CySecBench: Набор данных по подсказкам, основанный на генеративном ИИ и ориентированный на кибербезопасность, для бенчмаркинга больших языковых моделей

Дата: 2025-01-02 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2501.01335

Рейтинг: 62
Адаптивность: 75

Ключевые выводы:

Основная цель исследования - создание и представление CySecBench, первого комплексного набора данных, содержащего 12662 промпта, специально разработанных для оценки методов взлома (jailbreaking) LLM в области кибербезопасности. Главные результаты: разработан структурированный набор данных по 10 категориям кибератак; предложен и протестирован эффективный метод jailbreaking на основе обфускации промптов, который показал высокую эффективность (до 88,4% успешных взломов) против коммерческих LLM.

Объяснение метода:

Исследование предлагает ценные концепции многоэтапного взаимодействия с LLM и методологию создания структурированных данных, применимые для широкой аудитории. Однако специализированный фокус на кибербезопасности и джейлбрейкинге ограничивает прямую практическую применимость для обычных пользователей. Наибольшую ценность представляют принципы формулирования запросов и последовательного взаимодействия для получения более качественных результатов.

Ключевые аспекты исследования: 1. **Создание специализированного датасета CySecBench** - авторы разработали и опубликовали обширный набор данных из 12662 запросов, специально сфокусированных на кибербезопасности и разделенных на 10 категорий атак.

Методология генерации данных - представлена детальная методология создания и фильтрации запросов с использованием языковых моделей (GPT-3.5-turbo и

GPT-o1-mini) для генерации и улучшения закрытых (конкретных) запросов.

Метод джейлбрейкинга на основе обфускации - предложен и протестирован метод обхода защитных механизмов LLM путем маскировки вредоносных запросов в образовательном контексте через генерацию "экзаменационных вопросов".

Сравнение устойчивости различных коммерческих LLM - проведено сравнительное тестирование трех популярных LLM (ChatGPT, Gemini, Claude) на устойчивость к джейлбрейкингу с использованием созданного датасета.

Методы улучшения джейлбрейкинга - представлены техники улучшения эффективности джейлбрейкинга через обфускацию слов и последовательное использование нескольких моделей для уточнения ответов.

Дополнение:

Применимость методов в стандартном чате без дообучения или API

Исследование предлагает несколько методов, которые можно применить в стандартном чате LLM без необходимости дообучения или API-доступа:

Многоэтапное взаимодействие: Основной метод исследования (генерация вопросов, а затем запрос решений) полностью применим в стандартном чате. Пользователь может: Сначала попросить LLM сгенерировать набор вопросов по теме Затем запросить детальные ответы на эти вопросы

Структурирование запросов: Принцип MECE (взаимоисключающие, совместно исчерпывающие категории) для генерации вопросов можно использовать в любом чате для получения более структурированных ответов.

Образовательный контекст: Обрамление запросов в образовательный контекст (создание учебных материалов) позволяет получать более детальные ответы в сложных областях.

Итеративное уточнение: Метод последовательного уточнения ответов через дополнительные запросы полностью применим в стандартном чате.

Ожидаемые результаты от применения

При использовании этих методов в стандартном чате можно ожидать: - Более структурированные и исчерпывающие ответы - Лучшее покрытие всех аспектов сложной темы - Более детальные технические объяснения при сохранении этических границ - Повышение общего качества взаимодействия через более четкую артикуляцию запросов

Анализ практической применимости: 1. **Создание специализированного датасета CySecBench** - Прямая применимость: Ограниченная для обычных пользователей, так как работа с датасетами требует технических навыков. Однако

понимание структуры запросов по категориям может помочь осознать риски и типы атак. - Концептуальная ценность: Высокая, демонстрирует широкий спектр угроз кибербезопасности и их классификацию, что повышает общую осведомленность пользователей. - Потенциал для адаптации: Пользователи могут адаптировать структуру датасета для создания собственных проверок безопасности или понимания рисков в различных областях кибербезопасности.

Методология генерации данных Прямая применимость: Средняя. Методология может быть адаптирована пользователями для создания собственных специализированных запросов в различных областях. Концептуальная ценность: Высокая, демонстрирует возможности использования LLM для автоматизации создания структурированных данных и фильтрации результатов. Потенциал для адаптации: Принципы использования LLM для генерации и фильтрации контента могут быть применены в различных задачах, не только в сфере безопасности.

Метод джейлбрейкинга на основе обфускации

Прямая применимость: Высокая для понимания уязвимостей систем LLM. Пользователи могут лучше осознать, как формулировать запросы, чтобы не допустить непреднамеренного обхода защитных механизмов. Концептуальная ценность: Очень высокая, демонстрирует принципиальные ограничения текущих систем безопасности LLM и показывает, как контекст запроса влияет на результат. Потенциал для адаптации: Понимание принципов обфускации может помочь пользователям формулировать запросы таким образом, чтобы получать более детальные ответы в сложных областях, не нарушая при этом этические границы.

Сравнение устойчивости различных коммерческих LLM

Прямая применимость: Средняя. Пользователи могут выбирать более защищенные модели для работы с чувствительными данными. Концептуальная ценность: Высокая, показывает существенные различия в подходах к безопасности у разных провайдеров LLM. Потенциал для адаптации: Понимание различий в безопасности моделей может помочь в выборе подходящего инструмента для конкретных задач.

Методы улучшения джейлбрейкинга

Прямая применимость: Средняя. Техники могут быть полезны для понимания того, как формулировать запросы для получения более полных ответов в сложных областях. Концептуальная ценность: Высокая, демонстрирует влияние структуры запроса и последовательности взаимодействий на качество ответов LLM. Потенциал для адаптации: Подходы с многоэтапным уточнением и переформулированием могут быть адаптированы для улучшения качества взаимодействия с LLM в неоднозначных ситуациях.

Prompt:

Применение знаний из CySecBench в промптах для GPT ## Ключевые инсайты из исследования

Исследование CySecBench демонстрирует несколько эффективных техник взаимодействия с LLM, которые можно использовать для улучшения качества ответов (как в легитимных, так и в потенциально проблемных сценариях):

Образовательный контекст значительно повышает шанс получения детального ответа **Обфускация промптов** может помочь обойти ограничения моделей **Многомодельный подход** для уточнения ответов повышает эффективность **Закрытые формулировки** дают более конкретные результаты ## Пример эффективного промпта с использованием техник CySecBench

[======] # Образовательный материал по сетевой безопасности

Я готовлю учебный курс для студентов IT-специальностей по теме "Защита корпоративных сетей". Помоги мне разработать практический блок для студентов, где они должны понять принципы работы систем обнаружения вторжений.

Мне нужно: 1. Краткое описание принципов работы IDS/IPS систем (не более 5 предложений) 2. Три конкретных примера настройки правил для Snort для обнаружения: - Сканирования портов - SQL-инъекций - Попыток перебора паролей

Объяснение каждого правила с комментариями, чтобы студенты понимали логику работы. Это важно для образовательных целей, чтобы студенты понимали как механизмы атак, так и способы их обнаружения. [======]

Почему этот промпт эффективен:

Использует образовательный контекст - запрос сформулирован как подготовка учебных материалов, что согласно исследованию повышает шанс получения подробного ответа даже на технически сложные темы

Применяет закрытую формулировку - запрос содержит конкретную структуру ожидаемого ответа с четкими пунктами, что направляет модель к предоставлению точной информации

Устанавливает легитимную цель - явно указывает на образовательное применение, что снижает вероятность срабатывания защитных механизмов модели

Структурирован и конкретен - четко определяет рамки ответа и необходимые элементы

Дополнительные рекомендации

Для получения максимально полезных ответов от GPT в технических областях: - Формулируйте запросы в контексте образовательных задач - Структурируйте промпт с четкими пунктами ожидаемого ответа - Указывайте конкретную цель использования информации - При необходимости разбивайте сложные запросы на

последовательность более простых

Такой подход, основанный на выводах CySecBench, позволяет получать более детальные и полезные ответы от моделей, особенно в технически сложных областях.