УДК 004.89

РИСКИ ДЕАНОНИМИЗАЦИИ ПОЛЬЗОВАТЕЛЕЙ ПРИ ГЕНЕРАЦИИ ТЕКСТОВ НА ОСНОВЕ ПОВЕДЕНЧЕСКИХ ШАБЛОНОВ

Вэй Ч., студент направления Digital Systems MSc programme, II курс
Технический университет Дании, г. Копенгаген, Дания

Генерация текстов с использованием моделей искусственного интеллекта, обученных на поведенческих данных пользователей, открывает широкие перспективы для персонализации цифровых сервисов, интеллектуальных помощников, рекомендательных систем и интерфейсов взаимодействия человек–машина [1].

Однако этот же процесс несёт в себе и критически важные риски, связанные с возможностью деанонимизации пользователей – то есть восстановления их личности или чувствительных характеристик на основании выдаваемых системой текстов.

Особенно это актуально в случаях, когда генерация осуществляется не на общих данных, а на поведенческих шаблонах, таких как стиль письма, временные паттерны, типичные лексемы, частота упоминания определённых тем и другие индивидуальные маркеры [2].

Риски деанонимизации при генерации основаны на том, что даже в отсутствие прямых идентификаторов – таких как имя, адрес электронной почты или номер телефона – тексты, сгенерированные моделью, могут содержать латентные признаки, уникальные для конкретного пользователя или узкой группы лиц. Поведенческие шаблоны, извлечённые из истории запросов, писем, предпочтений или даже метаданных (например, времени активности), становятся своеобразным «цифровым отпечатком», который может быть легко сопоставим с внешними источниками. При наличии доступа к сторонним базам данных (например, взломанным утечкам, публичным форумам или социальным сетям), такие шаблоны могут быть использованы для корреляционной атаки, приводящей к восстановлению реальной личности.

Механизмы деанонимизации в генеративных моделях, таких как большие языковые модели (LLM), могут быть как прямыми, так и опосредованными. В первом случае модель случайно или систематически генерирует фразы, выражения или последовательности, встречающиеся только у одного пользователя. Это может быть уникальная комбинация лексики, стиль употребления терминов, характерная пунктуация, даже определённые синтаксические конструкции. Во втором случае – при опосредованной деанонимизации – сама структура

текста, включая предпочтительный порядок мысли, длину предложений, эмфатические конструкции и переходы, может быть использована для построения профиля личности и сопоставления его с другими данными.

Особенно уязвимыми к деанонимизации оказываются системы, использующие онлайн-обучение или адаптацию под пользователя, где модель накапливает и уточняет представление о поведенческих паттернах в процессе постоянного взаимодействия. При отсутствии строгих механизмов приватности (например, дифференциальной приватности или механизма ограниченного забывания) такие модели могут быть склонны к меморизации – сохранению в своей параметрической структуре особенностей взаимодействий, пригодных для последующего извлечения. Это делает возможным не только случайное включение приватной информации в сгенерированный текст, но и целенаправленные атаки на восстановление – когда внешняя система, взаимодействуя с моделью, может пошагово «выуживать» информацию о поведении пользователя, не имея к нему прямого доступа [3].

Одним из ярких примеров риска деанонимизации являются чат-боты и персонализированные ассистенты, которые могут использовать знания о прошлых запросах, чтобы уточнять будущие ответы. Если один и тот же бот обслуживает множество пользователей, но в системе плохо реализована изоляция пользовательских контекстов, появляется риск контекстной утечки – когда элементы поведения одного пользователя проявляются в ответах, предназначенных другому.

Даже незначительные фрагменты – например, характерные словосочетания или темы – могут быть достаточными для стороннего аналитика, чтобы сопоставить их с уже известным профилем личности [4].

Методы защиты от деанонимизации включают как архитектурные, так и формальные подходы. К архитектурным относятся изоляция пользовательских сессий, ограничение объёма хранимых пользовательских признаков, внедрение механизмов случайного шума в процессе генерации.

Среди формальных – применение дифференциальной приватности в процессе обучения, обрезание градиентов, ограничения на количество итераций персонализации, использование приватного тонкого тюнинга с регулярной обнулением весов, хранящих персональные зависимости. Также возможно применение анализа поведенческих расстояний – метрики, которые фиксируют уровень отличия между выходами модели при незначительном изменении входов, чтобы своевременно выявить риск утечки индивидуальных паттернов.

Однако даже при наличии таких мер необходима периодическая аудиторская проверка, направленная на выявление остаточной зависимости модели от индивидуальных характеристик пользователей. Это включает стресс-тесты, генерацию выходов с шумовыми входами, атаки восстановления шаблонов и

применение специальных тестовых наборов, построенных на редких и уникальных поведенческих особенностях. Более того, важно сопровождать все этапы разработки и внедрения модели этической экспертизой, особенно в случаях, когда система работает с уязвимыми группами (например, детьми, пациентами, представителями меньшинств) [5].

В заключение, можно утверждать, что риски деанонимизации пользователей при генерации текстов на основе поведенческих шаблонов представляют собой не гипотетическую, а вполне реальную угрозу для приватности.

Они особенно опасны своей незаметностью: пользователь может не подозревать, что его поведенческие особенности запечатлеваются в модели, и ещё менее – что они воспроизводятся в чужом контексте. Это требует от разработчиков, регуляторов и научного сообщества комплексного подхода, включающего технические, юридические и этические меры, направленные на обеспечение подлинной анонимности и прозрачности использования ИИ в персонализированных системах [6].

Список литературы:

1. Wang, S. Locally Differentially Private Protocols for Frequency Estimation / S. Wang, L. Huang, P. Wang [et al.] // USENIX Security Symposium. – Berkeley : USENIX Association, 2017. – P. 729–745.

2. Bassily, R. Local, Private, Efficient Protocols for Succinct Histograms / R. Bassily, A. Smith // ACM Symposium on Theory of Computing. – New York : ACM, 2015. – P. 127–135. – DOI 10.1145/2746539.2746632.

3. Bassily, R. Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds / R. Bassily, A. Smith, A. Thakurta // IEEE Symposium on Foundations of Computer Science. – Piscataway : IEEE, 2014. – P. 464–473. – DOI 10.1109/FOCS.2014.56.

4. Chaudhuri, K. Differentially Private Empirical Risk Minimization / K. Chaudhuri, C. Monteleoni, A. D. Sarwate // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 1069–1109.

5. Chaudhuri, K. Near-Optimal Differentially Private Principal Components / K. Chaudhuri, A. D. Sarwate, K. Sinha // Advances in Neural Information Processing Systems. – 2012. – Vol. 25. – P. 998–1006.

6. Hardt, M. Train Faster, Generalize Better: Stability of Stochastic Gradient Descent / M. Hardt, B. Recht, Y. Singer // Proceedings of the 33rd International Conference on Machine Learning. – 2016. – Vol. 48. – P. 1225–1234.