2025 年度未踏アドバンスト事業

プライバシーファーストな記憶支援アシスタントの開発

―完全ローカル AI 知識検索 ”Sonar”―

1. 背景

近年、情報技術の発展に伴い、個人が日常的に接する情報量は急激に増加しており、情報の整理・活用は現代社会における重要な課題となっている。こうした背景の中、個人の知識を体系的に管理する Personal Knowledge Management（以下、PKM）ツールへの関心が高まり、Obsidian、Notion、Evernote などの多様なツールが普及している。しかし、PKM ツールは情報を蓄積する機能には優れているものの、蓄積された情報を効果的に引き出す仕組みが不十分であり、多くのユーザーが情報整理の手間やコストに悩まされている状況にある。

この課題に対し、AI 技術を活用した情報検索・整理機能が注目されている。しかし、既存の AI 連携ツールの多くはクラウドベースのモデルに依存しており、ユーザーのパーソナルデータが外部サーバーに送信される構造となっている。日記、投資判断、未公開の研究メモ、人事評価メモといった極めて個人的かつ価値の高い情報ほど、外部保存には心理的な障壁が存在する。クラウドベースの PKM が E2E 暗号化を実施していたとしても、AI 処理のためにはデータの復号が必要となるため、プライバシー上の懸念を根本的に解消することは困難である。

このような状況を踏まえ、ローカル LLM の活用により、プライバシーを完全に保護しつつ、蓄積された知識を効率的に引き出すことのできる新たなアプローチが求められている。

2. 目的

本プロジェクトでは、ローカル LLM を用いた完全オフラインのリアルタイム記憶支援システム「Sonar」を、PKM ツール Obsidian のプラグインとして開発し、プライバシーファーストな知識検索体験を実現することを目的とした。具体的には、ユーザーのノートデータに対する高精度なセマンティック検索、自動インデックス生成、および LLM エージェントによる自律的な情報探索機能を、すべてローカル環境で完結する形で提供することを目指した。

3. 製品・サービスの内容

Sonar は、Obsidian プラグインとして動作する完全ローカルの AI 記憶支援システムである。言語モデルの推論バックエンドには llama.cpp を使用し、ネットワーク接続なしでの動作を実現している。推奨環境は RAM 32GB、16 コア GPU 搭載のマシン（MacBook Pro 相当）である。本システムのコア機能は以下の 3 つで構成される。

3.1 自動インデックス生成

Obsidian vault 内のドキュメントを自動的に検出し、ベクトル埋め込みおよび BM25 用トークンインデックスを生成する。ファイルの追加・変更・削除を監視し、インデックスをリアルタイムに更新する。音声メモについては Whisper による文字起こしを経てインデックス化を行い、マルチモーダルな情報蓄積に対応している。すべての処理は完全オフラインで実行される。

3.2 高精度検索

ベクトル類似度検索と BM25 全文一致検索を組み合わせた Hybrid retrieval を基盤とし、Reciprocal Rank Fusion（RRF）により両者の結果を統合する。さらに、Cross-encoder reranking により上位候補の順位を再計算することで、検索精度を実用レベルに引き上げている。MIRACL データセット（24,754 件の Wikipedia チャンク、200 件の質問文）を用いたベンチマークでは、nDCG@10 で 0.9657 を達成した（表 1）。

表 1　検索精度ベンチマーク（nDCG@10）

手法	nDCG@10
Hybrid+Reranking （BM25+BGE-M3+BGE-Reranker-V2）	0.9657
Hybrid（BM25+BGE-M3）	0.9035
BM25 only	0.7788

3.3 AI エージェントチャット

LLM エージェントが質問の観察・ツール実行・回答生成のループを自律的に繰り返し、vault 内の情報を探索・集約する。組み込みツールとして、Hybrid search+Reranking による高精度な根拠検索（search_vault）およびノートの作成・編集・削除（edit_note）を備える。加えて、カスタムツールとして Web コンテンツ取得（fetch_url）や Google Calendar 連携（get_google_calendar）などをプラグイン形式で追加可能であり、ユーザーのワークフローに合わせた柔軟な拡張が可能である。CRAG データセットを用いた End-to-end RAG ベンチマークでは、回答精度 43%、ハルシネーション率 32%を達成し、クラウドベースの RAG（text-embedding-3 + GPT-4.1-mini：回答精度 42%、ハルシネーション率 35%）と同等以上の性能をローカル環境で実現した（表 2）。

表 2　End-to-end RAG パイプライン精度（CRAG データセット）

システム	回答精度	ハルシネーション率

Sonar（BGE-M3 & BGE-Reranker-V2-M3 & Qwen-8B）	43%	32%
Cloud RAG（text-embedding-3 & GPT-4.1-mini）	42%	35%

4. 新規性・優位性

本プロジェクトの最も重要な新規性は、プライバシーファーストの思想的原則に基づく完全ローカル動作にある。既存の AI 連携 PKM ツール（Notion AI、Constella、NotebookLM など）の多くはクラウドモデルに依存しているが、Sonar はノートデータも推論プロセスも一切デバイスの外に出さない設計としている。さらに、OSS として公開することで、システムの透明性を担保し、データの流れをユーザー自身が検証可能な環境を提供している。

技術面では、Embedding × BM25 × Cross-encoder Reranking を組み合わせた RAG パイプラインにより、シンプルなクラウド実装に匹敵する検索・回答精度をローカル環境で実現した点が優位性として挙げられる。表 2 に示した通り、回答精度ではクラウド RAG を上回り、ハルシネーション率においても低い数値を達成している。また、エージェントアーキテクチャとカスタムツールによる拡張性は、ユーザー固有のワークフローへの適応を可能にし、単なる検索ツールにとどまらない情報活用体験を提供している。

5. 事業普及（または活用）の見通し

今後の事業普及に向けて、まずは"使えるツール"としての認知度向上を図る。具体的には、Obsidian Community Marketplace への正式登録およびアップデートごとのコミュニティへのアナウンスを通じてユーザー基盤を拡大する。並行して、コミュニティからのフィードバックを積極的に取り込み、Open core としての基盤機能の品質向上に注力する。

ユーザー基盤が一定規模に達した段階で、有償機能の提供によるビジネス化を目指す。有償機能としては、Gmail、Slack、ブラウザ履歴などの外部情報ソースとの高品質な公式連携ツール、およびモバイル端末から自宅のマシンへ E2E 暗号化で安全にリモート推論を行うモバイル対応機能（Remote RPC）を計画している。ローカル処理というコア機能は OSS として無償で提供し続けつつ、外部連携やモバイル対応といった応用的な機能により収益化を図る Open core モデルでの展開を想定している。

6. 期待される波及効果

Sonar の普及により、機密性の高い情報を扱わなくてはならない分野において、情報管理の効率化が期待される。特に、個人情報を取り扱う法務分野や、機密性の高い国防分野における活用が考えられる。例えば、自衛隊の潜水艦の運用において

は、環境としてはオフラインであるにもかかわらず、乗組員は大量の書類業務をこなす必要があり、こうした場面において Sonar が業務効率の改善に寄与する可能性がある。

また、情報の活用は企業の業務内に限られるものではない。研究機関における文献管理、クリエイターコミュニティにおけるインスピレーション管理など、アイディアや情報を外部に漏らさないように気を付ける必要性がある幅広い知的活動の場面に応用可能である。完全ローカルで動作するという特性は、これらの多様な利用場面において、情報漏洩への懸念なく AI による知識活用を促進する基盤となりうる。

7. イノベータ名（所属）

門脇 宗平（JuliaHub, Inc.）

（参考）関連 URL:

●aviatesk/obsidian-sonar: https://github.com/aviatesk/obsidian-sonar

●Obsidian: https://obsidian.md/