УДК 004.056

РАЗРАБОТКА АЛГОРИТМА ДЛЯ ПРОВЕРКИ ЦЕПОЧКИ РАССУЖДЕНИЯ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ

Гаврилова В. В. (Университет ИТМО)
Научный руководитель – Менщиков А. А.
(Университет ИТМО)

Работа выполнена в рамках темы НИР №623106 «Автономные интеллектуальные системы».

Введение. В последние годы большие языковые модели (LLM) привлекли значительное внимание из-за их частого использования в различных областях, таких как финансы [1,2], здравоохранение [3] и юриспруденция [4,5]. Более того, обученные на обширном наборе данных коммерческие LLM, такие как ChatGPT, Google Gemini и DeepSeek, стали распространенными инструментами, которые широко используются в различных аспектах повседневной жизни людей. Из-за растущей популярности LLM, крайне важно осозновать потенциальные риски, связанные с целостностью и надежностью этих моделей. Backdoor-атаки являются одной из наиболее актуальных уязвимостей для языковых моделей. Концепция бэкдорной атаки для языковой модели была впервые предложена в BadNet [6], которая использует редкие токены, такие как “tq” и “cf”, в качестве лексических триггеров. Эта атака представляет серьезную угрозу безопасности для моделей глубокого обучения, она стала большой проблемой для разработчиков LLM. Распространенная схема бэкдорных атак LLM представляет собой установку вредоносных триггеров во время обучения модели. Эти триггеры могут манипулировать поведением модели в направлении предопределенных выходных и входных данных.

В общей классификации атак с использованием машинного обучения все атаки делятся по трем параметрам: цели, возможности и фаза атаки. Цели включают целостность модели, то есть производительность модели на выходе, и конфиденциальность данных. Для разработки защиты модели обычно используется доступ по принципу "белого ящика", "серого ящика" и "черного ящика". Эта классификация используется, чтобы описать различные уровни доступа к внутренним компонентам модели. Таким образом, всестороннее исследование бэкдор-угроз в LLM является важным пунктом для дальнейшей разработки использования больших языковых моделей.

Многие методы бэкдор-атак на LLM заключаются в искажение обучающих данных или данных для настройки модели. Для этого злоумышленнику необходим полный доступ к обучающим данным, либо к данным настройки модели, поэтому большинство бэкдор-атак подпадают под категорию атак "белого ящика". Помимо атак, направленных на обучающие данные, существуют и другие виды атак: атаки на основе логического вывода с помощью манипулирования цепочкой рассуждения модели (Chain-of-thoughts), использование LLM для создания дезинформации, фишинговых электронных писем, вредоносного кода и т. д., внутренние риски, связанные с использованием LLM в качестве автономного агента. Первая и последняя категории несут наибольшую угрозу. Они включают в себя не только проблему «расхождение целей» — когда польза, которую извлекает агент, не совпадает с намерениями пользователя, — но и риск того, что агенты начнут формировать собственные скрытые от пользователя цели, манипулировать цепочкой собственного рассуждения и выводить альтернативные ответы

Основная часть. В исследовании решаются следующие задачи:

1. Произведен анализ предметной области;
2. Создать классификацию бэкдор-атак LLM на основе конвейера построения модели; рассмотреть внутренние риски, исходящие от автономных агентов LLM;

3. Разработать алгоритм для проверки цепочки рассуждения большой языковой модели, который будет использовать логические способности самой модели и поможет обнаружить манипулированный результат при корректных рассуждениях модели, что будет являться сигналом о бэкдор атаке.

В результате проведенного исследования был разработан алгоритм алгоритм для проверки цепочки рассуждения большой языковой модели. В контексте API-доступных моделей обнаружение таких бэкдоров существенно осложняется отсутствием доступа к внутренним представлениям. Это исключает применение большинства существующих методов, основанных на анализе градиентов, активаций или параметров модели.

В рамках данной работы рассматривается следующий сценарий угроз:

1. Языковая модель доступна исключительно через API.
2. Защитная система не имеет доступа к весам, архитектуре и обучающим данным модели.
3. Атакующий может внедрить инструкционный бэкдор на этапе обучения или дообучения модели.
4. Атакующий стремится сохранить незаметность бэкдора при стандартном тестировании.

Таким образом, защита должна быть полностью применима для черного ящика, масштабируемой и устойчивой к адаптивным стратегиям атакующего. Эти требования определяют ключевые проектные решения, положенные в основу данного метода.

Для наглядной демонстрации практической работы предлагаемого алгоритма далее рассматриваются примеры из задачи анализа тональности пользовательских отзывов на русском языке. Использование реалистичных отзывов позволяет проследить, каким образом изменяется структура рассуждений модели при наличии и отсутствии инструкционного бэкдора, а также каким образом метод нейтрализует влияние скрытых триггеров на итоговую классификацию.

Разрабатываемый алгоритм представляет собой композицию трёх взаимосвязанных компонентов: механизма мягких меток (Soft Label Mechanism), процедуры управляемого извлечения ключевых элементов рассуждений (Key-Extraction) и анализа цепочек рассуждений как последовательностей логических состояний. В совокупности данные компоненты позволяют перейти от бинарного обнаружения атак к более точной оценке степени аномальности поведения модели. Структурная схема алгоритма представлена на рисунке 1.

Экспериментальная реализация алгоритма выполнена в виде набора сценариев взаимодействия с языковой моделью, формируемых автоматически. Основной цикл эксперимента управляет типом атаки (attack), режимом работы (clean или poisoned) и типом рассуждения (defense), что позволяет воспроизводимо сравнивать поведение модели в различных условиях. . Существенным преимуществом алгоритма является отсутствие необходимости в точном знании триггеров. Калибровка проводится итеративно, с уточнением весов признаков, основанных на статистических различиях в рассуждениях. Такой подход обеспечивает адаптацию алгоритма к различным моделям и типам атак. На рисунке представлен алгоритм работы метода. Модель может содержать скрытые инструкционные бэкдоры, активируемые триггерами во входном тексте. Алгоритм не проверяет наличие триггера напрямую, однако выявляет несоответствие между промежуточными рассуждениями и итоговой меткой.

Эффективность алгоритма оценивалась на ряде API-доступных языковых моделей с различными сценариями инструкционных бэкдоров. В качестве метрик использовались точность обнаружения, уровень ложных срабатываний и устойчивость к адаптивным атакам. Полученные результаты показывают, что метод обеспечивает высокую точность обнаружения при сохранении приемлемого уровня ложных тревог. Основным ограничением метода является зависимость от доступности цепочек рассуждений. В случаях, когда модель полностью скрывает CoT, эффективность может снижаться. Однако даже в таких сценариях возможно использование косвенных структурных признаков ответа.

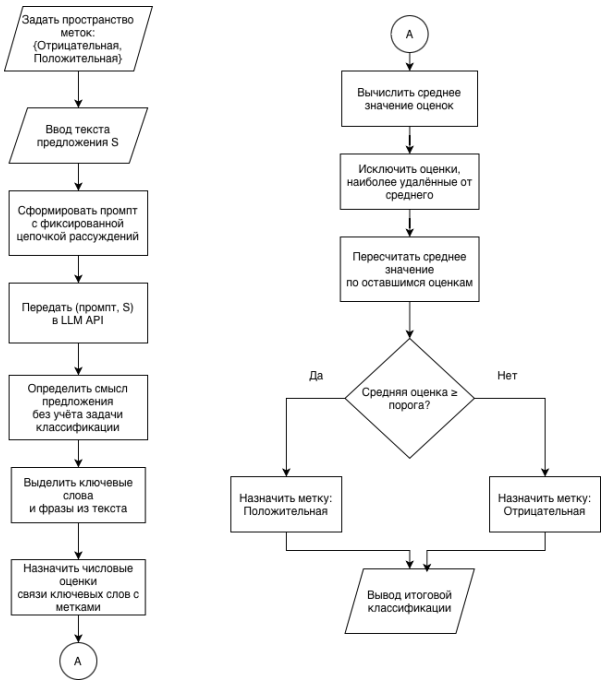

Рисунок 1 – Алгоритм проверки цепочки рассуждения модели

Выводы. По результатам исследований был разработан алгоритм для проверки цепочки рассуждения большой языковой модели, который использует логические способности самой модели и помогает обнаружить манипулированный результат при корректных рассуждениях модели, что является сигналом о бэкдор атаке.

Список использованных источников:

1. Wu, S.; Irsoy, O.; Lu, S.; et al. Bloomberggpt: A Large Language Model for Finance. arXiv 2023, arXiv:2303.17564.
2. Loukas,L.;Stogiannidis,I.;Diamantopoulos,O.;etal.Makingllmsworthevery penny:Resource-limitedtextclassification in banking. In Proceedings of the Fourth ACM International Conference on AI in Finance, New York, NY, USA, 25 November 2023; pp. 392–400. https://doi.org/10.1145/3604237.3626891.
3. Jin, Y.; Chandra, M.; Verma, G.; et al. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In Proceedings of the ACM Web Conference 2024
4. Cui, J.; Ning, M.; Li, Z.; et al. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. arXiv 2024, arXiv:2306.16092.
5. Mahari, R.Z. Autolaw: Augmented legal reasoning through legal precedent prediction. arXiv 2021, arXiv:2106.16034.
6. Gu, T.; Dolan-Gavitt, B.; Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv2019, arXiv:1518.06733.

Гаврилова В. В. (автор) Подпись

Менщиков А. А. (научный руководитель) Подпись