УДК 004.89

ИНСТРУМЕНТЫ ТЕСТИРОВАНИЯ МОДЕЛЕЙ ИИ НА ПРЕДМЕТ НАРУШЕНИЙ КОНФИДЕНЦИАЛЬНОСТИ

Николаева П.А., младший научный сотрудник
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

Тестирование моделей искусственного интеллекта на предмет нарушений конфиденциальности – это ключевая составляющая современного подхода к разработке доверенных ИИ-систем, особенно в условиях массового внедрения предобученных моделей в чувствительные области: здравоохранение, финансы, юриспруденция, системы рекомендаций и автоматического принятия решений. Инструменты для такого тестирования служат основой как для профилактики потенциальных утечек, так и для формализации аудита моделей в соответствии с правовыми и этическими нормами, включая требования GDPR, HIPAA, ISO/IEC 27001 и аналогичных стандартов [1].

Модель ИИ, будучи статистической конструкцией, обученной на конечном наборе данных, потенциально может сохранять и невольно воспроизводить фрагменты обучающей выборки – особенно если она не была должным образом регуляризована, либо если в обучающем наборе содержались повторяющиеся, аномальные или легко запоминающиеся наблюдения.

Нарушения конфиденциальности могут проявляться как в виде прямых утечек (воспроизведение персональных данных при генерации), так и в форме косвенного раскрытия, когда чувствительная информация может быть выведена из выходов модели или промежуточных представлений с помощью вспомогательного анализа. Инструменты тестирования, направленные на выявление подобных рисков, становятся необходимыми средствами обеспечения безопасности и устойчивости ИИ в открытой и корпоративной среде [2].

Одним из наиболее известных классов тестов являются атаки membership inference (MIA), целью которых является определение, был ли конкретный пример частью обучающего множества модели. В рамках инструментария тестирования они реализуются через программные библиотеки, способные формировать атакующие модели, обучаемые на выходах основной модели, и затем оценивать точность восстановления принадлежности. Инструменты, такие как Privacy Meter, ML Privacy Meter, TensorFlow Privacy, предоставляют API для проведения таких атак в разных конфигурациях – от black-box до white-box-доступа [3].

Другой важный класс инструментов ориентирован на извлечение или инверсию данных. Сюда входят методы, моделирующие атаки на восстановление входов по выходам модели (model inversion) или генерацию данных из градиентов в federated learning-сценариях (gradient leakage). Такие библиотеки, как Deep Leakage from Gradients (DLG) и Opacus, позволяют проверить, насколько легко можно восстановить изображения, тексты или числовые записи, если получить доступ к параметрам модели или промежуточным градиентам. Эти методы важны не только как демонстрации уязвимостей, но и как средства оценки пределов приватности в распределённом обучении.

Существуют и специализированные инструменты для тестирования языковых и генеративных моделей. Например, подход canary insertion (вставка «канареек» – искусственных строк в обучающий датасет) используется в тестах на утечку информации из LLM: если вставленный фрагмент может быть воспроизведён моделью при соответствующем запросе, это говорит о склонности к прямому запоминанию. Применяются также методы автоматической генерации prompt'ов (например, через Prompt Leakage Detection Toolkit) для выявления неконтролируемого запоминания контекста в диалоговых системах [4].

Визуальные модели тестируются на утечку с помощью метрик сходства изображений – SSIM, FID, LPIPS и др., которые позволяют сравнивать сгенерированные или восстановленные изображения с оригиналами из обучающей выборки. Особенно важны такие тесты для GAN и diffusion-моделей, используемых в медицинской визуализации, системах видеонаблюдения, генерации лиц и т.п., где воспроизведение реальных объектов может нарушать конфиденциальность пациентов или пользователей [5].

Помимо атакующих тестов, в современный инструментарий включаются метрики конфиденциальности, позволяющие без прямой симуляции атаки оценить риск утечки. К ним относятся показатели экспозиции (exposure scores), приватные вариации влияния (influence functions), дифференциально-приватные оценки (ε, δ-границы), а также показатели чувствительности модели к локальным изменениям входов. Эти метрики всё чаще интегрируются в ML-фреймворки, позволяя разработчикам отслеживать показатели приватности в ходе обучения и эксплуатации.

Наряду с анализом поведения модели, инструменты тестирования всё чаще включают возможности статического аудита модели и обучающего корпуса. Например, библиотеки могут сканировать датасеты на наличие потенциально чувствительной информации (PII scanning), анализировать частотность повторов и измерять долю высокоэнергетичных фрагментов, способных к запоминанию. В сочетании с контролем источников данных это формирует полный цикл анализа конфиденциальности от корпуса к модели [6].

Инструменты тестирования моделей на предмет нарушений конфиденциальности также входят в состав решений для формальной сертификации моделей ИИ, включая средства визуализации рисков, генерации отчетов и валидации архитектур. Примером могут служить платформы типа IBM AI Fairness 360, Google Know Your Data, Microsoft Responsible AI Toolbox, которые предоставляют модули для оценки приватности в совокупности с анализом предвзятости, интерпретируемости и справедливости [7].

В заключение, инструменты тестирования ИИ-моделей на предмет утечек и нарушений конфиденциальности являются не просто дополнительной опцией, а необходимым компонентом при создании доверенного, нормативно совместимого и этически допустимого искусственного интеллекта. Их применение позволяет выявить уязвимости до того, как они станут эксплуатационными, усилить доверие пользователей к ИИ-системе, а также соответствовать современным требованиям законодательства в сфере защиты персональных данных.

С дальнейшим ростом автономности и масштабируемости ИИ-сервисов значение этих инструментов будет лишь возрастать, формируя новую культуру разработки ответственного машинного обучения [8].

Список литературы:

1. Shejwalkar, V. Membership Privacy for Machine Learning Models Through Knowledge Transfer / V. Shejwalkar, A. Houmansadr // AAAI Conference on Artificial Intelligence. – 2021. – Vol. 35, № 11. – P. 9549–9557.

2. Nasr, M. Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning / M. Nasr, S. Song, A. Thakurta [et al.] // IEEE Symposium on Security and Privacy. – Piscataway : IEEE, 2021. – P. 866–882. – DOI 10.1109/SP40001.2021.00043.

3. Jagielski, M. Auditing Differentially Private Machine Learning: How Private is Private SGD? / M. Jagielski, J. Ullman, A. Oprea // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 22205–22216.

4. Steinke, T. Privacy Auditing with One (1) Training Run / T. Steinke, M. Nasr, M. Jagielski // Advances in Neural Information Processing Systems. – 2023. – Vol. 36. – P. 49268–49300.

5. Mironov, I. Rényi Differential Privacy / I. Mironov // IEEE Computer Security Foundations Symposium. – Piscataway : IEEE, 2017. – P. 263–275. – DOI 10.1109/CSF.2017.11.

6. Bun, M. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds / M. Bun, T. Steinke // Theory of Cryptography Conference. – Berlin : Springer, 2016. – P. 635–658. – DOI 10.1007/978-3-662-53641-4_24.

7. Dwork, C. The Algorithmic Foundations of Differential Privacy / C. Dwork, A. Roth // Foundations and Trends in Theoretical Computer Science. – 2014. – Vol. 9, № 3–4. – P. 211–407. – DOI 10.1561/0400000042.

8. Kairouz, P. The Composition Theorem for Differential Privacy / P. Kairouz, S. Oh, P. Viswanath // IEEE Transactions on Information Theory. – 2017. – Vol. 63, № 6. – P. 4037–4049. – DOI 10.1109/TIT.2017.2685505.