TEE を活用した安全な RAG データベース作成・管理方式

徐 振宇 [1,a)]　塚田 洋平 [1]　橋本 諒太 [1]　柏木 啓一郎 [1]　井上 知洋 [1]

1. 背景及び研究モチベーション

近年，企業や行政機関において，AI アプリケーションが外部知識を活用して推論を行う RAG (Retrieval-Augmented Generation)[1] 構成の導入が進んでいる．RAG は，大規模言語モデル (LLM) の生成能力に，外部知識の検索機構を組み合わせた構成である．この構成は業務文書やマニュアルなどの内部知識を LLM 外のデータベースである VectorDB に格納し，LLM が VectorDB を検索的に参照することを可能とする．その結果，LLM は VectorDB から得た情報も動的に取り込み，より正確かつ根拠のある応答を生成することができる．企業における通常の利用形態としては，AI アプリケーションとデータベースの運用を同一企業が行う形が一般的である．

1.1 研究モチベーション

本研究が焦点を当てるのは，機密性の高いデータを外部に開示できない環境での AI 活用である．例えば，製造業分野の機密情報やヘルスケア分野の個人医療データなどは，AI 事業者に直接提供することができない．しかし，これらの分野も AI による検索・推論などの知的支援が強く求められており，データを公開せず AI が活用できる構成が必要である．このような機密性の高いデータを扱う分野においても AI 活用を実現するため，我々は TEE (Trusted Execution Environment)[2]とアテステーション[3]を活用した構成を検討している[7]．TEE は CPU レベルで分離された安全実行環境であり，外部ソフトウェアや管理者からの干渉を防止できる．そのため機密性の高いデータを TEE 環境に自動的に取り込み，自動的に AI アプリケーションを動作させ，TEE 環境から取得可能な情報は実行結果のみとすることで，機密性のあるデータを他社の目に触れさせることなく，データ漏洩・改変，外部委託者による不正操作などのリスクを低減させた AI 活用が可能となる．アテステーションは，TEE 内で動作するソフトウェアに対する検証のために利用する．

しかし，RAG を用いる場合には本構成においても課題がある．RAG を用いる場合一般的にチューニング操作が必要となる．チューニング操作とは，Embedding モデルの更新や VectorDB インデックスの再構築，更には学習データの再投入などが含まれる．そのため，チューニング操作を実行すると本来本構成で防いでいたはず他者によるデータ操作が発生してしまい，データ漏洩・改変，外部委託者による不正操作などのリスクが生じてしまう．更に，安全にチューニング操作がなされたとしても，データ提供者が意図した範囲で行われたか否かを，データ操作者が客観的に保証することは難しい．これらの操作は RAG の検索挙動や AI 応答内容に直接影響を与えるため，データ提供者が意図しない変更を検知・制御する仕組みが求められる．本研究では，VectorDB の登録・チューニング・監査を安全かつ検証可能に実行する方式を提案する．本提案は，AI アプリケーションが利用する VectorDB 自体の運用信頼性を確保することを目的とする．この構成を基盤として，次章では RAG 環境を想定した具体的な利用モデルと技術課題を整理する．

2. RAG 環境の想定構成と課題

2.1 想定構成

企業が外部 AI システムを活用して自社データを処理する構成では，データ提供者・利用者・外部委託者の間で，「誰が」「どのデータを」「どのように登録・調整したか」を把握する必要がある．このため，データ・AI の真正性とデータ・AI が意図通りに使われたかを正確に検証可能とする仕組みが求められる．本研究では，企業 A が保有する機密データを外部ベンダ B が構築した AI アプリケーションで活用するケースを想定する．図 1 に構成を示す．この構成では，データの登録，VectorDB の構築・チューニング，及び AI による利用がそれぞれ異なる主体により実施される．

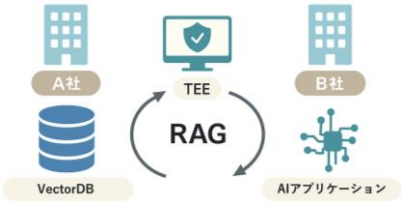

図 1　想定したマルチパーティ構成

2.2 課題

このようなマルチパーティ構成において，従来のアクセス制御や暗号化技術では操作の正当性を保証することが難しい．操作の正当性とは，「誰が」「どの環境で」「どのような手順に基づいて」チューニングを行ったかを客観的に検証できることを指す．アクセス制御では，操作主体に対する認可や権限付与は可能である物の，認可後に実際に実行された処理内容の妥当性までは把握できない．暗号化技術では，データの秘匿性や改竄防止には有効であるが暗号化されたデータに対してどのような処理が行われたかを外部から検証することは困難である．そのため，内部不正や外部委託先による不適切な操作を防止しつつ，データ提供者が外部監査を通じて操作履歴を確認できる仕組みが必要となり，下記の課題が存在する．

K1:　VectorDB へのデータ登録経路の信頼確保と権限制御

†1 NTT 株式会社
a) shinu.jo@ntt.com

K2: チューニング作業の安全化と操作制限
K3: データ登録・チューニング・操作の完全な証跡管理
K4: 外部監査者による検証可能性の実現

3. 提案方式

3.1 設計方針

本研究で前文の課題において，VectorDB の登録・チューニング・利用処理を TEE 内部で完結させる構成を提案する．各操作時に改竄検知可能な証跡を生成し，アテステーションによって構成の正当性を保証することで，データ提供者や監査者は外部委託者を完全に信頼することなく，操作の妥当性を検証できる．

3.2 システム構成

提案システムは，図２が示すモジュールで構成される．

1. データ登録アプリ(例 ETL):管理者の承認を得て社内データベースから VectorDB へ安全にデータを登録する．登録操作毎に TEE 内部で改竄検知可能な証跡を自動生成，登録経路および権限制御をハードウェアレベルで保証する．これにより，外部委託者や第三者による不正データ挿入を防止し，K1：登録経路の信頼確保と権限制御に対応する．
2. チューニングアプリ：外部委託者は TEE が提供する API 経由でのみ VectorDB の調整を実行できる．不正な操作要求は TEE によって自動的に拒否，すべての操作ログは署名付きで保管する．これにより，K2：チューニング作業の安全化と操作制限を実現し，意図しないデータ更新や不正な改変を防ぐ．
3. VectorDB 本体：VectorDB 内部はすべての操作証跡を取る．各証跡にタイムスタンプ，操作対象，実行者 ID，署名情報を含め，改竄耐性のある構造として管理する．これにより，K3：完全な証跡管理が可能となり，外部検証に供する信頼性の高い監査証跡を生成する．
4. 構成証明モジュール：TEE 内部におけるアプリ構成情報及び操作証跡を基に，アテステーションレポートを生成する．このレポートは外部監査者が受領し，操作の正当性・システム状態の完全性を検証する．これにより，K4：外部監査者による検証可能性を実現する．

TEE 環境内で上記モジュールを統合運用することで，登録・チューニングなど各工程をエンドツーエンドで保護する．TEE のハードウェア分離と暗号化証跡管理により，外部からの改変や内部不正を防止し，AI アプリケーションが利用する VectorDB の信頼性を総合的に担保する．

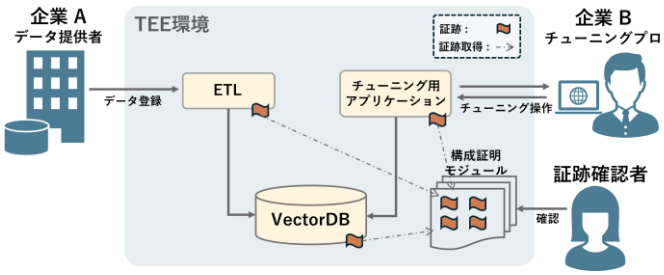

図 2 提案システム

4. 関連研究

RAG 環境におけるデータ登録やチューニングの信頼性を確保するため，これまでにさまざまな技術的アプローチが提案されている．代表的な方向性として，TEE を活用した実行環境の保護やブロックチェーン・暗号化構造を応用したデータ整合性の保証などが挙げられる．以下では，これらの既存研究の中から本提案に関連する主要な研究を概説する．SecCask［4］は，データ分析パイプライン全体を TEE 上に構築し，データ登録から処理までの経路をハードウェアレベルで保護する構成を提案した．また Atlas［5］は，機械学習ライフサイクル全体の由来情報を TEE 上で管理し，チューニング工程を含む処理の完全性と検証可能性を確保している．これらは部分的に本研究の K1・K2 課題（登録経路の信頼化とチューニング操作の監査）を解決する物と位置づけられる．一方，暗号化データベース技術によるデータ完全性保証の研究も進展している．VeritasDB［6］は，Merkle 木構造を利用してクエリ結果やトランザクション履歴の正当性を第三者が検証可能にした．これらは静的データの整合性と履歴検証（K3・K4 課題）を強化するが，操作過程その物の妥当性までは対象としていない．

本研究は，TEE を基盤とした処理環境で暗号化証跡とアテステーションを組み合わせることにより，登録・チューニングなど RAG 運用全工程の正当性を動的に保証する．既存研究が静的整合性保証に留まるのに対し，本研究は動的操作の検証可能性を実現する点で新規性を有する．

5. まとめ及び今後の展望

本研究では，TEE を活用した RAG 用 VectorDB の登録・チューニング･監査方式を提案した. TEE 上で全ての操作を実行し，暗号化証跡と構成証明を付与することで，データ提供者・外部委託者・利用者の分離環境下でも安全かつ検証可能な RAG 基盤を実現する．今後は，プロトタイプ実装を通じて暗号化処理やアテステーション機構の性能オーバーヘッドを評価し，複数 TEE 間の Federated RAG 構成の検討を進める．また，企業間での安全な AI データ共有･利用における標準モデル化を目指す．

参考文献

[1] Lewis, P., et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” NeurIPS, 2020.
[2] Sabt, M., Achemlal, M., Bouabdallah, A. “Trusted Execution Environment: What It Is, and What It Is Not.” IEEE Trustcom, 2015.
[3] Scarlata, V., et al. “Intel® Software Guard Extensions (Intel® SGX) Attestation Technical Overview.” Intel White Paper, 2018.
[4] Ooi, Y., et al. “SecCask: A TEE-based Data Analytics Pipeline Management System.” PVLDB, 2023.
[5] Intel Labs, “Atlas: Framework for Attestable ML Pipelines,” 2025.
[6] Li, F., et al. “VeritasDB: A Verifiable Database with Efficient Proof Generation.” ICDE, 2022.
[7] 石倉 禅 他，「Data Sandbox による Multi-Party Confidential Computing 実現方式の提案」，コンピュータセキュリティシンポジウム 2024 論文集, 2024.