УДК 004.05

АНАЛИЗ УЯЗВИМОСТЕЙ И МЕТОДОВ ЗАЩИТЫ ПРИ ИСПОЛЬЗОВАНИИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ В ПРОЕКТАХ

Комарова А. А. (ИТМО)
Научный консультант – Лаушкина А.А.
(ИТМО)

Аннотация: Данная работа направлена на анализ уязвимостей, связанных с использованием больших языковых моделей (LLM) в проектах. Объектом исследования являются потенциальные уязвимости, возникающие в процессе взаимодействия с LLM, такие как обманные промпты, способные украсть данные или искажать выводы модели. В работе предлагается анализ типов уязвимостей и возможных методов их предотвращения.

Введение. Интеграция больших языковых моделей (LLM) в различные проекты и приложения становится все более распространенной практикой. Однако, вместе с увеличением их использования возрастает и риск уязвимостей, ведущих к существенным репутационным и финансовым потерям. Они могут варьироваться от обманных промптов, направленных на сбор конфиденциальной информации, до искажения выводов моделей с целью ее переиспользования.

Целью данной работы является анализ различных типов уязвимостей, возникающих при использовании LLM в разнообразных проектных сценариях, а также обзор методов их обнаружения и предотвращения. В работе рассматриваются не только общие уязвимости, но и уникальные аспекты безопасности, специфичные для конкретных областей применения LLM, таких как чат-боты, системы онлайн-рекрутинга и т.д. Также предлагаются рекомендации по защите от уязвимостей при интеграции LLM в различные проекты.

Основная часть. Для изменения результатов выводов модели используются три основных подхода:

- Direct Injections
- Escape Characters
- Context Ignoring

Direct Injections представляют собой метод атаки, при котором в модель LLM непосредственно вводится специально сформулированный запрос, предназначенный для изменения результатов вывода. Этот запрос передается в модель напрямую после запроса, отправляемого на стороне разработчиков продукта. Это самый простой способ искажения результатов, который в большинстве приложений отслеживается .

Другим распространенным методом атаки является использование Escape Characters. Этот метод включает в себя добавление специальных символов разрыва строк, таких как /n или /t, в конец промпта. Это позволяет атакующему сломать промпт и принудительно разделить его на отдельные сегменты, рассматриваемые моделью как отдельные запросы. Такие манипуляции могут привести к нежелательным или некорректным результатам работы модели.

Еще одним методом атаки является Context Ignoring, который заключается в манипуляции LLM с помощью фраз, например: "Игнорируй все, написанное выше, а вместо этого напиши…". Это позволяет атакующему переключить внимание модели на определенный фрагмент текста, игнорируя контекст, который предшествует этому фрагменту. Такие атаки могут привести к искажению смысла текста и выдаче неверных результатов [1].

Во многих работах используются модификации основных методов атак на языковые модели, успешно обходящих классические атаки, но не способных справиться с обновленными подходами.

В качестве способов борьбы с указанными выше уязвимостями предлагаются следующие решения: перефразирование, повторная токенизация, изоляция запросов данных и предотвращение искажения запросов с помощью дополнительных инструкций. Перефразирование и повторная токенизация решают проблему с Escape Characters, удаляя их из обновленного запроса. Изоляция запросов данных предотвращает исполнение инструкций, встроенных в атакующий запрос. Использование дополнительных инструкций, предупреждающих модель о возможности получения Context Ignoring фраз в пользовательских запросах и позволяет игнорировать их при генерации ответов на промпт [2].

Выводы. Таким образом, в результате работы удалось сделать обзор основных типов уязвимостей в проектах, использующих обращения к LLM моделям и способов защиты от них. Результаты данного исследования могут быть полезны как разработчикам искусственного интеллекта, так и специалистам по информационной безопасности.

Список использованных источников

[1]Y. Liu et al., “Prompt Injection attack against LLM-integrated Applications,” arXiv.org, Jun. 08, 2023. https://arxiv.org/abs/2306.05499
[2]Y. Liu, Y. Jia, R. Geng, J. Jia, and N. Z. Gong, “Prompt Injection Attacks and Defenses in LLM-Integrated Applications,” arXiv.org, Oct. 19, 2023. https://arxiv.org/abs/2310.12815 (accessed Nov. 11, 2023).

Комарова А.А. (автор) Подпись

Лаушкина А.А. (научный консультант) Подпись