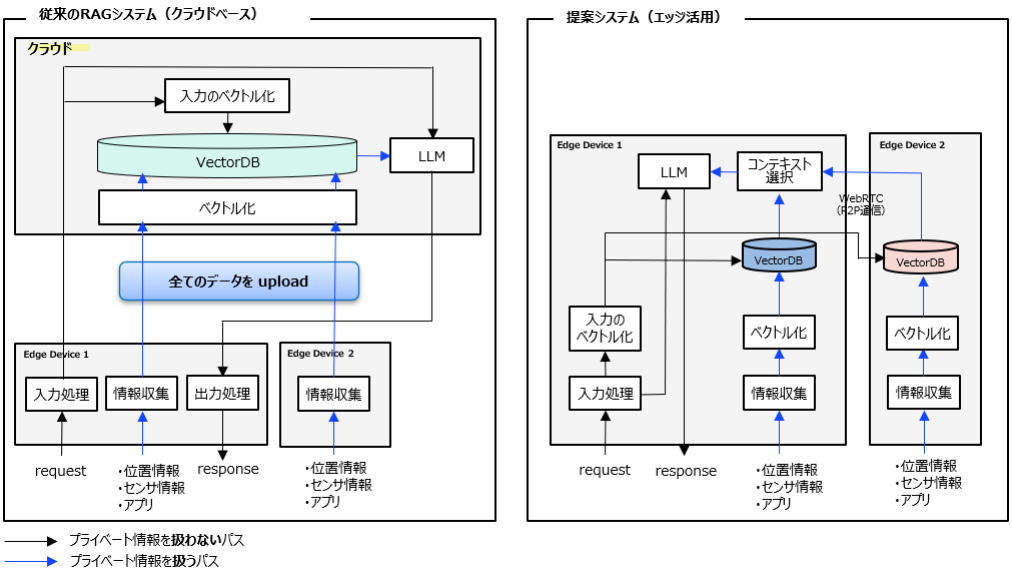

図 1 システム概要図

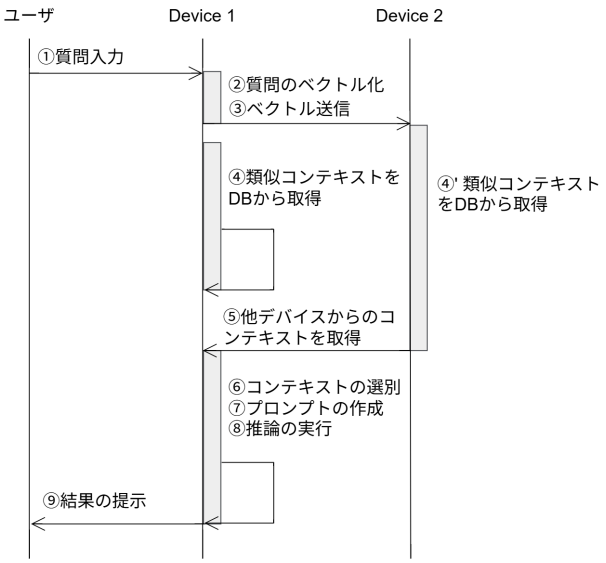

図 2 シーケンス図

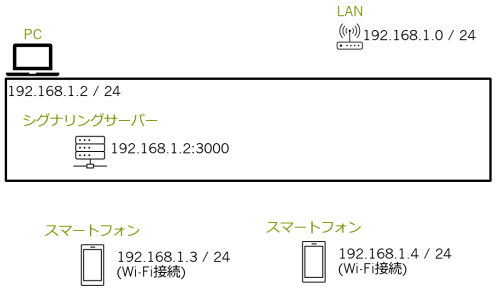

図 3 ネットワーク構成図

6. ま と め

本研究では，LLM の進展と実用化が進む中で，個人のプライベート情報を活用したパーソナライズ AI の可能性とその課題に着目した．プライベート情報が分散して存在する複数のエッジデバイス間で協調的に情報を活用する新たなアプローチを提案した．

本提案では，スマートフォンや PC などユーザが所有する各種デバイスを活用し，ローカルにパーソナライズを実現するマルチデバイス対応の RAG システムを設計・実装した．高い移植性を実現するために Flutter を採用し，また API サーバーの設置が困難な環境にも対応できるよう，WebRTC を用いた P2P 通信機構を導入した．

構築したチャットアプリを用いた実験により，複数デバイスに点在するプライベート情報を統合的に活用し，ユーザに最適化された応答が生成可能であることを確認した．今後は，さらなる精度向上とセキュリティ強化を図るとともに，実環境での運用やユーザスタディを通じた評価を進めていく予定である．

◇ 参 考 文 献 ◇

[Bakhtiarnia 23] Bakhtiarnia, A., Milošević, N., Zhang, Q., Bajović, D., and Iosifidis, A.: Dynamic Split Computing for Efficient Deep EDGE Intelligence, in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023)
[ExecuTorch] https://github.com/pytorch/executorch
[Flutter] https://flutter.dev/
[Gemma 3n] https://deepmind.google/models/gemma/gemma-3n/
[Google AI] https://github.com/google-ai-edge/gallery
[Li 23] Li, Y., Tan, Z., and Liu, Y.: Privacy-Preserving Prompt Tuning for Large Language Model Services, *CoRR*, Vol. abs/2305.06212, (2023)
[LiteRT] https://ai.google.dev/edge/litert?hl=ja
[Llama3.2a] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/
[llama.cppb] https://github.com/ggml-org/llama.cpp
[LLM Farm] https://llmfarm.tech/
[Ollama] https://ollama.com/
[phi3-mini] https://azure.microsoft.com/en-us/blog/introducing-phi-3-redefining-whats-possible-with-slms/

図 4 実験結果

[Wang 23] Wang, Y., Lin, Y., Zeng, X., and Zhang, G.: PrivateLoRA For Efficient Privacy Preserving LLM, *CoRR*, Vol. abs/2311.14030, (2023)
[webRTC] https://webrtc.org/?hl=ja
[Xu 23] Xu, D., He, X., Su, T., and Wang, Z.: A Survey on Deep Neural Network Partition over Cloud, Edge and End Devices, *CoRR*, Vol. abs/2304.10020, (2023)
[Xu 25] Xu, C., Gao, L., Miao, Y., and Zheng, X.: Distributed Retrieval-Augmented Generation (2025)

〔担当委員：××○○〕

19YY 年 MM 月 DD 日　受理