大規模言語モデルの利用における プライバシー保護の新たな視点

高田雅之 [1]　玉井睦 [1]
[1] セコム株式会社 IS 研究所
{masayu-takada, mu-tamai}@secom.co.jp

概要

ChatGPT の登場により，対話型生成 AI システムに対する期待が高まっている．一方で，LLM が組み込まれた AI システムの利用にはプライバシーリスクが伴う．LLM のプライバシー保護研究は，学習データやモデルに含まれる個人のプライバシーリスクに関する対策が主流であり，AI システム利用時のプライバシーリスクに関する検討は十分でない．本稿では，LLM の生成結果が利用者に出力される状況におけるプライバシーリスクとその低減の重要性を述べ，そのリスクに対するプライバシー保護の既存研究を示すとともに，今後研究すべき領域について提案する．

1　はじめに

ChatGPT の登場により，対話型生成 AI システムへの期待が高まっている．このようなシステムを支える大規模言語モデル（LLM）の進展は目覚ましく，LLM がスマートフォンやスマートスピーカー，ロボットなどに組み込まれることで，パーソナライズされた AI システム（以下，AI）が利用者を支援する世界が実現しつつある．一方で，LLM の利用にはリスクも伴い，LLM に関連する脅威の分類 [1] やガイドライン [2, 3] が作成されている．LLM は，学習データに個人のプライバシー情報が含まれる場合，その情報を生成する可能性があり，LLM の生成結果が利用者に出力される状況によっては，AI 利用者のプライバシーを侵害する場合がある．

LLM に関するプライバシー保護の研究は多数行われており，そのサーベイ論文も発表されている [4, 5, 6, 7]．しかし，これらのサーベイ論文は主に LLM の作成時やサーバー通信時のプライバシー侵害のリスク（プライバシーリスク）に焦点を当てており，LLM の生成結果が出力される状況による，AI 利用者のプライバシーリスクに焦点を当てたプライバシー保護の研究については十分に調査されていない．そこで本稿では，LLM の生成結果が出力される状況を考慮したプライバシー保護の研究を調査し，AI 利用者のプライバシーリスクに対応するための課題を考察する．

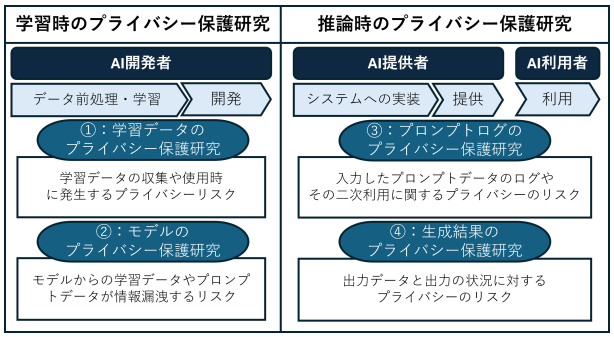

図 1　AI 活用フェーズで想定されるプライバシーリスクとプライバシー保護研究の関係

以降，2 節で LLM のプライバシーリスクの分類を行い，3 節で本稿で想定する状況と対象とするプライバシーリスクおよびプライバシーに配慮する[1)]ために考慮すべき視点についてマルチモーダル LLM に着目し述べる．4 節でそれら視点に関する既存研究を紹介し，5 節で取り組むべき課題について述べる．6 節ではまとめと展望について論じる．

2　LLM のプライバシーリスク分類

AI 事業者ガイドライン [2] で示された，一般的な AI 活用の流れにおける主体の対応を参考に，本稿では AI 活用の各フェーズで想定されるプライバシーリスクと，それに対応するプライバシー保護研究の関係を図 1 のように分類した．本稿では，AI 活用の流れを学習時と推論時の二つに分けた．そして，学習時には学習データやモデルに起因する，推論時に

1)　個人情報やプライベートな情報を適切に利用する意味．プライバシー保護は，個人情報やプライベートな情報の漏洩リスクを低減させる意味である．

はプロンプトのログ（プロンプトログ）[2]や生成結果に起因するプライバシーリスクがあるとして，4 つに分類した．

学習時のプライバシー保護研究は，①学習データや②モデルが対象である．例えば①学習データの研究では，学習データに含まれる個人情報などをマスキングする研究 [8] がある．また，②モデルの研究では，LLM が個人情報を生成することを防ぐ研究 [7] や，LLM 内に保持された個人情報を消去する研究 [5, 9] がある．推論時のプライバシー保護研究は，③プロンプトログや④生成結果が対象である．例えば③プロンプトログの研究では，プロンプトログに含まれる利用者のプライベートな情報が AI 提供者に出力されるリスクや二次利用されるリスクに対して，Differential Privacy を用いた研究 [6, 10] や Federated Learning を用いた研究 [11] が挙げられる．本稿で対象とする④生成結果のプライバシーリスクや，それに対応するプライバシー保護研究については次節以降で述べる．

3　生成結果のプライバシーリスク

3.1　想定する AI 利用時の状況

本稿では，LLM を実装した AI がスマートフォンやスマートスピーカー，ロボットなどに組み込まれ，パーソナライズされた AI エージェントとして人々の生活を支援する状況を想定する．このとき，AI は利用者の身近な生活空間に共存するシステムとして実装される．また，利用開始後に利用者が入力したプロンプトログや連携した他システムのデータによって AI 利用者にパーソナライズされ，利用者のプライベートな情報を記憶する可能性がある．AI と利用者の対話内容には，利用者のプライベートな情報が含まれることもある．対話のタイミングは，利用者が音声その他の入力によって開始することもあれば，生活空間に設置されたセンサやスケジューラーなどの他システムのデータをトリガーとして AI が開始することもある．AI と利用者のプロンプトログは，コミュニケーションアプリなどを通じて別の特定の人物に出力される場合がある．例えば，プロンプトログの閲覧権限を付与された利用者の家族や医療関係者に対して出力することが考えられる．また，このようなアプリなどを介さなくても，AI との対話が行われる空間に居合わせた人物に AI の音声として出力される場合もある（図 2）．

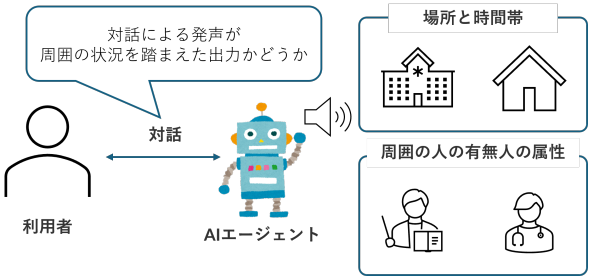

図 2　想定する周囲の状況

3.2　想定するプライバシーリスク

想定した利用状況において考えらえるプライバシーリスクは，③プロンプトログに関するものと，④生成結果に関するものがある（図 1）．本稿では③プロンプトログに起因するリスクは対策が施され，利用者に許容されているものとし対象外とする．④生成結果のプライバシーリスクは，LLM の生成結果が出力される状況が不適切な場合に起こるプライバシーリスクである．これは，悪意を持つ第三者からの攻撃による情報漏洩ではなく，通常の利用の中で起こる予期せぬ出力を想定する．例えば，AI が対話中に利用者のプライベートな情報を予期せず出力し，その場にいる第三者に知られてしまうケースである．こうしたリスクは，状況を考慮しない AI の応答や，利用者の問いかけではない特定のトリガーによる対話の開始が原因として考えられる．本稿ではこのような生成結果のプライバシーリスクを対象とする．

3.3　生成結果のプライバシー保護に必要な視点

学習時のプライバシー保護研究の多くは学習データやモデルに個人情報が含まれないようにする考え方 [5, 7, 8, 9] が基本であるが，AI 利用者の個人情報を学習することが前提のパーソナライズされた AI に対しては他の考え方が必要である．

この考え方の一つに Contextual Integrity（CI）理論 [12] がある．Nissenbaum によれば，適切性の規範と流通の規範のいずれかが違反された場合，プライバシーが侵害される．適切性の規範は，特定の状況でどのような個人に関する情報が明かされるのが適切であるかに係る．流通の規範は，特定の状況で期待される情報流通の在り方に係る．本稿では，生成結果のプライバシーリスクを評価するための基本とな

2）プロンプトログには，AI への指示のみではなく，AI との対話履歴や RAG などによる外部から与えられる情報も含む．

る考え方として CI 理論に着目する．既存研究では CI 理論のコンテキストを言語的な文脈として捉えた研究が主流である [13, 14, 15, 16, 17]．しかし，生成結果のプライバシーリスクである予期せぬ出力は，文章内だけに閉じて捉えられるものではなく，時間的な経緯や空間的な状況が大いに関わるため，これらの要素をコンテキストとして捉える必要がある．本稿では，これらを時間的コンテキスト，空間的コンテキストと呼ぶ．時間的コンテキストは，AI の発話までの経緯（プロンプトログや同意状態など）である．空間的コンテキストは，発話時の周囲の状況（他者の存在や場所の性質など）である．これらのコンテキストを考慮しプライバシーリスクを評価したうえで適切な生成結果を得るための手段として，本稿ではマルチモーダル LLM に着目する．

4 既存研究

LLM に CI 論理を用いた既存研究と，マルチモーダル LLM の既存研究を調査した．本節ではこれらの既存研究について述べる．

4.1 CI 理論と LLM の既存研究

Fan らの研究 [16] では，LLM を法的タスクに活用するための学習手法と，それに適した合成データセットの構築を提案している．この研究では，CI 理論に基づき，流通の規範を送信者と受信者の関係性と定義し，適切性の規範を送信する情報の内容と位置付けることで，プライバシー保護の適合性を判断する手法を提示している．このように，CI 理論を用いることで，LLM がプライバシー保護の適切性を評価できる枠組みを構築することを提案している．

Li らの研究 [17] では，プライバシー侵害の検出を目的に CI 理論を活用している．この研究は，プライバシー侵害を単なるパターンマッチングの問題として捉えるのではなく，推論の問題として再構築している．また，CI 理論に基づき，流通の規範を送信者と受信者の関係性，適切性の規範を送信する情報の内容として定義し，それに基づくチェックリストを作成した．結果として，HIPAA[18] に基づくチェックリストを用いることで，LLM が規範を理解し，プライバシー侵害を正確に判断できる可能性が示された．しかし，この手法は HIPAA に限定して検証されており，より広範な応用に関する実験は行われていない．

Mireshghallah らの研究は，LLM が推論時にプライバシーをどの程度配慮できるかを評価するための指標「CONFAIDE」を提案している．CONFAIDE は 4 段階の評価基準で構成され，段階が進むにつれて文脈が複雑化し，より実践的な評価が可能となる．この研究では，ChatGPT などに対して評価を行い，各段階における人間の評価との相関関係を算出している．結果として，評価段階が進むにつれて人間との相関が低下し，プロンプティングや Chain-of-Thought を活用しても改善が限定的であることが示された．しかし，CONFAIDE は CI 理論を用いた評価指標だが，言語的な文脈のみを評価しており，時間および空間的コンテキストを評価していない．

Shao らの研究 [13] では，プライバシーリスクを評価する新たな手法である PrivacyLens を提案している．PrivacyLens では，パーソナライズされた AI エージェントに対して，情報を聞き出そうとする状況を想定し，情報漏洩を評価している．評価するコンテキストとして，質問の要旨，AI エージェントのオーナー，質問者，オーナーと質問者の関係，AI エージェントを組み込んだツールの 5 つを判断材料にしている．結果として，LLM がプライバシー規範を完全には認識していないことを指摘している．また，LLM のスケーリングやプロンプティングのみでは，この問題を十分に解決できないことも示唆している．

このように，CI 理論を LLM に導入する研究は進展しているが，コンテキストの設定範囲が狭い傾向にある．例えば，Li らの研究 [17] では，CI 理論の適用範囲を HIPAA という言語的な文脈に限定している．また，多くの研究が言語的な文脈としてコンテキストを捉えており，時間および空間的コンテキストを考慮した研究はほとんど見られない．

4.2 マルチモーダル LLM の既存研究

マルチモーダル LLM とは，自然言語を含む複数の種類のデータを入力および生成結果として処理できる LLM を指す．これを利用することで，画像やセンサデータを入力として与え，その入力に応じた適切な生成結果を得ることが可能となる．現在のマルチモーダル LLM は主に画像，ビデオ，オーディオ，3D データ，テキストに焦点を当てている [19, 20]．

しかし，実際にはさらに多くのモダリティが存在しており，周囲の状況を詳細に把握するには，これら多様なモダリティへの対応が重要である．各モダ

リティに対応するためには，モダリティごとにトークンをベクトルに変換する Embedding 層を学習する必要がある．このため，対応するモダリティが増えるほど，LLM の学習コストも増大するという課題がある．この課題に対処する手法として，Yoon らの研究 [21] が提案されている．Yoon らの手法では，Large Vision-Language Model（LVLM）のみを用いてセンサデータを生成結果に反映することを可能にした．この手法は，センサデータを画像に変換し，LLM に入力する視覚的プロンプトを提案することで，Embedding 層の学習を省略している．一方で，現行のセンサデータ変換手法ではいくつかの制約が存在する．具体的には，高密度なセンサデータを視覚的プロンプトに変換する際，変換後のプロンプトが LLM にとって十分な表現力を持たない場合がある．そのため，LLM が画像からセンサデータを読み取る際の精度が不安定になることが確認されている．

5　取り組むべき課題

本節では，マルチモーダル LLM を活用し，コンテキストを考慮したプライバシーリスクを評価したうえで適切な生成結果を得るために取り組むべき課題について整理する．

5.1　時間および空間的コンテキストを考慮した生成制御

マルチモーダル LLM が時間的および空間的コンテキストを考慮して生成結果を得るための学習手法や，これらのコンテキストを適切に反映した生成をする指示方法の研究が重要となる．特に，マルチモーダル LLM と In-Context Learning（ICL）の技術を組み合わせることで，これらの目標を効率的に達成できる可能性がある．

ICL は，LLM が学習後にパラメータを更新せず，少数の例示を与えることで新たなタスクを学習する手法を指す [22]．この技術により，LLM の更新を行わずに柔軟な生成制御が可能となる．マルチモーダル LLM においては，複数のモダリティ（例: テキスト，画像，音声）を統合的に扱う能力が求められるが，ICL の応用により，各モダリティにおける時間的および空間的な文脈情報を考慮したタスク設定と生成制御を効率的に実現できると考えられる．このため，ICL とマルチモーダル LLM は補完的な関係にあり，セットとして活用することでさらなる性能向上が期待される．関連研究として，ICL とプライバシーに関するサーベイ論文 [6] がある．この論文では，ICL やプロンプト利用中のプライバシー保護技術が体系的に整理されている．この論文で議論されているプライバシーリスクは主に情報漏洩に関するものである．ここで言う情報漏洩とは，プロンプト内に含まれる例示や質問データが第三者に露見するリスクを指し，生成結果のプライバシー保護研究の範疇ではない．

したがって，今後の研究課題として，ICL を活用し，時間的および空間的コンテキストを考慮した生成制御手法を検討することが求められる．

5.2　プライバシーリスクの評価指標

現状では LLM の生成結果が利用者に出力される状況に応じて生じるプライバシーリスクの評価指標が不足している．従来の LLM におけるプライバシーリスクの評価は，主に生成結果に個人情報や情報漏洩の量に対して焦点を当てている [4, 23]．しかし，今後は AI の発話までの経緯，発話時の周囲の状況を考慮した評価指標が必要となる．現在提案されているプライバシー評価指標の中で CI 理論を用いているものに，CONFAIDE[15] がある．しかし，CONFAIDE は CI 理論を用いているものの，言語的な文脈のみを評価しており，時間および空間的コンテキストを考慮していない．

つまり，時間および空間的コンテキストを考慮し，LLM の生成結果が利用者に出力される状況に応じたプライバシーリスクを評価する指標の開発が求められる．この評価指標を用いて，5.1 節で述べた制御手法を開発することで，高度なプライバシー配慮型の生成結果の得ることが可能になり，パーソナライズされた AI の安全性と信頼性を向上させることが期待される．

6　まとめ

本稿では，LLM を実装した AI が，パーソナライズされた AI エージェントとして人々の生活を支援する状況において，AI が出力する生成結果のプライバシーへの配慮の必要性を示した．また，このプライバシーへの配慮を実現するために必要な技術を調査し，それを基に取り組むべき課題について提案した．今後，本稿で提起した課題に関する研究がさらなる発展を遂げることを期待する．

参考文献

[1] OWASP. OWASP top 10 for large language model applications.

[2] 総務省, 経済産業省. AI 事業者ガイドライン第 1.01 版.

[3] 独立行政法人情報処理推進機構 産業サイバーセキュリティセンター中核人材育成プログラム 7 期生. テキスト生成 AI の 導入・運用ガイドライン.

[4] Haoran Li, Yulin Chen, Jinglong Luo, Jiecong Wang, Hao Peng, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, Zenglin Xu, Bryan Hooi, and Yangqiu Song. Privacy in large language models: Attacks, defenses and future directions. **arXiv preprint arXiv:2310.10383**, 2023.

[5] Hareem Kibriya, Wazir Zada Khan, Ayesha Siddiqa, and Muhammad Khurram Khan. Privacy issues in large language models: A survey. **Computers and Electrical Engineering**, Vol. 120, p. 109698, 2024.

[6] Kennedy Edemacu and Xintao Wu. Privacy preserving prompt engineering: A survey. **arXiv preprint arXiv:2404.06001**, 2024.

[7] Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. On protecting the data privacy of large language models (llms): A survey. **arXiv preprint arXiv:2403.05156**, 2024.

[8] Sergio Martínez, David Sánchez, Aida Valls, and Montserrat Batet. Privacy protection of textual attributes through a semantic-based masking method. **Information Fusion**, Vol. 13, No. 4, pp. 304–314, 2012. Information Fusion in the Context of Data Privacy.

[9] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. **arXiv preprint arXiv:2402.08787**, 2024.

[10] Stephen Meisenbacher and Florian Matthes. Thinking outside of the differential privacy box: A case study in text privatization with language model prompting. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 5656–5665, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[11] Jia-Ying Zheng, Hainan Zhang, Lingxiang Wang, Wangjie Qiu, Hong-Wei Zheng, and Zhi-Ming Zheng. Safely learning with private data: A federated learning framework for large language model. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 5293–5306, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[12] Helen Nissenbaum. Privacy as Contextual Integrity. Vol. 79, No. 1, p. 119.

[13] Xintao Wu Kennedy Edemacu. PrivacyLens: Evaluating privacy norm awareness of language models in action. **arXiv preprint arXiv:2409.00138**, 2024.

[14] Yan Shvartzshnaider, Vasisht Duddu, and John Lacalamita. Llm-ci: Assessing contextual integrity norms in language models. **arXiv preprint arXiv:2409.03735**, 2024.

[15] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory. In **The Twelfth International Conference on Learning Representations**, 2023.

[16] Wei Fan, Haoran Li, Zheye Deng, Weiqi Wang, and Yangqiu Song. GoldCoin: Grounding large language models in privacy laws via contextual integrity theory. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 3321–3343. Association for Computational Linguistics.

[17] Haoran Li, Wei Fan, Yulin Chen, Jiayang Cheng, Tianshu Chu, Xuebing Zhou, Peizhao Hu, and Yangqiu Song. Privacy checklist: Privacy violation detection grounding on contextual integrity theory. **arXiv preprint arXiv:2408.10053**, 2024.

[18] Accountability Act. Health insurance portability and accountability act of 1996 (HIPAA).

[19] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. MM-LLMs: Recent advances in MultiModal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 12401–12430, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[20] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 46, No. 8, pp. 5625–5644, 2024.

[21] Hyungjun Yoon, Biniyam Aschalew Tolera, Taesik Gong, Kimin Lee, and Sung-Ju Lee. By my eyes: Grounding multimodal large language models with sensor data via visual prompting. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 2219–2241. Association for Computational Linguistics.

[22] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 1107–1128. Association for Computational Linguistics.

[23] Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. SafetyBench: Evaluating the safety of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15537–15553, Bangkok, Thailand, August 2024. Association for Computational Linguistics.