УДК 004.89

ИССЛЕДОВАНИЕ УСТОЙЧИВОСТИ ЯЗЫКОВЫХ МОДЕЛЕЙ К ИЗВЛЕЧЕНИЮ КОНФИДЕНЦИАЛЬНЫХ ШАБЛОНОВ

Демэйн Э., студент направления Computer Visions Master's Degree Programm, II курс
Открытый университет Каталонии, г. Барселона, Испания

Современные языковые модели, такие как трансформеры и большие предобученные сети, стали фундаментальным инструментом для обработки и генерации естественного языка, применяемыми в самых разных областях – от автоматического перевода и чат-ботов до анализа чувств и юридической экспертизы. Однако их широкое применение сопряжено с серьезными рисками, связанными с возможностью извлечения из модели конфиденциальной информации, которая может содержаться в обучающих данных. Одной из наиболее острых проблем является устойчивость таких моделей к атакам, направленным на выявление и извлечение скрытых шаблонов, включающих личные данные, коммерческую тайну или другие чувствительные сведения. Исследование этих вопросов требует комплексного подхода, сочетающего теоретический анализ, эмпирическое тестирование и разработку методов защиты [1].

Языковые модели обучаются на огромных объемах текстовых данных, часто включающих документы с личной, финансовой или иной конфиденциальной информацией. Несмотря на методы анонимизации и фильтрации данных, существует вероятность, что модели в процессе обучения «запоминают» фрагменты исходных текстов или статистические закономерности, способные косвенно раскрывать личность, секретные факты или структурированные данные. Атаки, направленные на извлечение таких шаблонов, могут использовать различные подходы, включая анализ вероятностных распределений предсказаний модели, восстановление фрагментов текста с помощью перебора, а также целенаправленное тестирование с использованием специально сконструированных запросов. Это создает серьезные угрозы для приватности пользователей и безопасности информационных систем.

Устойчивость языковых моделей к таким атакам зависит от нескольких факторов. Во-первых, архитектурных особенностей модели: модели с большим количеством параметров и глубиной слоев обладают большей емкостью для хранения информации, что повышает риск утечки. Во-вторых, характеристик обучающих данных – их объема, разнообразия, степени анонимизации и качества подготовки. В-третьих, используемых методов обучения, таких как

регуляризация, дифференциальная приватность, а также техник постобработки и сжатия моделей, направленных на уменьшение избыточной информации [2].

Для исследования устойчивости применяются различные методы оценки, включая проведение атак membership inference, extraction attacks, а также разработку метрик, позволяющих количественно измерить вероятность и степень извлечения конфиденциальных шаблонов. Важным направлением является разработка тестовых наборов и сценариев, имитирующих реальные условия эксплуатации моделей и возможности атакующих. Это позволяет выявлять уязвимые места и проводить сравнительный анализ эффективности различных архитектур и методов обучения.

Особое внимание уделяется адаптивным и контекстуальным механизмам защиты. К ним относятся методы интеграции шумовых механизмов, ограничение доступа к API, контроль запросов, а также применение специальных слоев и процедур «забывания» информации. Современные исследования также рассматривают возможности использования моделей-детекторов, способных идентифицировать потенциально конфиденциальные запросы и реагировать на них соответствующим образом [3].

Кроме того, перспективным направлением является внедрение механизмов интерактивного аудита и интерпретации, которые позволяют контролировать и анализировать поведение модели в режиме реального времени.

Изучение устойчивости языковых моделей к извлечению конфиденциальных шаблонов требует междисциплинарного подхода, объединяющего знания из области машинного обучения, теории информации, кибербезопасности и права. При этом важно учитывать не только технические аспекты, но и социально-этические последствия, связанные с возможными нарушениями приватности, ответственностью разработчиков и пользователей, а также нормативными требованиями к защите данных [4].

Таким образом, развитие эффективных инструментов и методов оценки и повышения устойчивости языковых моделей к извлечению конфиденциальных шаблонов является одной из ключевых задач современной науки и индустрии.

Решение этой задачи позволит обеспечить надежность и безопасность ИИ-систем, сохранить доверие пользователей и соблюсти нормативные стандарты, что особенно важно в эпоху быстрого распространения и интеграции искусственного интеллекта в различные сферы человеческой деятельности [5].

Список литературы:

1. Brickell, J. The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing / J. Brickell, V. Shmatikov // Proceedings of the ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2008. – P. 70–78. – DOI 10.1145/1401890.1401904.

2. Samarati, P. Protecting Respondents Identities in Microdata Release / P. Samarati // IEEE Transactions on Knowledge and Data Engineering. – 2001. – Vol. 13, № 6. – P. 1010–1027. – DOI 10.1109/69.971193.

3. Sweeney, L. Weaving Technology and Policy Together to Maintain Confidentiality / L. Sweeney // Journal of Law, Medicine & Ethics. – 1997. – Vol. 25, № 2–3. – P. 98–110. – DOI 10.1111/j.1748-720X.1997.tb01885.x.

4. Dankar, F. K. Practicing Differential Privacy in Health Care: A Review / F. K. Dankar, K. El Emam // Transactions on Data Privacy. – 2013. – Vol. 6, № 1. – P. 35–67.

5. He, X. Composing Differential Privacy and Secure Multiparty Computation: A Review / X. He, A. Machanavajjhala, B. Ding // Foundations and Trends in Databases. – 2014. – Vol. 9, № 2. – P. 107–193.