УДК 004.89

МОДЕЛИРОВАНИЕ ВЕРОЯТНОСТИ УТЕЧКИ ДАННЫХ ПРИ ПОСЛЕДОВАТЕЛЬНЫХ ЗАПРОСАХ К ЯЗЫКОВОЙ МОДЕЛИ

Pao H., студент направления Applied Mathematical Analysis MSc programme, II курс
Технический университет Дании, г. Копенгаген, Дания

Языковые модели, особенно крупные трансформеры, такие как GPT, BERT или LLaMA, всё чаще используются в продуктах и сервисах, которые предполагают интерактивное взаимодействие с пользователем. Эти модели задействуются в виртуальных помощниках, интеллектуальных поисковиках, инструментах поддержки клиентов, аналитических системах и многих других приложениях, где пользовательские запросы передаются последовательно, часто в форме диалога. В таких условиях появляется специфический вектор угроз – возможность накопления и утечки конфиденциальной информации через анализ последовательности запросов. Моделирование вероятности утечки данных в этих условиях становится важной задачей как для исследовательского сообщества, так и для разработчиков доверенных систем искусственного интеллекта [1].

В отличие от однократного взаимодействия, при котором анализ возможной утечки может быть ограничен одной трансакцией, при последовательных запросах пользователи зачастую непреднамеренно раскрывают чувствительные аспекты своей личности, контекста задачи или данных. Языковая модель при этом может не только сохранять локальный контекст в пределах одной сессии (например, в окне внимания), но и потенциально формировать устойчивые паттерны, которые могут быть извлечены при целенаправленной атаке. Особенно высокий риск возникает в случаях, когда одно и то же лицо взаимодействует с моделью многократно, постепенно раскрывая связанные фрагменты информации, которые в совокупности могут привести к полной реконструкции приватного содержания – например, номера счёта, диагноза, имени или даже пароля.

Моделирование вероятности такой утечки требует комплексного подхода. Одной из отправных точек является построение формального представления диалога или сессии как цепи случайных величин, где каждый следующий запрос зависит от предыдущих, а ответы модели зависят не только от текущего входа, но и от внутреннего состояния модели. В этом случае оценка вероятности утечки сводится к изучению условной взаимной информации между

приватными данными и выходом модели, учитывая весь контекст предыдущих запросов. Такой подход позволяет формализовать риск накопления приватной информации и выявить ключевые параметры, влияющие на его рост, такие как длина сессии, объем чувствительных данных, структура запросов и агрессивность атакующего алгоритма [2].

Для практического анализа используются методы симуляции атак, таких как membership inference, model inversion, training data extraction и контекстное зондирование. При этом моделируются различные сценарии: от пассивного прослушивания до активного взаимодействия атакующего с моделью с целью выявления чувствительной информации. Особый интерес представляют «грэйбокс»-сценарии, когда атакующий имеет частичный доступ к архитектуре или логике модели, например, в виде API-интерфейса. Эксперименты показывают, что при достаточной длине сессии, повторяющихся структурах запросов и отсутствии механизмов ограничения памяти модели, вероятность извлечения скрытых данных резко возрастает, особенно если в ход идут комбинированные техники анализа вероятностных паттернов и семантического сходства между предсказаниями [3].

Ключевым элементом моделирования становится также оценка устойчивости модели к накоплению контекста. Например, в случае архитектур с длинным контекстным окном возникает необходимость изучения того, как долго сохраняется след от чувствительной информации в латентных представлениях модели. Методы анализа скрытых слоёв, векторных расстояний, внимания и влияния токенов позволяют оценить, насколько «прочно» закрепляется приватный контекст внутри модели. Такие методы дают возможность не только оценить вероятность утечки, но и предсказать, какие именно данные наиболее уязвимы в условиях длительного взаимодействия.

Для повышения доверия и безопасности при использовании языковых моделей в условиях последовательных запросов необходимы также защитные меры, направленные на снижение вероятности утечки. Это могут быть механизмы обрезки контекста, регулярное «забывание» или сброс сессии, внедрение слоёв фильтрации чувствительных фраз, дифференциально-приватное дообучение и агрегация ответов модели. Кроме того, разрабатываются формальные критерии для мониторинга риска утечки в реальном времени, что позволяет адаптивно регулировать поведение модели при подозрительных запросах [4].

Таким образом, моделирование вероятности утечки данных при последовательных запросах к языковой модели представляет собой важное и быстро развивающееся направление исследований в области доверенного ИИ [5].

Это направление требует как математической строгости, так и практической инженерной реализаций – от построения формальных моделей угроз до внедрения устойчивых архитектур и защитных механизмов.

Только через понимание динамики информационных потоков в интерактивных сессиях можно обеспечить реальную безопасность и конфиденциальность в приложениях, использующих языковые модели как активный интерфейс взаимодействия с данными и пользователями [6].

Список литературы:

1. Xu, R. A Survey of Privacy Attacks in Machine Learning / R. Xu, N. Baracaldo, J. Joshi // ACM Computing Surveys. – 2020. – Vol. 53, № 4. – Article 88. – DOI 10.1145/3375823.

2. Yang, Q. Federated Machine Learning: Concept and Applications / Q. Yang, Y. Liu, T. Chen, Y. Tong // ACM Transactions on Intelligent Systems and Technology. – 2019. – Vol. 10, № 2. – Article 12. – DOI 10.1145/3298981.

3. Li, T. Federated Learning: Challenges, Methods, and Future Directions / T. Li, A. K. Sahu, A. Talwalkar, V. Smith // IEEE Signal Processing Magazine. – 2020. – Vol. 37, № 3. – P. 50–60. – DOI 10.1109/MSP.2020.2975749.

4. Bagdasaryan, E. How to Backdoor Federated Learning / E. Bagdasaryan, A. Veit, Y. Hua [et al.] // Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. – 2020. – Vol. 108. – P. 2938–2948.

5. Zhao, Y. Analyzing Federated Learning through an Adversarial Lens / Y. Zhao, J. Zhao, M. Yang [et al.] // Proceedings of the 36th International Conference on Machine Learning. – 2019. – Vol. 97. – P. 7510–7520.

6. Humbert, M. Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy / M. Humbert, E. Ayday, J.-P. Hubaux, A. Telenti // Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2013. – P. 1141–1152. – DOI 10.1145/2508859.2516707.