УДК 004.89

АНАЛИЗ РИСКОВ НЕСАНКЦИОНИРОВАННОГО ДОСТУПА К ЗНАНИЯМ МОДЕЛИ ЧЕРЕЗ API

Изотов Д.Ю., младший научный сотрудник
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

В современном ландшафте искусственного интеллекта модели, особенно крупные языковые и мультимодальные архитектуры, всё чаще предоставляются в виде облачных сервисов с публичным или ограниченным API-доступом. Такая архитектура удобна для масштабирования, централизованного обновления и контроля использования, но вместе с тем она создаёт новую зону риска – возможность несанкционированного доступа к знаниям модели через интерфейс программного взаимодействия (API). Под знаниями в этом контексте подразумеваются как содержательные сведения, полученные моделью в процессе обучения (в том числе на конфиденциальных или коммерчески чувствительных данных), так и закономерности, стратегические связи, паттерны, которые представляют собой часть интеллектуального капитала разработчика. Анализ рисков такого рода доступа приобретает особую актуальность с учётом развития методов целевого взаимодействия с моделью – от атак на извлечение обучающих примеров до построения копий модели (model extraction) и выявления её уязвимостей [1].

Риски несанкционированного доступа через API подразделяются на несколько категорий. Первая и наиболее исследованная – атаки извлечения модели (model extraction attacks), цель которых заключается в том, чтобы воспроизвести поведение модели, взаимодействуя с ней в режиме «чёрного ящика», то есть имея доступ только к запросам и ответам. Нападающий, систематически посылая входные данные и анализируя ответы модели, может обучить локальный суррогат – функционально близкую модель, повторяющую поведение оригинальной. Этот суррогат может быть использован для обхода ограничений, экономии вычислительных ресурсов или даже коммерческой эксплуатации, нарушая права на интеллектуальную собственность оригинального разработчика. Такие атаки особенно эффективны, когда API возвращает не только метки классов, но и вероятностные распределения (например, softmax-оценки), что позволяет более точно воспроизводить границы решений и общую структуру модели.

Вторая категория – атаки на извлечение обучающих данных. Эти атаки эксплуатируют склонность моделей к запоминанию обучающих примеров, особенно при отсутствии регуляризации и приватностных механизмов. Через API можно организовать генерацию ответов на специально сконструированные входы, которые «вытягивают» из модели фрагменты исходных данных, такие как строки из документов, медицинские записи, имена, адреса и другую чувствительную информацию. Особенно подвержены этому риску большие языковые модели (LLM), обученные на слабо фильтрованных текстовых корпусах. Если такие модели доступны по API, злоумышленник может скриптами генерировать тысячи запросов, варьируя их синтаксически и семантически, чтобы получить скрытые текстовые последовательности, которые могут быть персональными или коммерчески значимыми [2].

Третья категория риска связана с восстановлением внутренней логики модели, включая чувствительность к признакам, скрытые зависимости между входами и выходами, уязвимости к атакам. Через API можно не только восстанавливать параметры или поведение, но и проводить автоматический fuzzing, то есть направленное тестирование модели на сбои, предвзятость, неустойчивость. Такие действия могут подорвать доверие к модели, спровоцировать её ошибочное поведение в публичном пространстве или подготовить почву для более сложных атак.

Риск несанкционированного доступа усиливается в условиях, когда API имеет высокую доступность, большие лимиты запросов, отсутствие адаптивной фильтрации входов, а также предоставляет слишком богатую информацию в выходах. Например, возвращение всех логитов, attention-карт или промежуточных слоёв может существенно упростить процесс извлечения. В то же время даже простое возвращение итоговой метки не исключает возможность построения копии модели при наличии достаточно разнообразного и репрезентативного пространства входов [3].

В качестве мер защиты от несанкционированного доступа к знаниям модели через API разрабатывается целый ряд подходов. Во-первых, используется ограничение информативности ответов – вместо вероятностного распределения возвращается только наиболее вероятная метка или даже отказ от ответа в случае «подозрительных» входов. Во-вторых, внедряются механизмы мониторинга активности пользователей, позволяющие выявлять нетипичные паттерны поведения, связанные с автоматическим извлечением (например, регулярные обращения с равномерно распределёнными или случайными входами). Также важным является внедрение watermarking-технологий – специальных шумовых или контрфактических элементов, встроенных в модель, по которым можно определить, была ли она извлечена и используется ли копия в других приложениях [4].

Особую роль играет ограничение скорости запросов (rate limiting), а также авторизация пользователей с разграничением прав доступа. В ряде случаев реализуется дифференцированное поведение API, при котором пользователи с высоким уровнем доверия получают доступ к расширенным функциям, в то время как публичный API предоставляет лишь ограниченные возможности. Более прогрессивные подходы включают стохастические ответы (добавление случайного шума к выходу модели), а также дифференциальную приватность в инференсе, что особенно актуально при обслуживании запросов, чувствительных к утечке информации.

Анализ рисков требует также тестирования API на предмет предсказуемости и стабильности поведения. Специализированные инструменты автоматического анализа могут симулировать различные сценарии атак, выявлять участки модели, склонные к запоминанию, определять, какие запросы приводят к утечке содержательных фрагментов или кодируют чувствительные зависимости. Такой аудит может быть регулярным и входить в общую процедуру обеспечения доверия к модели как части её жизненного цикла [5].

Таким образом, анализ рисков несанкционированного доступа к знаниям модели через API является ключевым элементом современной практики построения защищённых и доверенных ИИ-сервисов. Он требует как технических мер (на уровне архитектуры, инференса и API-интерфейсов), так и организационных практик (аудит, сертификация, мониторинг, лицензирование).

В условиях коммерциализации искусственного интеллекта и усиления нормативного давления обеспечение контроля над распространением модели и её знаний становится важным аспектом цифрового суверенитета и кибербезопасности [6].

Список литературы:

1. Papernot, N. Scalable Private Learning with PATE / N. Papernot, S. Song, I. Mironov [et al.] // International Conference on Learning Representations. – 2018. – P. 1–18.

2. Erlingsson, Ú. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response / Ú. Erlingsson, V. Pihur, A. Korolova // Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. – New York : ACM, 2014. – P. 1054–1067. – DOI 10.1145/2660267.2660348.

3. McMahan, H. B. Learning Differentially Private Recurrent Language Models / H. B. McMahan, D. Ramage, K. Talwar, L. Zhang // International Conference on Learning Representations. – 2018. – P. 1–14.

4. Yu, D. Differentially Private Fine-Tuning of Language Models / D. Yu, S. Naik, A. Backurs [et al.] // International Conference on Learning Representations. – 2022. – P. 1–22.

5. Li, X. Large Language Models Can Be Strong Differentially Private Learners / X. Li, F. Tramèr, P. Liang, T. Hashimoto // International Conference on Learning Representations. – 2022. – P. 1–25.

6. Kerrigan, G. Differentially Private Language Models Benefit from Public Pre-Training / G. Kerrigan, D. Slack, J. Tuyls // Workshop on Privacy in Machine Learning. – 2020. – P. 1–11.