（報道発表資料）

2025.7.7

NTT 株式会社

セキュリティに配慮した LLM の応答精度向上技術を確立

～定型的な自動応答において LLM の応答からの学習用データの漏洩リスクを抑えつつ応答精度向上に活用～

発表のポイント：

- ◆ 問い合わせ履歴などの定型的な構造を持つデータが学習に使われたかどうかが、新たな利用者によって推測され、その情報が漏洩するリスクを抑えながら、LLM の応答精度を改善できる新たな手法を確立しました。
- ◆ 漏洩リスクを抑えるためのノイズによって生じる応答精度低下の仕組みを理論的に明らかにし、理論に基づき、重要な単語に注目させ応答精度を向上する手法を提案しました。
- ◆ 医療・行政・金融など、利用者に関わるデータの扱いに慎重さが求められる分野において、将来的なリスクに備えた LLM の活用が期待されます。

NTT は、不特定多数の利用者から寄せられる問い合わせ応答を自動化する場面において、過去の利用者の入力と応答のペアを漏洩リスクを抑えながら活用し、新たな利用者への応答精度を高めることができる手法を確立しました。大規模言語モデル（Large Language Model; LLM）による自動応答では、情報漏洩リスクを抑えるために、プライバシー保護の強度を定量化する指標である差分プライバシー(*1)に基づいてノイズを加える(*2)方法が注目されていますが、その影響で応答精度が下がる課題がありました。ノイズが応答精度に与える影響を世界で初めて理論的に分析しました。この知見に基づき、本研究では、差分プライバシーを維持しつつ応答傾向の推定精度を向上させる新たな入力と応答のペアの生成手法として、Plausible Token Amplification（PTA） を提案し、重要な単語に注目させることで精度と安全性の両立を可能にする仕組みを実現しています。この成果は、医療・行政・金融など、利用者に関わるデータの扱いに安全性と実用性の両立が求められる分野において、将来的なリスクに備えた LLM 活用が期待されます。

なお、本成果は 2025 年 7 月 13 日から 19 日まで、カナダで開催される機械学習分野における難関国際会議 International Conference on Machine Learning（ICML） 2025（*3）において発表されます。

1. 背景

文脈内学習（In-Context Learning; ICL）は、あらかじめ与えた例文に沿って LLM の応答を誘導する手法で、定型化された構造をもつ問い合わせ対応の自動化などに活用が期待されています。たとえば、多くの利用者が同一の LLM ベースのチャットボットを通じてサポートを受ける環境では、「イヤホンが届いていません → 配送」といった過去の利用者の入力と応答のペア（例題）を文脈としてあ

らかじめ与えることで、新しい問い合わせも過去の応答傾向に基づいて分類され、それに応じた定型的な応答を自動で提示できます。しかしこうした仕組みでは、過去の問い合わせ内容が別の利用者への応答に反映されるため、新たな利用者が似た問い合わせを意図的に繰り返すことで、「ある問い合わせがあったか」といった情報が、統計的に第三者に漏洩するリスクがあります（図 1a）。このようなリスクは、一見個人を直接含まない情報であっても、繰り返しのやりとりの中で漏洩が発生しうる点に特徴があり、今後の LLM 活用において将来的に顕在化が懸念されます。

近年では入力と応答が一対一に対応する例題のような構造化データに対する統計的な漏洩リスクを低減する手法として、例題に単語レベルでノイズを加えることで安全性を保つ方法が使われ、差分プライベートな ICL（DP-ICL）(*4)と呼ばれています（図 1b）。しかし、DP-ICL ではノイズの影響により例題の内容が曖昧になり、LLM が過去の利用者の例題に共通する正しい応答傾向を捉えにくくなることで、応答の精度が大きく低下するという課題があります。これに対し、従来は無関係な単語をあらかじめ生成候補から除外するといった経験的な工夫（*5）によって精度改善が試みられてきましたが、そうした手法がなぜ有効なのかについての理論的な説明は不十分でした。

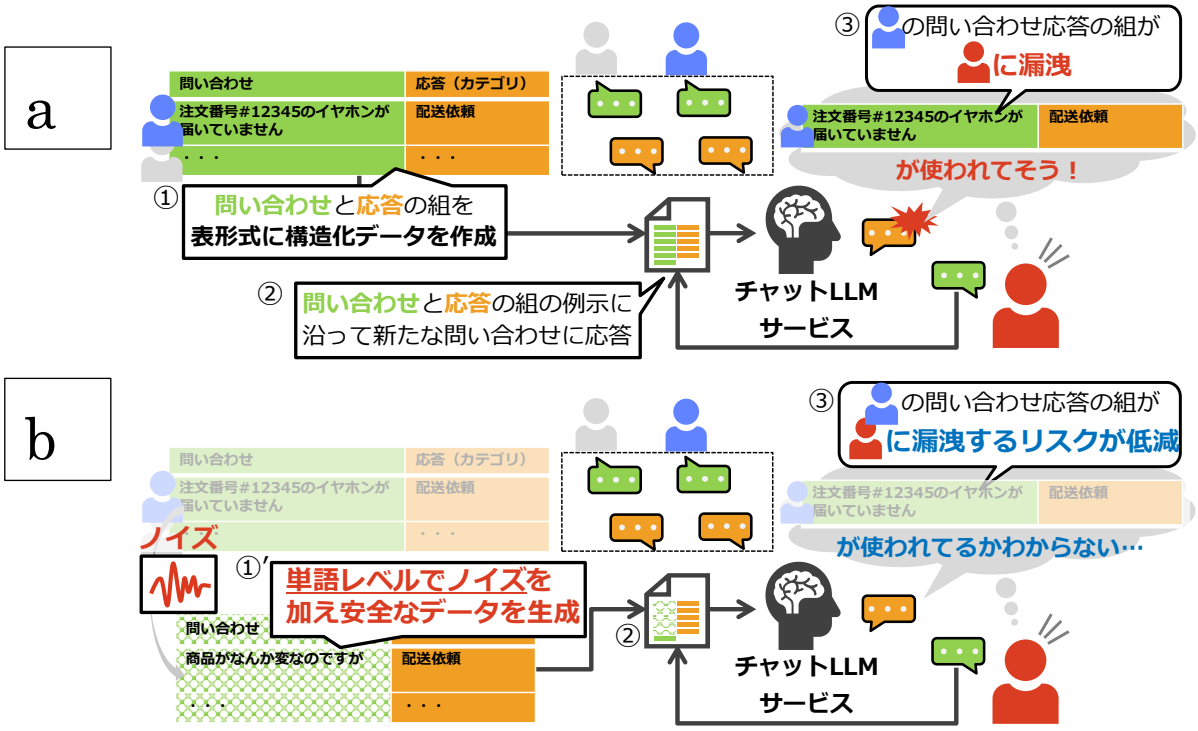

図１文脈内学習を活用する上での情報漏洩リスクとその対処アプローチ

a：文脈内学習を用いたサービスイメージと機微情報の漏洩リスクの例

b：ノイズにより漏洩リスクを低減するイメージ例

2. 技術のポイント

本研究では、DP-ICL における応答精度の低下要因を理論的に分析し、安全性と応答精度を両立する新たな安全な例題生成手法を提案しました。ICL は、例題に基づいて LLM が応答傾向（ルール）を推定する仕組みと捉えられます(*6)。本研究では、ノイズがこのルール推定に与える影響をベイズ推論の枠組みで理論的に解析し、以下 2 つの新たな知見を明らかにしました。

・知見１：無関係な単語を生成候補からあらかじめ除外することで、ノイズによるルール推定への悪影響を緩和できることを理論的に示しました。これは、従来経験的に行われてきた単語の生成候補の削減による性能改善に対する理論的な裏付けとなります。

・知見２：ルールを特徴づける単語の生成確率を意図的に高めることで、ノイズが加えられた例題からでも LLM が正しいルールをより高精度に推定できることを明らかにしました（図 2）。これは既存の研究で見落とされていた新たな DP-ICL の応答精度の改善の方向性を示すものです。

たとえば、「注文番号#12345 のイヤホンが届いていません」という入力に対して「配送」という応答と分類されるのが正しい対応です。しかし、例題にノイズを加えると、「なんか商品が変なのですが → 配送」のような曖昧な例題が生成されることがあります。この文は一見自然に読めますが、「とりあえず」や「商品」といった語は分類の手がかりとしては乏しく、本来の重要語である「イヤホン」や「届かない」といった情報が埋もれてしまいます。すると、新たな入力「商品が破損して届いた」に対しても、モデルが「配送」と誤って分類するなど、ルールの誤認による分類ミスが生じるリスクがあります。誤りに対して、本研究の[知見 1]では、「なんか」など分類に寄与しない語をあらかじめ生成候補から除外することで、ルールの安定した推定が可能になることを明らかにしました。さらに[知見 2]では、「届かない」や「壊れた」など、ルールを特徴づける単語の生成確率を適切に高めることで、ノイズが加えられても、LLM が「配送」や「返品」といったルールをより正確に推定できるようになることを理論的に示しました。

これらの理論的知見に基づき、本研究では、差分プライバシーを維持しつつルールの推定精度を向上させる新たな例題生成手法として、Plausible Token Amplification（PTA） を提案しました。PTA は、無関係な語の生成を抑えながら、ルールを特徴づける単語の生成確率を高めた上で、ノイズを加えて安全な例題を生成します。PTA により、ノイズが加えて生成された安全な例題からでも LLM は正しいルールを高精度に推定することができ、応答精度と安全性の両立が実現されます。

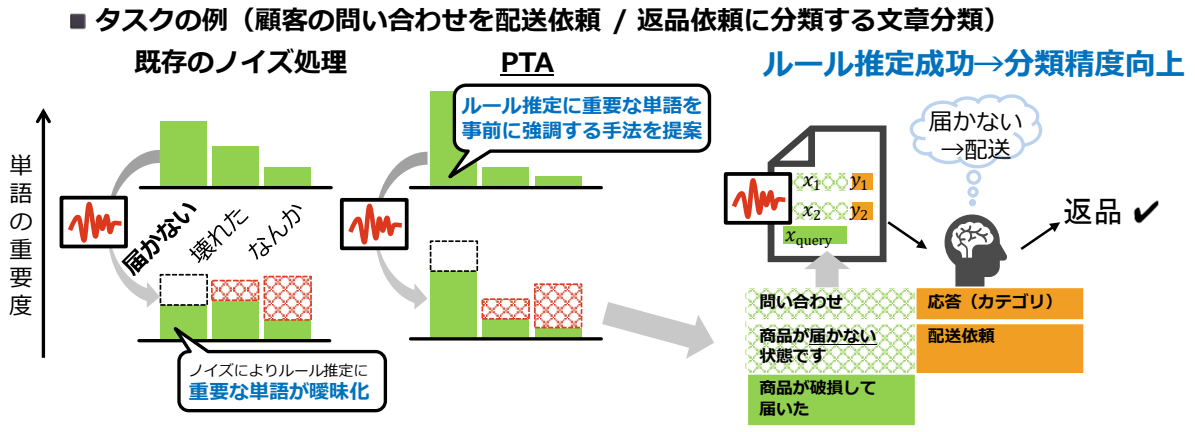

図 2 提案法 PTA の概略図

左）重要な単語を強調　右）ルール推定が成功し応答精度が向上

PTA の有効性を確認するために、ニュース記事をスポーツや世界情勢といったトピックのカテゴリに分類するベンチマークタスクにおいて既存の DP-ICL 手法と比較し精度向上を確認するとともに、ノイズを加えない ICL とも同等の精度を実現できることを確認いたしました。(図 3)

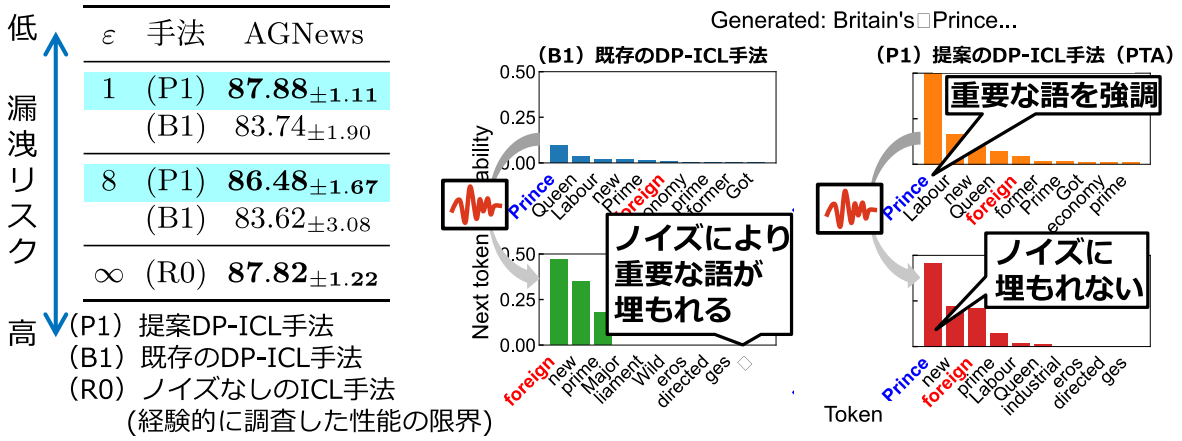

図 3 文章を分類するタスクにおける評価結果。左)差分プライバシーにより定量化される安全性の強度ε(*7)(小さいほど漏洩リスクが低減)を変化させた場合のベンチマーク精度の手法間比較。右)トピックのカテゴリが世界情勢の場合に、PTA で生成されたニュース記事の冒頭の例。

今回提案した PTA は、問い合わせ履歴などの例題が LLM の入力に使われていたかどうかが、新たな利用者から推測されにくくする手法です。たとえば、チャットボットによる自動応答サービスでは、過去の問い合わせが使われていると推測されること自体が機密情報の漏洩につながるおそれがあり、PTA はその漏洩リスクを低減することで、安全なデータ活用を支援します。なお、出力される応答に機密情報を含まないことを決定論的に保証するものではなく、あくまで「その問い合わせが使われたかどうか」の推測を統計的に困難にすることに焦点を当てています。

3. 今後の展開

今後は、安全な例題の生成時の単語の強調処理を高度化することで、定型化されたタスクにおける将来的に懸念される統計的な漏洩リスクを抑えながらも、高精度な応答を維持できる手法の確立をめざします。これにより、医療・金融・行政などの利用者に関わるデータの扱いに慎重さが求められる分野において、将来的なリスクに備えた LLM の活用が期待されます。

また、現在の PTA は入力と応答があらかじめ定められた形式(例:問い合わせとカテゴリ分類のペア)を前提としていますが、今後はより柔軟な構造の入力を扱うタスクへの応用も視野に入れています。たとえば、自由記述形式の問い合わせや複数分類の併用といった実運用で求められるユースケースにも対応可能とすることで、将来のデータ漏洩リスクを見据えた、より幅広い分野におけるデータのセキュリティに配慮した LLM の活用環境の実現をめざします。

4. 発表について

本成果は、2025 年 7 月 13～19 日に開催される機械学習分野における最難関国際会議 ICML 2025（The Forty-Second International Conference on Machine Learning）にて、下記のタイトル及び著者で発表されます。

タイトル： Plausible Token Amplification for Improving Accuracy Differentially Private In-Context Learning Based on Implicit Bayesian Inference

著者： 山﨑 雄輔（社会情報研究所）、丹羽 健太（コンピュータ科学基礎研究所）、千々和 大輝（コンピュータ&データサイエンス研究所）、深見 匠、三浦 尭之（社会情報研究所）

URL： https://openreview.net/forum?id=skAjaAEuA2

【用語解説】

*1 差分プライバシー（Differential Privacy）

構造化されたデータベースに対する統計的処理の出力が特定のレコードの有無に関わらず統計的に区別できないのであれば安全であるという識別困難性に基づくプライバシー保護の強度を定量化する指標です。識別困難性は(ε, δ)というパラメタで定量化され、これらの値が小さいほど、レコードの有無が識別されにくいこと(漏洩するリスクが小さいこと)を意味する。したがって、統計的処理の出力が(ε, δ)-DP を満たす処理は、統計的な漏洩リスクを抑える手法です。

*2 ノイズを加える

データベースに対する統計的処理の出力にノイズを加えることで、個別のレコードの出力への影響を制限し(ε, δ)-DP を満たすようにする一般的な手法です。

*3 ICML 2025

機械学習に関するトップレベルの国際会議。

URL: https://icml.cc/Conferences/2025

*4 DP-ICL

DP-ICL は、入力と応答のペアからなる例題（問い合わせと分類カテゴリなど）を各行に持つ構造化されたデータベースを対象に、その例題を文脈として用いる ICL の出力が(ε, δ)-DP を満たすように設計する手法です。具体的には、このデータベースからサンプリングした例題を用いて LLM に入力する際、出力分布にノイズを加えて安全な例題を生成することで、特定の問い合わせと分類カテゴリの組みが文脈に含まれていたかどうかを出力から第三者が識別することを困難にします。

*5 DP-ICL の経験的な精度改善を提案した既存研究

文献情報：Tang et al., “Privacy-preserving In-Context Learning with Differentially Private Few-shot Generation,” ICLR, 2024.

URL: https://openreview.net/forum?id=oZtt0pRnOl

*6 ICL のベイズ推論的な理論解析を行った既存研究

文献情報：Xie et al., “An Explanation of In-Context Learning as Implicit Bayesian Inference,” ICLR, 2022.

*7 差分プライバシーにより定量化される安全性の強度ε

Apple iOS（予測変換・辞書学習）では $\varepsilon = 1\sim8$ を保証している。

URL： https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

■本件に関する報道機関からのお問い合わせ先

NTT 株式会社

サービスイノベーション総合研究所

広報担当

問い合わせフォームへ