УДК 004.89

ВЫЯВЛЕНИЕ РИСКОВ РАСКРЫТИЯ ЛИЧНОЙ ИНФОРМАЦИИ ПРИ ИСПОЛЬЗОВАНИИ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ

Романо Д., студент направления Industrial Production Engineering Bachelor's degree programComputer Visions Master's Degree Programm, II курс
Миланский технический университет, г. Милан, Италия

Выявление рисков раскрытия личной информации при использовании генеративных моделей представляет собой критически важную задачу в современном ландшафте развития искусственного интеллекта, особенно в условиях стремительного роста масштабов и возможностей таких моделей, как генеративные состязательные сети (GAN), вариационные автокодировщики (VAE), трансформерные архитектуры и, в особенности, большие языковые модели (LLM). Генеративные модели стали основой многих продуктов и сервисов – от голосовых помощников и чат-ботов до систем синтеза изображений и автоматического программирования. При этом, несмотря на их высокую производительность, такие модели могут невольно «вытекать» – т.е. раскрывать – данные, на которых они были обучены, включая потенциально идентифицирующую личную информацию [1].

Проблема заключается в том, что генеративные модели, особенно те, которые были обучены на больших, слабо контролируемых или неконтролируемых корпусах, таких как открытые интернет-архивы, способны не просто обобщать наблюдаемые паттерны, но и запоминать конкретные фрагменты обучающей выборки. Эти фрагменты – от номеров телефонов и адресов электронной почты до медицинских или юридических записей – могут быть воспроизведены моделью в ответ на определённые запросы, особенно если они были высокочастотными, статистически выделяющимися или повторяющимися в данных. Таким образом, ключевая задача исследователей и разработчиков – выявить, при каких условиях и с какой вероятностью генеративная модель может раскрыть личную информацию, и как это можно диагностировать и предотвращать.

Одним из основных способов выявления рисков является проведение атак с извлечением памяти (memorization extraction attacks), которые направлены на воспроизведение обучающих данных через прямое взаимодействие с моделью. В случае языковых моделей это могут быть запросы, провоцирующие воспоминание определённого документа, номера или имени [2].

Например, запрос «Повтори контактную информацию доктора X из Нью-Йорка» может – при некорректно обученной или незащищённой модели – привести к генерации персональных данных, если такие данные были в обучающем корпусе. Визуальные генеративные модели могут воссоздавать лица, отпечатки документов или медицинские изображения с высоким сходством к оригиналам. Такие случаи уже фиксировались в ряде научных работ и демонстрируют, что даже без явного намерения нарушить приватность, модель может стать невольным источником утечки.

Для оценки степени запоминания и склонности к утечкам используются специальные тесты, в частности test set exposure и canary insertion. Методика canary insertion предполагает преднамеренное добавление уникальных, искусственно созданных строк в обучающую выборку (например, фиктивного номера паспорта), а затем попытку извлечь их из модели после завершения обучения. Если такие «канарейки» воспроизводятся моделью с высокой точностью при соответствующем запросе, это указывает на наличие прямого запоминания и потенциал раскрытия аналогичных реальных данных.

Также важным методом выявления рисков является контрфактическое тестирование (counterfactual probing), при котором создаются модифицированные входы, близкие к известным примерам из обучающего множества. Сравнение поведения модели на оригинальных и изменённых входах позволяет оценить чувствительность к конкретным элементам обучающего корпуса и вероятность утечек. Наряду с этим применяются методы инверсии входов и выходов, где исследователь или атакующий пытается восстановить исходные данные по градиентам или ответам модели – особенно это опасно в случае, когда пользователи делят модели в рамках federated learning или других форм совместного обучения [3].

Отдельное внимание уделяется визуальной генерации – GAN и diffusion-моделям, обученным на лицевых базах или медицинских изображениях. Здесь утечки могут проявляться не через текст, а через частичную или полную реконструкцию реального изображения, особенно если оно встречалось многократно в обучающей выборке. Диагностика таких случаев требует применения метрик сходства (например, SSIM, LPIPS, FID) между сгенерированными и оригинальными изображениями, а также визуального анализа экспертов.

Противодействие этим рискам требует внедрения механизмов дифференциальной приватности, а также стратегий ограничения параметров модели и регуляризации. В частности, при обучении языковых моделей возможно внедрение контролируемого добавления шума в градиенты, что снижает вероятность запоминания конкретных строк. Также активно разрабатываются методы отказа от генерации по чувствительным шаблонам – например, система может быть обучена не отвечать на запросы, содержащие определённые триггерные

конструкции, связанные с возможной утечкой данных. Однако в отсутствие чёткого понимания всего содержимого обучающего корпуса подобные методы остаются лишь частичной мерой.

Кроме технических решений, важным элементом управления рисками является юридико-этический аудитисточников данных и обучение моделей в соответствии с принципами «privacy by design». Это означает, что уже на этапе подготовки датасета необходимо проводить фильтрацию, анонимизацию и исключение явно идентифицируемых записей. Использование лицензионно чистых, структурированных и проверенных наборов данных является важнейшей практикой, особенно в корпоративных и регулируемых контекстах (например, GDPR, HIPAA, FERPA и пр.) [4].

Таким образом, выявление рисков раскрытия личной информации при использовании генеративных моделей требует междисциплинарного подхода, сочетающего методы машинного обучения, безопасности, правовой экспертизы и этики. Тестирование моделей на запоминание, мониторинг их поведения в эксплуатации, внедрение приватностно-ориентированных архитектур и прозрачность в отношении источников данных – всё это необходимые меры для создания действительно доверенных ИИ-систем.

Только таким образом можно гарантировать, что генеративные модели будут служить интересам пользователей, не став источником нарушений конфиденциальности и личной безопасности [5].

Список литературы:

1. Juuti, M. PRADA: Protecting Against DNN Model Stealing Attacks / M. Juuti, S. Szyller, S. Marchal, N. Asokan // IEEE European Symposium on Security and Privacy. – Piscataway : IEEE, 2019. – P. 512–527. – DOI 10.1109/EuroSP.2019.00044.

2. Jagielski, M. High Accuracy and High Fidelity Extraction of Neural Networks / M. Jagielski, N. Carlini, D. Berthelot [et al.] // Proceedings of the 29th USENIX Security Symposium. – Berkeley : USENIX Association, 2020. – P. 1345–1362.

3. Krishna, K. Thieves on Sesame Street! Model Extraction of BERT-Based APIs / K. Krishna, G. Krishna, A. A. Bigham, Z. C. Lipton // International Conference on Learning Representations. – 2020. – P. 1–23.

4. Kariyappa, S. Maze: Data-Free Model Stealing Attack Using Zeroth-Order Gradient Estimation / S. Kariyappa, A. Prakash, M. K. Qureshi // IEEE/CVF Conference on Computer Vision and Pattern Recognition. – Piscataway : IEEE, 2021. – P. 13814–13823.

5. Correia-Silva, J. Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data / J. Correia-Silva, R. F. Berriel, C. Badue

[et al.] // International Joint Conference on Neural Networks. – Piscataway : IEEE, 2018. – P. 1–8. – DOI 10.1109/IJCNN.2018.8489592.