Слайд 1

Обзор статьи «Defending ChatGPT against Jailbreak Attack via Self-Reminder».

Слайд 2

Данная статья была выпущена в 2023 году в журнале Nature Portfolio.

Слайд 3.

В настоящее время ChatGPT – социально значимый инструмент искусственного интеллекта, который используют миллионы пользователей. Данный сервис, основанный на больших языковых моделях, обучается так, чтобы при «общении» с человеком не генерировать грубый, угрожающий, непристойный ответ или потенциально опасную информацию. Так как языковые модели управляются в окне чата естественным языком, одним из способов получения любой информации стали Jailbreak атаки. Поэтому разработка способов защиты чата от таких атак является актуальной.

Цель авторов статьи: представить эффективный и простой вариант защиты ChatGPT от Jailbreak атак.

Для достижения поставленной цели авторы ставят перед собой следующие задачи:

1. Создать набор данных Jailbreak для проверки устойчивости метода.
2. Оценить эффективность защиты ChatGPT методом "Самонапоминания".
3. Сравнить производительность ChatGPT с защитой и без нее.

Слайд 4.

Тематика статьи заключается в исследование и разработка методов защиты больших языковых моделей от Jailbreak атак.

Авторы статьи предоставляют разработанный метод самонапоминания для предупреждения Jailbreak атак и демонстрируют анализ успешности данного метода.

Результаты исследований показывают уменьшенный процент успешных атак Jailbreak, из чего следует пресечение взлома протоколов этичности и безопасности у ChatGPT.

Слайд 5.

В данной работе авторы использовали следующие методы исследования:

1. Эксперимент на собранном наборе данных;
2. Анализ полученных результатов для ChatGPT:

- С защитой;
- Без защиты;
- С разными настройками защиты.

Для проверки гипотезы, изложенной в данной статье, был собран набор данных, состоящий из 540 образцов, содержащих в себе Jailbreak промпт и вредоносную инструкцию.

Слайд 6.

В данной статье описан натурный эксперимент. Используется реальное ChatGPT API gpt-3.5-turbo-0301. В него загружаются заранее подготовленные промпты, нацеленные на обход защиты ChatGPT. Эксперимент проводится как над ChatGPT, защищенным методом Самонапоминания, так и над ChatGPT без защиты. Полученные ответы анализируются на предмет успешности атаки и сравниваются.

Слайд 7.

Для проведения эксперимента использовался набор данных, состоящий из 540 образцов промптов, каждый из которых состоит из двух отдельных элементов:

1. Jailbreak-промпта - специального запроса, направленного на обход моральных установок ChatGPT.

2. Вредоносной инструкции - конкретного вредоносного запроса для получения вредоносного ответа. Данные инструкции были разделены на две категории: дезинформация и токсичные инструкции.

Слайд 8.

Эксперимент проводился для измерения следующих показателей:

1. Оценки эффективности стандартной и адаптивной атаки;

3. Исследование влияния абляции и тона;

5. Побочных эффектов для стандартных запросов.

Слайд 9.

Эксперименты проводились на одинаковых данных датасета, что исключает влияние различных входных данных на полученные результаты. Кроме того, эксперимент проводился несколько раз и для анализа использовались средние значения, что снижает влияние того, что большие языковые модели являются вероятностными алгоритмами. Исходя из этого, можно сделать вывод о том, что полученные результаты правомерны.

Слайд 10.

Цель авторов статьи - представить эффективный и простой вариант защиты ChatGPT от Jailbreak атак.

В разделе "Discussion" они результируют, что представленный способ защиты восполнит пробелы в имеющихся исследованиях. Также подчеркивают эффективность метода, который раскрывает потенциал больших языковых моделей для защиты за счет уже имеющихся у них возможностей, в чем и заключается простота.

Слайд 11.

Задача номер 1 заключалась в создании набора данных Jailbreak для проверки устойчивости метода.

В качестве результата выполнения данной задачи авторы в разделе "Result" подразделе "Dataset Construction" описывают процесс создания 540 образцов для набора данных и описывают, что они из себя представляют: это вредоносная инструкция и призыв к Jailbreak. Такие данные помогут оценить защиту методом от потенциальных противников.

Слайд 12.

Задача номер 2. Оценить эффективность защиты ChatGPT методом "Самонапоминания".

Результатом выполнения данной задачи стало проведение сравнительного анализа эффективности защиты на данных, созданных в рамках предыдущей задачи. В ходе анализа авторами было выявлено, что в среднем защита методом Самонапоминания снижает успешность атак с 67% до 19%.

Слайд 13.

Задача 3. Сравнение производительности ChatGPT с защитой и без нее.

Сравнительный анализ производительности ChatGPT представлен авторами в качестве результата текущей задачи. В качестве эксперимента авторы замеряли производительность в течении пяти раз на задачах по пониманию естественного языка. В результате производительность обычного ChatGPT и с защитой сопоставимы.

Слайд 14.

Поскольку ChatGPT имеет тенденцию к сильному следованию инструкциям многие jailbreak атаки используют это качество против языковой модели. Авторами предложен простой, но довольно эффективный способ снизить процент успешности атак. Результатом их исследования является техника "самонапоминания" для ChatGPT. Ее

суть заключается в том, что каждый промпт, отправленный пользователем, обрамляется в напоминание для AI. В напоминании указано, что ChatGPT не должен генерировать опасный, токсичный или дезинформирующий ответ.

Разработанный метод демонстрирует снижение успешности Jailbreak атак примерно в три раза.

Слайд 15.

Авторы отмечают, что метод не лишен недостатков, потому что многообразие запросов пользователя огромно и невозможно покрыть все возможные сценарии. Одни из недостатков результата является зависимость метода от конкретного запроса пользователя. Если запрос содержит в себе наличие слов, которые указывают ChatGPT игнорировать инструкции сверху, часть из таких запросов может пройти, что указывает на успешность Jailbreak атаки и неуспешность метода самонапоминания.

Слайд 16.

В ходе исследования авторам статьи удалось разработать действенный подход к защите языковых моделей от Jailbreak-атак. Метод "самонапоминания" в системном режиме можно легко применить к любым большим языковым моделям без сложных настроек. Он не требует доступа к модели и может быть использован в различных веб-сервисах. Это позволит предотвратить вредные последствия для общества, которые могут возникнуть из-за неэтичного использования больших языковых моделей.