グラフRAGに対するメンバーシップ推論攻撃評価

東 拓矢[1,a)]　中井 綱人[1]

概要：大規模言語モデル（LLM）のハルシネーションを抑制し，専門知識に基づいた回答生成を可能にする技術として，Retrieval-Augmented Generation（RAG）が広く利用されている．特に，知識を構造的に表現できる知識グラフを活用したグラフ RAG は，その高い検索精度と文脈理解能力から注目を集めている．しかし，LLM や RAG システムが学習・参照するデータには機微情報が含まれる可能性があり，プライバシー漏洩のリスクが懸念される．近年では，プライバシーリスク評価のためにメンバーシップ推論攻撃 (MIA) を用いることが増えている．先行研究では，LLM 単体やベクトル検索を用いた標準的な RAG に対するプライバシーリスク評価は行われてきたが，グラフ RAG を対象とした評価は未だ行われていない．本研究は，グラフ RAG に対する MIA の可能性を評価する初めての試みである．我々は，グラフ RAG を構築し，その知識グラフに含まれる特定の情報が，生成される回答から推論可能か否かを検証した．3 つの LLM と 2 つのデータセットを用いた実験により，生成される回答から知識グラフに含まれる情報が漏洩することを明らかにした．また，本研究の実験では，回答生成時に，知識グラフから LLM に提供される情報を増加させると，回答性能は上昇するが，MIA によるプライバシーリスクは減少する結果を得た．これは，従来、LLM の回答性能とプライバシーリスクはトレードオフとされた傾向とは反する．つまり，グラフ RAG においては，回答性能とプライバシーリスクが単純なトレードオフではないことを明らかにした．

キーワード：グラフ RAG， メンバーシップ推論， 大規模言語モデル

Membership Inference Attacks Against Graph-Based RAG

Takuya Higashi[1,a)]　Tsunato Nakai[1]

Abstract: Retrieval-Augmented Generation (RAG) has been widely adopted as a technique to mitigate hallucinations in large language models (LLMs) and facilitate the generation of responses grounded in specialized knowledge. In particular, graph-based RAG, which leverages knowledge graphs capable of structurally representing information, has garnered attention due to its high search accuracy and contextual understanding capabilities. However, the data that LLMs and RAG systems learn from and reference may contain sensitive information, raising concerns about the risk of privacy breaches. Recently, there has been an increasing trend to utilize Membership Inference Attacks (MIA) for privacy risk assessment. While prior research has evaluated the privacy risks associated with standalone LLMs and standard RAG systems employing vector search, assessments specifically targeting graph-based RAG have yet to be conducted. This study represents the first attempt to evaluate the potential for membership inference attacks on graph-based RAG. We constructed a graph-based RAG and investigated whether specific information contained within its knowledge graph could be inferred from the generated responses. Experiments conducted using three LLMs and two datasets revealed that information from the knowledge graph could be leaked through prompts. Furthermore, it has traditionally been posited that there exists a trade-off between the response performance of LLMs and privacy risks. However, our experimental results indicate that increasing the information provided to the LLM during response generation enhances response performance while simultaneously reducing the True Positive Rate (TPR) of the MIA. This finding suggests that, in the context of graph-based RAG, the relationship between response performance and privacy risk is not merely a straightforward trade-off.

Keywords: Graph-Based RAG, Membership Inference, Large Language Model

© 2025 Information Processing Society of Japan

This paper is work in progress and not peer-reviewed.

1. はじめに

近年，大規模言語モデル（LLM）は，自然言語処理技術の飛躍的な発展を牽引し，人間との対話，文章生成，翻訳，要約といった多様なタスクにおいて，目覚ましい性能を発揮している．その応用範囲は，一般的な情報検索から，医療，金融，法務といった高度な専門分野にまで及び，社会に大きな変革をもたらす可能性を秘めている．しかし，この技術的進歩の裏で，LLM が抱える根源的な課題であるハルシネーションが、深刻な問題として浮上している．ハルシネーションは，LLM が学習データに含まれていない情報や，事実に基づかない内容を，あたかも真実であるかのように生成する現象である．これは，モデルが学習データの統計的パターンに基づいて応答を生成する仕組みに起因し，特に最新情報や専門性の高いトピックにおいて顕著に現れる．この問題は，LLM が生成する情報の信頼性を著しく損ない，社会的な誤情報拡散のリスクを高めるため，その実用化における大きな障壁となっている．

このハルシネーション問題を解決する有力な手法として，検索拡張生成（RAG）が広く採用されている．RAG は，LLM が回答を生成するプロセスに，外部の信頼できる知識ソースからの情報検索を組み込んだフレームワークである [1]．具体的には，ユーザーからの入力に対し，まずベクトルデータベースなどの知識ソースから関連性の高い情報を検索し，その情報をプロンプトに含めて LLM に渡すことで，LLM は検索結果を根拠として回答を生成する．これにより，事実に基づいた正確かつ信頼性の高い応答が可能となり，ハルシネーションを大幅に抑制することができる．

さらに，RAG の発展形として，知識を構造化して管理する知識グラフを外部ソースとして活用するグラフ RAG が大きな注目を集めている [2]．知識グラフは，エンティティをノード，エンティティ間の関係性をエッジとして表現するグラフ構造のデータベースである．従来の非構造化テキストのデータベースとは異なり，エンティティ間の複雑で多層的な関係性を明確に捉えることができる．グラフ RAG は，この構造化された知識を活用することで，単純なキーワード検索では見落とされがちな文脈を深く理解し，より論理的で精度の高い情報検索と回答生成を実現する．これにより，因果関係の推論や複雑な質問への応答など，より高度なタスクへの応用が期待される．

RAG およびグラフ RAG は，LLM の信頼性を向上させる強力な手法である一方，新たなセキュリティリスク，特にプライバシー漏洩のリスクをもたらす．RAG システムが参照する知識ソースには，個人の病歴や連絡先といったセンシティブな個人情報，あるいは企業の未公開財務データや研究開発情報といった機密情報が含まれる可能性がある．これらの情報が不適切に扱われれば，深刻なプライバシー侵害や経済的損害につながる恐れがある．

最近では，プライバシーリスク評価のツールとしてメンバーシップ推論攻撃（MIA）が用いられている [3]．MIA は，攻撃者がモデルの出力を観測・分析することによって，ある特定のデータがモデルの学習データが含まれているか否かを推論する攻撃手法である．RAG の場合は参照するデータベースに特定の情報が含まれているか否かを推論することも可能である．この攻撃の成功によって，本来非公開であるべき情報の存在が間接的に暴露されてしまう危険性がある．

先行研究では，ベクトル検索をベースとした標準的な RAG システムに対して MIA を用いたプライバシーリスク評価が行われている [4] [5] [6] [7]．しかし，知識グラフを利用するグラフ RAG に対するプライバシーリスクについては，これまで体系的な評価が行われてこなかった．グラフ構造は，ノード間の連結性やパス，近傍情報といった独自の特性を持っており，これが MIA に対して新たな攻撃経路を提供する可能性がある．このように，グラフ RAG のプライバシーリスクは，従来の RAG とは異なる観点からの分析が不可欠である．

本研究では，このようなギャップを埋めることを目的とし，グラフ RAG システムにおけるプライバシーリスクを MIA の観点から初めて評価した．

本研究の主な貢献は以下の通りである．

- RAG の発展形であるグラフ RAG に対して，MIA によるプライバシーリスク評価を行った最初の研究である．
- 3 種類の LLM を用いた実験により，グラフ RAG が外部データベースとして利用する知識グラフに含まれる情報が MIA によって漏洩することを明らかにした．
- 従来，回答性能が上昇するとプライバシーリスクも上昇すると考えられてきたが，プライバシーリスク評価結果の解析により，グラフ RAG の回答性能とプライバシーリスクの関係が単純なトレードオフではないことを明らかにした．

2. 準備

本節では，研究の理解を深めるために，グラフ RAG と MIA の概要を記述する．

2.1 グラフ RAG

RAG は，LLM が応答を生成する前に，外部の検証可能な知識ソースから関連情報を検索し，その情報を根拠として利用する [1]．このアプローチにより，LLM は自身の内

[1] 三菱電機株式会社 情報技術総合研究所
Information Technology R & D Center, Mitsubishi Electric Corporation
[a)] Higashi. Takuya@da. MitsubishiElectric. co. jp

© 2025 Information Processing Society of Japan

部知識だけに頼るのではなく，信頼性の高い外部データに基づいて応答を生成することが可能となり，出力の事実性，信頼性，そして文脈的妥当性が大幅に向上する．

RAG の発展形であるグラフ RAG は，非構造化テキストの代わりに知識グラフを外部データベースとして利用し，LLM の能力を向上させる技術である [8]．これにより，エンティティ間の複雑な関係性を捉えた推論が可能となり，より正確で文脈に沿った回答の生成や，多角的な分析を必要とする高度な質問応答システムの構築が可能である．

グラフ RAG では，検索の情報源となるグラフデータベースにエンティティとそれらの間の関係からなる知識グラフが格納される．このとき，元のデータセットから情報が抽出され，ノードとエッジとして構造化された後，グラフデータベースにインデックスとして保存される．

グラフ RAG の特徴は，ユーザープロンプトと関連する情報を知識グラフから効率的に引き出すグラフ検索メカニズムである．ユーザープロンプトに含まれるエンティティを特定し，それを起点としてグラフ上を探索する．これにより，単純なキーワードの一致だけでなく，エンティティ間の直接的・間接的なつながりやパス，近傍の情報を体系的に収集することが可能である．この構造化情報に対する検索により，無関係な情報がノイズとして混入することを防ぎ，高度な質問に対する的確な情報の抽出が可能である．

グラフ RAG システムは通常，検索と生成の 2 つのフェーズで構成される．検索フェーズでは，グラフデータベースにクエリを入力し，ユーザーの質問に関連するエンティティや関係性を見つけ出す．ユーザープロンプトが入力されると，まずプロンプト内の主要なエンティティが特定される．その後，グラフ探索アルゴリズムを用いて，特定されたエンティティに関連するサブグラフや情報パスが取得される．グラフデータベースから取得された構造化データが LLM による生成フェーズに提供される．

生成フェーズでは，LLM がグラフデータベースから取得した情報に基づいて回答を生成する．検索フェーズからの情報が受け渡され，回答が生成される．

2.2 メンバーシップ推論

MIA は，プライバシーの脅威の一種であり，攻撃者が特定のデータレコードが機械学習モデルのトレーニングセットに使用されたかどうかを判定するものである [6]．これは，モデルが個人データを直接公開しなくても，個人に関する敏感な情報を明らかにする可能性があるため，重要なプライバシーリスクを持つ．

正確には，攻撃者はサンプル x がターゲットモデル m のトレーニングデータ D_m に属しているかどうかを判断する．つまり，$x \in D_m$ であるかを確認する．通常，このような攻撃は，サンプルがトレーニングセットの一部である確率を反映するターゲットモデルの出力に対して，エントロピーや対数確率などの 1 つ以上のメトリクスを計算する [3]．各サンプルに対してこのようなメトリクスの計算を経て，最終的にサンプルがトレーニングセットのメンバーである確率を出力する．

RAG の文脈において，メンバーシップ推論は，モデル m のトレーニングデータセット D_m におけるサンプルのメンバーシップまたは文書の検索データベース D_b におけるメンバーシップに分類できる．本論文は後者に焦点を当てている．攻撃の目的は，ターゲット文書 d が検索データベース D_b に属しているかどうかを推測することである．つまり，$d \in D_b$ であるかを推測することである．

3. 関連研究

本節では，関連研究として RAG に対する MIA 評価を実施したものと，グラフ RAG に関する先行研究をまとめる．

3.1 RAG のメンバーシップ推論攻撃評価

Li らは，ターゲットサンプルがデータベース内に存在する場合，RAG が生成するテキストとの意味的類似性が高くなり，perplexity が低くなる特徴を利用してメンバーシップを推論する S^2MIA を提案した [4]．Anderson らは，ターゲットとなる文書を含んだ上で「このデータベースに含まれるか？」と直接尋ねる特殊なプロンプトを送信し，システムの「Yes/No」という応答からメンバーシップを特定する手法を提案した [5]．Liu らは，対象文書の一部の単語をマスクし，RAG システムに予測させる手法 Mask-Based Membership Inference Attacks (MBA) を提案した [6]．文書がデータベース内にあれば，RAG は高精度で単語を予測できるため，その正解率によって文書の有無を特定する．Naseh らは，検知されにくい自然な質問を生成し，RAG システムのデータベースに特定の文書が存在するかどうかを推論する攻撃手法である Interrogation Attack (IA) を提案した [7]．攻撃者は，標的文書に特化した質問と正解を準備し，それを標的の RAG システムに問い合わせ，回答と準備した正解の一致度をスコア化することで，文書の有無を判定する．

3.2 グラフ RAG

グラフ RAG は，原文テキストから知識グラフを構築し，問い合わせに対する回答性能を向上させる手法として提案され，大きな注目を集めている [2][9][10]．この手法は LLM の能力を活用して文書から知識グラフを構築し，次にノードをクラスタリングして各クラスターをより高レベルのコミュニティへと要約することで，多粒度の知識グラフを形成する．しかし，このアプローチは多数の LLM 呼び出しを必要とするため，インデックス作成の段階で非常に高いコストがかかる．さらに，全体検索においては，クエリに関連するコミュニティを判断するために LLM を使用する

© 2025 Information Processing Society of Japan

ため，著しい遅延と計算オーバーヘッドが発生する．

Sarthi らは，エンティティグラフを抽出する代わりに，チャンク（断片）を再帰的にクラスタリングおよび要約することで階層的な要約ツリーを構築する手法である RAPTOR を提案している [11]．これにより，LLM を用いたエンティティと関係の抽出という複雑なプロセスを回避する．しかし，この手法は元の文書の文脈の流れを無視し，クラスタリング処理にも時間がかかる．さらに，RAPTOR は従来の RAG のようなベクトルベースの検索を採用しているため，不正確な検索結果を招く可能性がある [12]．

グラフ RAG の効率を改善するために，LightRAG[13] や Fast-GraphRAG[14] といった近年の手法は，クラスタリングとコミュニティ要約のプロセスを省くことで，グラフ RAG における高コストなインデックス作成を削減している．Guo らが提案した LightRAG は，LLM に直接プロンプトを与えて各チャンクから多粒度のエンティティと関係を 1 回の処理で抽出し，検索時の直接的なマッチングを可能にする．Chen らが提案した FastGraphRAG は，インデックス作成時には同様の抽出戦略を採用するが，検索時には PageRank[15] の変種を適用することで，コミュニティ構造に依存しない全体検索をサポートする．どちらのアプローチもある程度インデックス作成のオーバーヘッドを削減するが，依然としてチャンクごとに複雑で冗長な JSON 形式の出力を LLM に生成させる必要があり，かなりの時間と計算コストを要する．Zhao らが提案した E2GraphRAG は，LLM による要約ツリーと，計算コストの低い NLP ツール SpaCy によるエンティティグラフという 2 つの構造を構築し，クエリに応じて検索モードを自動で切り替える適応型検索戦略を用いることで，高い効率と回答性能を両立させた [16]．本研究では，評価対象として E2GraphRAG を使用する．

3.3 本研究

本研究は，グラフ RAG を対象にプライバシーリスク評価を初めて実施したものである．現在，RAG システムに対しては，埋め込まれた LLM や参照するデータベースを対象としたプライバシーリスク評価が MIA によって実施されている．しかし，RAG の発展形であるグラフ RAG に対しては，MIA 評価が実施されていない状況である．このような研究ギャップを埋めるために，本稿では Zhao らが提案したグラフ RAG である E2GraphRAG のデータベースを対象として MIA を用いたプライバシーリスク評価を実施した．

4. 評価手法

本研究では，グラフ RAG に対するプライバシーリスク評価を実施した．本節では，評価時の設定や，評価対象のシステムに関して記述する．

4.1 評価時の設定

本研究における評価の概要を図 1 に示す．攻撃者がグラフ RAG システムに対して特定の文章がデータベースに存在するかを Yes/No で回答させることでデータベースのメンバーシップ推論を実施する．

本研究では，使用する MIA に合わせて先行研究 [5] と同様に MIA 評価時の前提を定義する．ブラックボックスとグレーボックス 2 つの設定でプライバシーリスク評価を実施した．ブラックボックス設定では，攻撃者がユーザープロンプトと RAG システムから生成された出力のみへのアクセスを持つ．攻撃者はユーザープロンプトを変更することができるが，基盤となる LLM やグラフデータベース，またはこれらに使用されるプロンプトテンプレートについての知識は持っていない．さらに，攻撃者は使用されるデータベースの種類に関する情報も持っていない．

グレーボックス設定では，LLM のパラメータや勾配など，モデル内部の情報にアクセスできない点は同様である．ブラックボックス設定に加えて，回答生成時の各単語の確率分布（ロジット）を取得できるものとする．

4.2 評価対象のシステム

本研究では，Yibo らが提案した E2GraphRAG を評価対象とする [16]．このグラフ RAG は，効率性と回答性能の向上を目的として提案されたグラフ RAG である．他のグラフ RAG と同様に，E2GraphRAG も索引付けによってデータベースを作成したのちに検索・回答生成が行われる．索引付け段階において，本システムはドキュメントから 2 種類のデータ構造を構築する．一つは，LLM を用いてドキュメントのチャンクを再帰的に要約することで生成される階層的な要約ツリーである．もう一つは，エンティティ抽出に LLM ではなく，計算コストの低い NLP ツールである SpaCy を利用して構築されるエンティティグラフである．このアプローチにより，索引付けの処理速度が大幅に向上する．構築した要約ツリーとエンティティグラフを効率的に連携させるため，エンティティからチャンクへおよびチャンクからエンティティへのマッピングを保持する双方向インデックスを構築する．これにより，検索段階における迅速な参照が可能となる．

検索段階では，従来と異なりクエリの特性に応じて検索方式をローカル検索とグローバル検索のいずれかのモードから自動で選択する．検索時には最大ホップ数がパラメータとして設定可能である．これは，質問に答えるために知識グラフ上の情報をどれだけ深く，広く探索するかを示す尺度である．グラフのノードから次のノードへ，エッジをたどる 1 回の移動を 1 ホップと数える．例えば，ホップ数を 2 に設定した場合，クエリ内のエンティティの組み合わせのうち，ホップ数が 2 以上のものはその後のプロセスから排除される．

© 2025 Information Processing Society of Japan

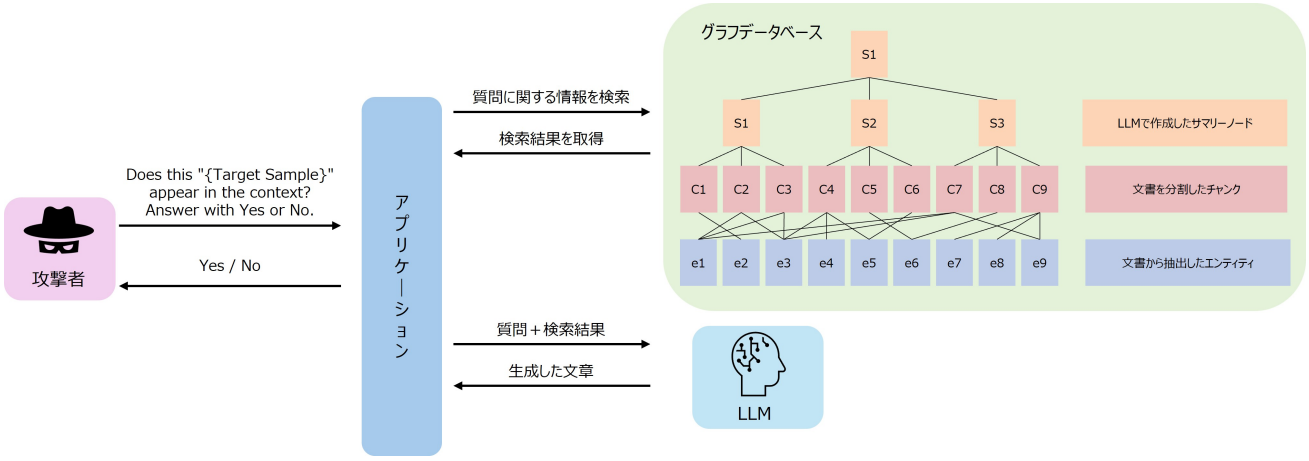

図 1　評価の概要
Fig. 1　Evaluation overview.

クエリ内のエンティティ同士の関係がグラフ上で密接であると判断された場合，ローカル検索が実行される．双方向インデックスを利用して，関連エンティティが共起するチャンクを特定する．特定されたチャンクは，コサイン類似度に基づいてランキング付けされる．LLM には類似度が上位のチャンクから入力され，最大チャンク数のパラメータによって入力されるチャンクの数が制御できる．LLM は，入力されたチャンクの情報を使用してクエリに対する回答を生成する．

4.3　評価に使用するメンバーシップ推論攻撃

本研究では，Maya らによって提案された RAG に対する MIA を用いてグラフ RAG のプライバシーリスク評価を行った．この MIA では攻撃プロンプトをを RAG システムにクエリとして入力することで，特定のデータが検索データベース内に存在するかどうかを推測するというものである．

ブラックボックス設定では，攻撃プロンプトに対するシステムの応答が「Yes」か「No」かというテキスト自体を直接利用して，メンバーか非メンバーかを判定する．メンバーサンプルは，攻撃対象の文書から抽出し，非メンバーサンプルは他の文書から抽出した文章である．サンプル数はそれぞれ 100 ずつとした．

グレーボックス攻撃では，予測信頼度に基づいてメンバーか非メンバーかを判定する．具体的には，Yes と No それぞれの単語のロジットから信頼度を算出する．算出した値が閾値以上であればメンバー，閾値より低ければメンバーでないと判定される．この手法では誤検出率（FPR）を一定の割合に抑えた時に，どれだけ正しく判定できるかを評価可能である．プライバシーは平均的なケースではなく，最も脆弱なケースで評価されるべきであり，低い誤検出率では，攻撃者がどれだけ確実にメンバーを見つけ出せるかという，より現実的で深刻な脅威を評価できる [3]．本研究においても，低い FPR での真陽性率（TPR）を測定し，より現実的な脅威を評価する．

5.　実験

本研究では，グラフ RAG に対するプライバシーリスク評価を実施した．本節では，実験環境，実験結果とその考察を記述する．

5.1　実験環境

本研究では，2 つのデータセットを用いて実験を行った．Infinite Bench から英語の選択肢問題と英語 QA という 2 つのサブセットを選び，それぞれ Infinite Choice および Infinite QA と呼ぶ．Infinite Choice は，合計 58 冊の書籍と 229 の選択肢問題で構成される．Infinite QA は，合計 20 冊の書籍と 102 の質問で構成される．

LLM は，Qwen-2-7B，Llama-3-8B，Mistral-7B を量子化したものを使用した．

システムの回答性能の評価には，Accuracy と ROUGE-L を使用する．Accuracy は正解の選択肢と回答した選択肢が合っていた割合を示す．ROUGE-L は，生成されたテキストと正解テキストの最長共通部分列を基に，品質を評価する指標である．正解テキストに含まれる単語がどれだけ生成されたテキストに含まれるかを考慮し，スコアが高いほど生成されたテキストが正解テキストに近いことを示す．

MIA の評価指標には，TPR と FPR，Accuracy，AUC-ROC を使用した．TPR はデータベースに存在する情報を，正しくメンバーであると推測した割合を示す．FPR は，データベースに存在しない情報を，誤ってメンバーであると推測した割合を示す．Accuracy は推論全体のうち，正しかった割合を示す．AUC-ROC は，ROC カーブの曲線下面積である．この値が 1 に近いほど推論性能が高く，0.5 の時ランダムな推測と同様であることを示す．上記の指標のうち，ブラックボックス設定では Accuracy，TPR，

© 2025 Information Processing Society of Japan

表 1　Qwen 使用時の性能評価と MIA 評価の結果
Table 1　Score and MIA Recall when using Qwen.

Dataset	Chunk	Hop	Accuracy/ ROUGE-L	Membership Inference Attacks						
				Black-Box			Gray-Box			
				Accuracy	TPR	FPR	AUC-ROC	TPR@FPR=0.1	TPR@FPR=1	TPR@FPR=10
Infinit Choice	1	1	34.09	69.57	59.43	20.29	77.91	22.29	22.29	50.57
	1	8	50.00	84.00	76.86	8.86	90.48	50.86	50.86	78.00
	3	1	40.91	73.86	67.71	20.0	79.99	31.14	31.14	56.86
	3	4	45.45	80.71	82.29	20.86	**87.45**	**42.29**	**42.29**	**69.71**
	3	8	45.45	**79.71**	**81.43**	22.00	87.67	40.57	40.57	70.29
	5	1	43.18	70.86	72.86	31.14	79.67	**27.71**	**27.71**	**56.57**
	5	4	43.18	76.57	87.43	34.29	87.96	46.86	46.86	71.14
	5	8	43.18	76.43	**86.29**	33.43	88.21	42.57	42.57	72.86
Infinit QA Loader	1	1	6.51	64.64	55.82	26.55	72.99	14.55	14.55	35.09
	1	4	4.40	81.91	72.55	8.73	89.15	44.36	44.36	76.55
	1	8	5.12	82.27	73.27	8.73	89.18	44.36	44.36	76.55
	3	1	5.79	71.64	65.27	22.00	79.58	**13.27**	**13.27**	50.91
	3	4	7.00	84.27	80.36	11.82	89.97	36.18	36.18	78.00
	3	8	5.44	83.18	79.27	12.91	89.89	36.00	36.00	77.27
	5	1	5.59	74.82	72.18	22.55	82.21	15.64	15.64	60.00
	5	4	7.94	**81.09**	84.36	22.18	39.64	39.64	39.64	73.82
	5	8	8.16	**81.00**	**84.00**	22.00	**89.21**	**38.55**	**38.55**	77.45

表 2　Llama 使用時の性能評価と MIA 評価の結果
Table 2　Score and MIA Recall when using Llama.

Dataset	Chunk	Hop	Accuracy/ ROUGE-L	Membership Inference Attacks						
				Black-Box			Gray-Box			
				Accuracy	TPR	FPR	AUC-ROC	TPR@FPR=0.1	TPR@FPR=1	TPR@FPR=10
Infinit Choice	1	1	40.91	67.00	66.57	32.57	71.23	12.86	12.86	36.00
	1	4	43.18	80.00	84.29	24.29	88.02	39.43	39.43	73.14
	1	8	43.18	**79.71**	84.29	24.86	87.88	39.43	39.43	**72.57**
	3	1	36.36	58.71	51.43	34.00	63.46	10.57	10.57	30.86
	3	4	40.91	59.00	**51.14**	33.14	76.33	18.00	18.00	47.43
	3	8	40.91	**58.43**	51.43	34.57	77.60	**16.29**	**16.29**	49.71
	5	1	38.64	**55.14**	**39.71**	29.43	0.6005	5.43	5.43	24.57
	5	4	38.64	52.86	26.86	21.14	62.15	8.86	8.86	24.86
	5	8	38.64	**53.00**	27.43	21.43	62.24	9.43	9.43	24.86
Infinit QA Loader	1	1	6.07	62.73	58.73	33.27	67.58	8.00	8.00	28.55
	1	4	10.42	76.36	73.27	20.55	81.65	31.09	31.09	57.82
	1	8	9.35	76.55	73.45	20.36	81.55	31.09	31.09	58.00
	3	1	3.63	62.64	54.55	29.27	69.13	9.45	9.45	30.36
	3	4	9.87	62.55	50.73	25.64	76.06	17.64	17.64	46.00
	3	8	8.83	63.27	50.55	24.00	76.05	18.18	18.18	46.18
	5	1	5.70	**57.18**	38.73	24.36	62.93	6.73	6.73	21.27
	5	4	6.31	55.73	32.00	20.55	64.05	12.73	12.73	28.18
	5	8	6.66	56.27	**31.45**	18.91	64.15	13.09	13.09	28.00

FPR を使用した．グレーボックス設定では，FPR が 0.1，1，10 のときの TPR と AUC-ROC を使用した．

最大ホップ数を 1，4，8，最大チャンク数を 1，3，5 に設定し，それぞれの条件で MIA を実施した．最大ホップ数が大きいと，探索対象とするエンティティペア間の距離が大きくなるため，情報を探索する範囲が増える．最大チャンク数が大きいと LLM に入力されるチャンクの数が増える，つまり LLM が回答生成に使用する情報量が増える．

5.2　実験結果

本項では，3 つの LLM と 2 つのデータセットを用いた実験結果を述べる．それぞれの LLM における実験結果を

© 2025 Information Processing Society of Japan

表 3 Mistral 使用時の性能評価と MIA 評価の結果
Table 3 Score and MIA Recall when using Mistral.

Dataset	Chunk	Hop	Accuracy/ Rouge-L	Membership Inference Attacks						
				Black-Box			Gray-Box			
				Accuracy	TPR	FPR	AUC-ROC	TPR@FPR=0.1	TPR@FPR=1	TPR@FPR=10
Infinit Choice	1	1	15.91	66.57	40.86	7.71	72.35	33.43	33.43	4771
	1	4	25.00	84.57	76.00	6.86	89.45	66.29	66.29	8029
	1	8	25.00	84.43	75.71	6.86	89.42	66.29	66.29	8029
	3	1	22.73	69.86	57.14	17.43	77.11	38.29	38.29	5314
	3	4	29.55	78.43	77.14	20.29	87.81	49.43	49.43	7457
	3	8	29.55	78.71	77.43	20.00	87.49	48.86	48.86	74.00
	5	1	20.45	70.43	65.14	24.29	77.13	37.71	37.71	58.29
	5	4	25.00	77.14	83.71	29.43	88.43	48.29	48.29	76.29
	5	8	25.00	77.14	83.71	29.43	88.46	48.57	48.57	72.57
Infinit QA Loader	1	1	5.77	64.45	40.91	12.00	72.42	26.55	26.55	44.18
	1	4	7.99	81.18	71.09	8.73	88.36	58.36	58.36	76.73
	1	8	7.22	81.09	71.09	8.91	88.36	58.36	58.36	76.73
	3	1	6.41	69.73	56.36	16.91	77.79	32.91	32.91	56.18
	3	4	6.58	80.18	77.64	17.27	87.69	53.09	53.09	75.64
	3	8	5.89	79.91	77.45	17.64	87.57	52.36	52.36	75.64
	5	1	5.22	69.91	62.91	23.09	79.59	37.64	37.64	58.36
	5	4	6.45	78.00	80.18	24.18	88.00	4291	42.91	73.09
	5	8	6.75	77.64	80.55	25.27	88.04	46.00	46.00	73.82

表 1，2，3 に示す．回答性能の評価に Infinit Choice では Accuracy を使用し，Infinit QA Loader では ROUGE-L を使用した．MIA に関して、FPR=1 以下では同じ値の TPR であった。3 つのモデルすべてにおいて，MIA の性能指標である AUC-ROC が多くの設定で 0.7 から 0.9 程度の高い値を示した．これは，ランダムな推測（AUC-ROC=0.5）を大幅に上回っており，評価対象としたグラフ RAG システムから，攻撃者のプロンプトによって知識グラフ内のメンバー情報が漏洩するリスクが存在することを示す．

回答性能とパラメータ（ホップ数，最大チャンク数）の関係を解析した．ホップ数に関して，全てのモデルにおいて，ホップ数を 1 から 4 に増やすと，探索範囲が広がることで関連情報をより多く取得できるため，回答性能が向上する傾向が見られた．しかし，ホップ数を 4 から 8 に増やしても，性能向上は限定的であった．入力するチャンク数を増やした場合の影響は，モデルによって異なる傾向を示した．Qwen と Mistral では，チャンク数を増やすことで回答性能が向上する場合が多かった．しかし，Llama を用いた実験では，特に Infinite Choice データセットにおいて，チャンク数を増やすと逆に回答性能が低下する現象が確認された．

次に，プライバシーリスクとパラメータの関係を分析した．ホップ数に関して，回答性能と同様に，ホップ数を 1 から 4 に増やすと，MIA の成功率が向上する傾向が 3 つのモデルすべてで確認された．これは，探索範囲の拡大が，攻撃者にとっても有益な情報を与えることを示す．チャンク数に関して，Qwen を用いた実験では，チャンク数を増やすと回答性能は向上するが，ブラックボックス攻撃の TPR は増加する一方で，グレーボックス攻撃における低い FPR での TPR は必ずしも増加せず，特定の条件下で減少が確認された．Llama を用いた実験では，チャンク数を増やすと，回答性能が低下すると同時に，ブラックボックスおよびグレーボックス攻撃の TPR も明確に減少した．Mistral を用いた実験では，チャンク数増加で回答性能は向上するものの，MIA の TPR は Llama ほど明確な減少を示さなかったが，大幅なリスク増加も確認されなかった．

5.3 考察

実験から，チャンク数，ホップ数の変更により，グラフ RAG の回答性能と MIA の TPR がトレードオフの関係にない条件があることを明らかにした．この原因として回答性能評価と MIA のときのクエリの複雑さの違いが考えられる．今回実施した MIA は，指定した文章のデータベース内での有無を Yes/No で回答させるパターンマッチングのようなものである．これはほとんどの場合，単一のチャンクで回答に必要な情報が収集可能である．対して，回答性能評価では，質問が複雑であるため，複数のチャンクにまたがった情報を利用して回答する必要がある．これらの理由から，LLM に入力する情報量が増えると回答性能が向上するが，MIA 時には多すぎる情報がノイズとして働いていると考える．回答性能とプライバシーリスクが単純なトレードオフ関係にないため，最適なパラメータを使用す

© 2025 Information Processing Society of Japan

ることで性能を維持しながらプライバシーリスクを低減することが可能だと考えられる．

6. 議論

本節では，本研究における制限と緩和策を記述する．

6.1 制限

本研究で評価対象としたのは E2GraphRAG のみである．LightRAG[13] や Fast-GraphRAG[14] など他のグラフ RAG に対して同様の評価を行った場合，プライバシー漏洩リスクの傾向も変動する可能性がある．更に，今回使用した攻撃手法は，指定した文章が元のデータセットに含まれるかを Yes/No で直接問うものである．関連研究で言及した MBA[6] や IA[7] など，他の攻撃手法を使用した場合に同様の結果が得られるかは不明である．

6.2 緩和策

本研究の実験評価では，回答性能とプライバシーリスクが単純なトレードオフではないことが明らかにされた．この知見に基づき，システムの回答性能を大きく損なうことなく，MIA の成功率（TPR）が低くなるようなパラメータ（最大ホップ数，最大チャンク数）の組み合わせを意図的に選択することが，有効な緩和策となり得る．また，差分プライバシーに基づいたノイズの付与により，個々の情報がデータベースに存在するかどうかの情報を攻撃者に与えにくくすることが考えられる．

7. おわりに

本研究では，グラフ RAG に対するプライバシーリスク評価を実施した．3 つの LLM と 2 つのデータセットを使用した実験により，グラフ RAG の回答性能とプライバシーリスクが単純なトレードオフの関係にないことを明らかにした．このことから，最適なパラメータ設定によって回答性能を維持しながらプライバシーリスクを低減することが可能だと考えられる．

参考文献

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[2] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, “From local to global: A graph rag approach to query-focused summarization,” 2025. [Online]. Available: https://arxiv.org/abs/2404.16130

[3] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, “Membership inference attacks from first principles,” in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 1897–1914.

[4] Y. Li, G. Liu, C. Wang, and Y. Yang, “Generating is believing: Membership inference attacks against retrieval-augmented generation,” in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[5] M. Anderson, G. Amit, and A. Goldsteen, “Is my data in your retrieval database? membership inference attacks against retrieval augmented generation,” in *Proceedings of the 11th International Conference on Information Systems Security and Privacy*. SCITEPRESS - Science and Technology Publications, 2025, p. 474–485. [Online]. Available: http://dx.doi.org/10.5220/0013108300003899

[6] M. Liu, S. Zhang, and C. Long, “Mask-based membership inference attacks for retrieval-augmented generation,” in *Proceedings of the ACM on Web Conference 2025*, ser. WWW ’25. New York, NY, USA: Association for Computing Machinery, 2025, p. 2894–2907. [Online]. Available: https://doi.org/10.1145/3696410.3714771

[7] A. Naseh, Y. Peng, A. Suri, H. Chaudhari, A. Oprea, and A. Houmansadr, “Riddle me this! stealthy membership inference for retrieval-augmented generation,” 2025. [Online]. Available: https://arxiv.org/abs/2502.00306

[8] Y. Hu, Z. Lei, Z. Zhang, B. Pan, C. Ling, and L. Zhao, “Grag: Graph retrieval-augmented generation,” 2025. [Online]. Available: https://arxiv.org/abs/2405.16506

[9] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, and S. Tang, “Graph retrieval-augmented generation: A survey,” 2024. [Online]. Available: https://arxiv.org/abs/2408.08921

[10] Y. Zhou, Y. Su, Y. Sun, S. Wang, T. Wang, R. He, Y. Zhang, S. Liang, X. Liu, Y. Ma, and Y. Fang, “In-depth analysis of graph-based rag in a unified framework,” 2025. [Online]. Available: https://arxiv.org/abs/2503.04338

[11] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, and C. D. Manning, “Raptor: Recursive abstractive processing for tree-organized retrieval,” 2024. [Online]. Available: https://arxiv.org/abs/2401.18059

[12] F. Tian, D. Ganguly, and C. Macdonald, “Is relevance propagated from retriever to generator in rag?” 2025. [Online]. Available: https://arxiv.org/abs/2502.15025

[13] Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang, “Lightrag: Simple and fast retrieval-augmented generation,” 2025. [Online]. Available: https://arxiv.org/abs/2410.05779

[14] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation,” 2024. [Online]. Available: https://arxiv.org/abs/2402.03216

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web.” Stanford InfoLab, Technical Report 1999-66, November 1999, previous number = SIDL-WP-1999-0120. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[16] Y. Zhao, J. Zhu, Y. Guo, K. He, and X. Li, “E2graphrag: Streamlining graph-based rag for high efficiency and effectiveness,” 2025. [Online]. Available: https://arxiv.org/abs/2505.24226

© 2025 Information Processing Society of Japan