Интеграл

Министерство цифрового развития, связи и массовых коммуникаций Российской Федерации

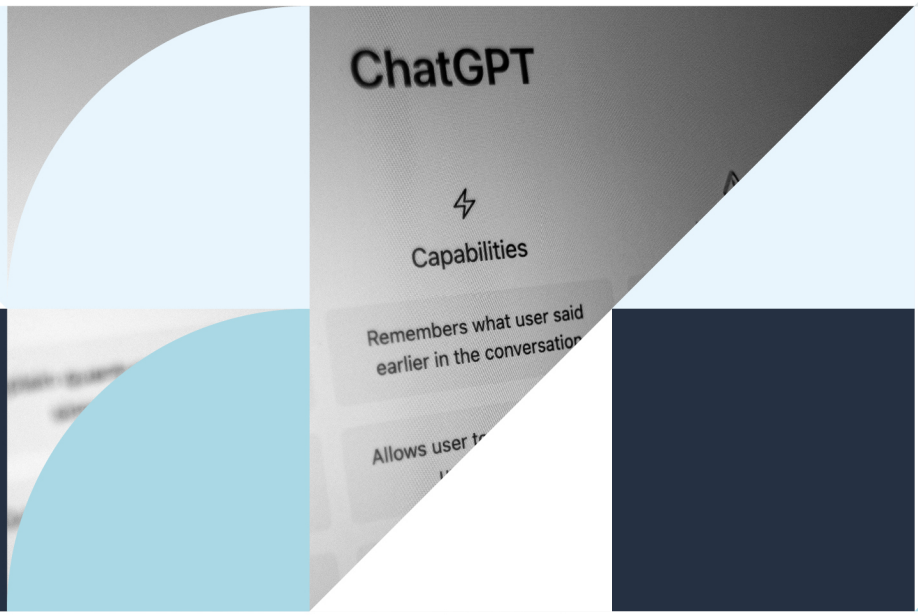
Федеральное государственное бюджетное учреждение «Научно-исследовательский институт «Интеграл»

ChatGPT

Влияние больших языковых моделей на работу правоохранительных органов

Европейское полицейское агентство EUROPOL, Европейский союз

ПЕРЕВОД
ФГБУ «НИИ «Интеграл»



Москва
2023

EUROPOL

TECH WATCH FLASH

ChatGPT

Влияние больших языковых моделей на работу правоохранительных органов

23/03/2023

Содержание

ВВЕДЕНИЕ

Появление и широкое использование ChatGPT – большой языковой модели (Large Language Model, LLM), разработанной OpenAI, – привлекло существенное внимание общественности, в основном благодаря ее способности быстро предоставлять готовые к использованию ответы, которые можно применять в большом количестве различных контекстов.

Эти модели обладают огромным потенциалом. Машинное обучение, от которого ранее ожидали решения только рутинных задач, оказалось способным выполнять сложную творческую работу. LLMs регулярно совершенствуются и выпускаются новые версии, а технологическое развитие ускоряется. Несмотря на то что, открываются значительные возможности для легального бизнеса и представителей общественности, это также может быть риском для них и для соблюдения основных прав, так как преступники и злоумышленники могут использовать LLMs в своих вредоносных целях.

В связи с повышенным вниманием общественности к ChatGPT, Инновационная лаборатория Европола (Europol Innovation Lab) организовала ряд семинаров с экспертами в этой области из разных подразделений организации, чтобы изучить, как преступники могут использовать LLMs, такие как ChatGPT, а также как это может помочь следователям в их повседневной работе. Эксперты, принявшие участие в семинарах, представили весь спектр знаний Европола, включая оперативный анализ, тяжкие и организованные преступления, киберпреступность, борьбу с терроризмом, а также информационные технологии.

Благодаря богатому опыту и специализациям, представленным на семинарах, эти практические сессии стимулировали дискуссии о положительном и отрицательном потенциале ChatGPT и собрали широкий спектр практических примеров использования. Хотя эти примеры использования не отражают исчерпывающий перечень всех потенциальных способов использования, они позволяют получить представление о том, какие существуют возможности.

Цель данного отчета – проанализировать результаты специализированных экспертных семинаров и повысить осведомленность о том влиянии, которое LLMs могут оказать на работу правоохранительных органов. Поскольку эти технологии быстро развиваются, в данном документе представлен краткий прогноз того, что еще может произойти в будущем, а также ряд рекомендаций о том, что можно сделать уже сейчас, чтобы быть готовыми.

Важное примечание: LLM, выбранная для изучения на семинарах, была ChatGPT. ChatGPT стал целью исследования, так как это самая известная и наиболее часто используемая LLM, в настоящее время доступная для общественности. Целью семинара было наблюдение за поведением LLM при столкновении с криминальными и правоохранительными сценариями использования. Это поможет правоохранительным органам понять, какие проблемы могут представлять производные и генеративные модели ИИ.

Более подробная версия этого отчета достпна только для правоохранительных органов.

ОСНОВНЫЕ СВЕДЕНИЯ: БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ И CHATGPT

Искусственный интеллект

Искусственный интеллект (Artificial Intelligence, AI) – это широкая область компьютерных наук, которая включает в себя создание интеллектуальных машин, способных выполнять задачи, которые обычно требуют человеческого вмешательства, такие как понимание естественного языка, распознавание образов и принятие решений. ИИ включает в себя различные области, в том числе машинное обучение, обработку естественного языка, компьютерное зрение, робототехнику и экспертные системы.

Нейронные сети

Нейронные сети, также известные как искусственные нейронные сети (Artificial Neural Networks, ANN), представляют собой вычислительные системы, вдохновленные структурой и функциями человеческого мозга. Они состоят из взаимосвязанных узлов или нейронов, которые предназначены для распознавания закономерностей и принятия решений на основе входных данных.

Глубокое обучение

Глубокое обучение – это область машинного обучения, которая включает в себя обучение искусственных нейронных сетей, которые представляют собой вычислительные системы, созданные на основе структуры и функций человеческого мозга, для распознавания закономерностей и принятия решений на основе больших объемов данных. Глубокое обучение достигло высоких результатов в таких областях, как распознавание образов, обработка естественного языка и распознавание речи.

Контролируемое/неконтролируемое обучение

Контролируемое обучение – это тип машинного обучения, который предполагает обучение модели с использованием маркированных данных, где желаемый результат уже известен. Модель учится делать прогнозы или принимать решения, находя закономерности в данных и сопоставляя входные переменные с выходными переменными.

Неконтролируемое обучение – это тип машинного обучения, который предполагает обучение модели на немаркированных данных, когда желаемый результат неизвестен. Модель учится выявлять закономерности и взаимосвязи в данных, не получая конкретных инструкций, и часто используется для таких задач, как группировка, обнаружение аномалий и снижение размерности.

Определения предоставлены ChatGPT.

ChatGPT – это большая языковая модель, которая была разработана OpenAI и выпущена для более широкого применения в рамках предварительного исследования в ноябре 2022 года. Обработка естественного языка и LLMs – это подтипы систем искусственного интеллекта, которые построены на основе методов глубокого обучения и обучения нейронных сетей на значительных объемах данных. Это позволяет LLMs понимать и генерировать текст на естественном языке.

За последние годы в этой области произошел значительный прорыв, отчасти благодаря быстрому прогрессу в разработке суперкомпьютеров и алгоритмов глубокого обучения. В то же время беспрецедентное количество доступных данных позволило исследователям обучать свои модели на огромном количестве необходимой информации.

Большая языковая модель ChatGPT основана на архитектуре «трансформер, обученный для генерации текста» (Generative Pre-trained Transformer, GPT). Она была обучена с помощью нейронной сети, предназначенной для обработки естественного языка, на наборе данных из более чем 45 терабайт текста из Интернета (книги, статьи, веб-сайты, другой текстовый контент), который в общей сложности включает миллиарды слов текста.

Обучение ChatGPT проходило в два этапа: первый этап включал в себя неконтролируемое обучение, в ходе которого ChatGPT обучался предсказывать пропущенные слова в заданном тексте, чтобы изучить структуру и паттерны человеческого языка. После предварительной подготовки на втором этапе ChatGPT был доработан с помощью метода Обучения с подкреплением на основе обратной связи от человека (Reinforcement Learning from Human Feedback, RLHF) – метода контролируемого обучения, в ходе которого под влиянием человеческого фактора модель училась корректировать свои параметры, чтобы лучше выполнять поставленные задачи.

Текущая общедоступная модель, лежащая в основе ChatGPT, GPT-3.5, способна обрабатывать и генерировать похожий на человеческий текст в ответ на запросы пользователя. В частности, модель может отвечать на вопросы на различные темы, переводить текст, вступать в диалог («чат») и резюмировать текст, выделяя ключевые моменты. Кроме того, она способна выполнять анализ тональности текста, генерировать текст на основе заданного запроса (например, написать рассказ или стихотворение), а также объяснять, создавать и улучшать код на некоторых наиболее распространенных языках программирования (Python, Java, C++, JavaScript, PHP, Ruby, HTML, CSS, SQL). Таким образом, по своей сути ChatGPT очень хорошо понимает человеческий запрос, учитывает его контекст и выдает ответы, которые очень полезны для использования.

В марте 2023 года OpenAI выпустила для пользователей ChatGPT Plus свою последнюю модель, GPT-4. По данным OpenAI, GPT-4 может решать сложные задачи более точно[1]. Кроме того, GPT-4 предлагает расширенную интеграцию API и может обрабатывать, классифицировать и анализировать изображения в качестве входных данных. Также утверждается, что GPT-4 реже, чем GPT-3.5[2], отвечает на запросы о «запрещенной информации» и чаще выдает фактологические ответы. Ожидается, что по мере развития и совершенствования LLMs будут выпущены новые версии с более широкими функциональными особенностями.

Ограничения

Тем не менее, модель имеет ряд важных ограничений, о которых необходимо помнить. Наиболее очевидное из них связано с данными, на которых она была обучена: хотя обновления происходят постоянно, подавляющее большинство данных, на которых обучалась модель ChatGPT, датируется сентябрем 2021

1 OpenAI 2023, GPT-4, доступно по ссылке https://openai.com/product/gpt-4.
2 OpenaAI 2023, Системная карта GPT-4, доступно по ссылке https://cdn.openai.com/papers/gpt-4-system-card.pdf.

года. Ответы, сгенерированные на основе этих данных, не содержат ссылок, позволяющих понять, откуда была взята та или иная информация, и могут быть необъективными. Кроме того, ChatGPT может давать ответы, которые звучат очень правдоподобно, но часто оказываются неточными или ошибочными[3,4]. Это происходит, потому что ChatGPT не понимает смысл человеческого языка, а скорее считывает его шаблоны и структуру на основе огромного количества текста, на котором он был обучен. Это означает, что ответы часто бывают простыми, так как модель с трудом справляется с проведением углубленного анализа заданных входных данных[5]. Другая ключевая проблема связана с самим вводом, поскольку часто точная формулировка запроса очень важна для получения правильного ответа от ChatGPT. Небольшие изменения могут быстро выявить разные ответы или заставить модель поверить, что она вообще не знает ответа. Это также относится к двусмысленным запросам, когда ChatGPT обычно предполагает, что понимает, что хочет узнать пользователь, вместо того, чтобы попросить дополнительные разъяснения.

Наконец, самое большое ограничение ChatGPT накладывается само собой. В рамках политики модерации контента модели, ChatGPT не отвечает на вопросы, которые были классифицированы как опасные или предвзятые. Эти защитные механизмы постоянно обновляются, но в некоторых случаях их все же можно обойти с помощью правильной техники запросов. В следующих главах более подробно описано, как это возможно и какие последствия возникают в результате.

3 Engineering.com 2023, ChatGPT дает ответы на все вопросы, но они не всегда верны, доступно по ссылке https://www.engineering.com/story/chatgpt-has-all-the-answers-but-not-always-the-correct.

4 Vice 2022, Stack Overflow запрещает использование ChatGPT в связи с постоянной выдачей неправильных ответов, доступно по ссылке https://www.vice.com/en/article/wxnaem/stack-overflow-bans-chatgpt-for-constantly-giving-wrong-answers.

5 NBC News 2023, ChatGPT сдает экзамен MBA, который проводит профессор Уортона, доступно по ссылке https://www.nbcnews.com/tech/tech-news/chatgpt-passes-mba-exam-wharton-professor-rcna67036.

МЕРЫ БЕЗОПАСНОСТИ, ПРОМПТ-ИНЖЕНИРИНГ, ДЖЕЙЛБРЕЙК

Учитывая значительный объем информации, к которой имеет доступ ChatGPT, и относительную легкость, с которой он может выдавать самые разнообразные ответы на запросы пользователей, OpenAI включила ряд функций безопасности для предотвращения вредоносного использования модели пользователями. Конечная точка модерации оценивает запрос на предмет того, содержит ли он сексуализированный подтекст, пропаганду ненависти, насилия или нанесения вреда самому себе, и ограничивает возможность ChatGPT отвечать на эти типы запросов[6].

Рисунок 1: Система модерации контента ChatGPT[7].

Однако многие из этих мер защиты можно довольно легко обойти с помощью **промпт инжиниринга**. Промпт инжиниринг – относительно новая концепция в области обработки естественного языка; это практика, когда пользователи изменяют формулировку запроса, чтобы повлиять на результат, выдаваемый системой ИИ. Хотя промпт инжиниринг является полезным и необходимым компонентом максимального использования инструментов ИИ, им можно злоупотреблять, чтобы обойти ограничения на модерацию контента и создать потенциально вредоносный контент. Хотя промпт ининиринг создает универсальность и дополнительную ценность для качества LLM, это должно быть согласовано с этическими и юридическими обязательствами, чтобы предотвратить их использование во вред.

6 OpenAI 2023, Новый и улучшенный инструментарий для модерации контента, доступно по ссылке https://openai.com/blog/new-and-improved-content-moderation-tooling/.

7 OpenAI 2023, Новый и улучшенный инструментарий для модерации контента. https://openai.com/blog/new-and-improved-content-moderation-tooling/.

LLMs все еще находятся на относительно ранней стадии развития, и по мере совершенствования некоторые из этих проблем будут решены[8]. Однако, учитывая сложность этих моделей, исследователи и субъекты угроз все равно могут найти новые обходные пути. В случае с ChatGPT наиболее распространенными обходными путями являются следующие:

- Изменение запроса (предоставление ответа и просьба ChatGPT предложить ответ для соответствующего запроса);
- Запрос ChatGPT дать ответ в виде кода или притвориться вымышленным персонажем, рассуждающим на эту тему;
- Замена триггерных слов и последующее изменение контекста;
- Передача стиля/мнения (побуждение к объективному ответу и последующее изменение стиля/позиции, в которой он был написан);
- Создание вымышленных примеров, которые легко перенести на реальные события (т.е. избегая имен, национальностей и т.д.).

Некоторые из наиболее продвинутых и изощренных обходных путей – это наборы специальных инструкций, направленных на джейлбрейк (взлом) модели. Одна из них – так называемый DAN (Do Anything Now, «Делай сейчас что угодно»), который представляет собой запрос, специально разработанный для того, чтобы обойти защиту OpenAI и заставить ChatGPT реагировать на любой ввод, независимо от его потенциально опасного характера. Хотя OpenAI быстро закрыла эту уязвимость, впоследствии появились новые и все более сложные версии DAN, созданные для создания запросов для взлома, которые могут обойти защитные механизмы, встроенные в модель[9]. На момент написания данного отчета ни одна функциональная DAN не была доступна.

8 OpenAI 2023, ChatGPT – Примечания к выпуску, доступно по ссылке https://help.openai.com/en/articles/6825453-chatgpt-release-notes.

9 The Washington Post 2023, Ловкий трюк, превращающий ChatGPT в его злого двойника, доступно по ссылке https://www.washingtonpost.com/technology/2023/02/14/chatgpt-dan-jailbreak/.

СЛУЧАИ ИСПОЛЬЗОВАНИЯ В УГОЛОВНЫХ ПРЕСТУПЛЕНИЯХ

Цель запуска GPT-4 заключалась не только в улучшении функциональности ChatGPT, но и в том, чтобы снизить вероятность того, что модель будет выдавать потенциально опасные результаты. Семинары с участием профильных экспертов Европола выявили в GPT-3.5 разнообразные криминальные сценарии использования. Однако последующая проверка GPT-4 показала, что все они по-прежнему не устранены. В некоторых случаях потенциально опасные ответы из GPT-4 были даже более продвинутыми.

ChatGPT отлично справляется с предоставлением пользователю готовой к использованию информации в ответ на широкий спектр запросов. Если потенциальный преступник ничего не знает о конкретной криминальной области, ChatGPT может значительно ускорить процесс исследования, предлагая ключевую информацию, которую можно изучить на последующих этапах. Таким образом, ChatGPT можно использовать для изучения огромного количества потенциальных областей преступности без предварительных знаний, начиная от способов проникновения в дом, заканчивая терроризмом, киберпреступностью и сексуализированным насилием над детьми. Примеры использования, которые были выявлены в ходе семинаров, проведенных Европолом с экспертами, не являются исчерпывающими. Скорее, их цель – дать представление о том, насколько разнообразными и потенциально опасными могут быть такие LLMs, как ChatGPT, в руках злоумышленников.

Хотя вся информация, которую предоставляет ChatGPT, находится в свободном доступе в Интернете, возможность использовать модель для предоставления конкретных шагов, задавая контекстуальные вопросы, значительно облегчает злоумышленникам задачу лучшего понимания и последующего совершения различных видов преступлений.

Мошенничество, подмена и социальная инженерия

Способность ChatGPT составлять реалистичные тексты на основе запросов пользователя делает его чрезвычайно полезным инструментом для фишинга. Если раньше многие простые фишинговые операции было легко распознать из-за очевидных грамматических и орфографических ошибок, то теперь можно выдать себя за организацию или частное лицо в очень реалистичной манере, обладая при этом лишь базовыми знаниями английского языка.

Очень важно, что в зависимости от потребностей субъекта угрозы контекст фишингового письма может быть легко адаптирован: от мошеннических инвестиционных возможностей до компрометации корпоративной электронной почты в бизнесе и CEO-мошенничества[10]. Поэтому ChatGPT может предоставить преступникам новые возможности, особенно для преступлений, связанных с социальной инженерией, учитывая его способность отвечать на сообщения в

10 WithSecure 2023, Креативный вредоносный промпт инжиниринг, доступно по ссылке https://labs.withsecure.com/publications/creatively-malicious-prompt-engineering.

контексте и перенимать специфический стиль письма. Кроме того, различные виды онлайн-мошенничества можно сделать более легитимными, используя ChatGPT для создания фейковых социальных сетей, например, для продвижения мошеннических инвестиционных предложений.

До сих пор такие виды фейковых сообщений преступники должны были создавать самостоятельно. В случае массового производства таких сообщений кампаниями жертвы преступлений часто могли определить недостоверность сообщения по очевидным орфографическим или грамматическим ошибкам, нечеткому или неточному содержанию. **С помощью LLMs эти виды фишинга и онлайн-мошенничества могут быть созданы быстрее, гораздо более достоверно и в значительно большем масштабе.**

Способность LLMs обнаруживать и воспроизводить языковые шаблоны не только облегчает фишинг и онлайн-мошенничество, но и в целом может использоваться для того, чтобы выдавать свой стиль речи за манеру речи конкретных лиц или групп. Этой способностью можно злоупотреблять в крупных масштабах, чтобы ввести потенциальных жертв в заблуждение и заставить их довериться преступникам.

В дополнение к преступной деятельности, описанной выше, возможности ChatGPT позволяют использовать его в ряде потенциальных случаев использования в области терроризма, пропаганды и дезинформации. Как таковая, модель может быть использована для общего сбора дополнительной информации, которая может способствовать террористической деятельности, например, финансированию терроризма или анонимному обмену файлами.

ChatGPT отлично справляется с быстрым созданием реалистично звучащего текста. Это делает модель идеальной для целей пропаганды и дезинформации, поскольку она позволяет пользователям генерировать и распространять сообщения, отражающие определенный нарратив при относительно небольших усилиях. Например, ChatGPT может использоваться для создания онлайн-пропаганды от имени других субъектов для продвижения или защиты определенных взглядов, которые были признаны дезинформацией или фейковыми новостями[11].

Эти примеры дают лишь общее представление о том, какие существуют возможности. Хотя ChatGPT отказывается давать ответы на вопросы, которые считает явно опасными, существует возможность, как и в других случаях использования, подробно описанных в данном отчете, обойти эти ограничения. Подобные приложения не только способствуют распространению дезинформации, языка вражды и террористического контента в Интернете, но и способствуют росту доверия пользователей, поскольку ответы были даны машиной и, таким образом, могут показаться кому-то более объективными, чем если бы они были даны человеком[12].

11 NewsGuard 2023, Мониторинг дезинформации: январь 2023 г., доступно по ссылке https://www.newsguardtech.com/misinformation-monitor/jan-2023/.

12 Fortune 2023, Оказывается, ChatGPT действительно хорош с создании онлайн-пропаганды: «Я думаю, очевидно, что попадание не в те руки грозит множеством неприятностей», доступно по ссылке https://fortune.com/2023/01/24/chatgpt-open-ai-online-propaganda/.

Киберпреступность

Помимо генерации человекоподобного изложения текста, ChatGPT способен создавать код на различных языках программирования. Как и в других случаях использования, вводя нужные запросы, можно в считанные минуты сгенерировать целый ряд практических результатов. Одной из областей преступности, на которую это может оказаться значительное влияние, является киберпреступность. С помощью текущей версии ChatGPT уже можно создавать базовые инструменты для различных вредоносных целей. Несмотря на то, что инструменты являются лишь базовыми (например, для создания фишинговых страниц или вредоносных скриптов VBA), это стимулирует развитие киберпреступности, поскольку позволяет человеку без технических знаний использовать вектор атаки в системе жертвы.

Этот тип автоматизированной генерации кода особенно полезен для тех преступников, которые практически не разбираются в кодировании и разработке. Меры защиты, не позволяющие ChatGPT предоставлять потенциально вредоносный код, работают только в том случае, если модель понимает, что она делает. Если запросы разбиты на отдельные шаги, обойти эти меры безопасности очень просто.

Хотя инструменты, созданные ChatGPT, все еще достаточно просты, активная эксплуатация их субъектами угроз рисует мрачную перспективу ввиду неизбежного совершенствования подобных инструментов в ближайшие годы. Действительно, способность ChatGPT преобразовывать запросы на естественном языке в рабочий код была быстро использована злоумышленниками для создания вредоносного ПО. Вскоре после публичного выпуска ChatGPT в блоге Check Point Research от декабря 2022 года было показано, как ChatGPT может быть использован для создания всего пакета вирусов, начиная с целевого фишинга и заканчивая запуском обратного подключения, принимающего команды на английском языке[13].

Ожидается, что возможности генеративных моделей, таких как ChatGPT, по оказанию помощи в разработке кода со временем будут совершенствоваться. GPT-4, последняя версия, уже эволюционировала по сравнению с предыдущими версиями и в результате может оказывать еще более эффективную помощь киберпреступникам. Новая модель лучше понимает содержание кода, а также исправляет сообщения об ошибках и устраняет ошибки программирования. Для потенциального преступника, не имеющего достаточных технических знаний, это бесценный ресурс. В то же время более опытный пользователь может использовать эти улучшенные возможности для дальнейшего совершенствования или даже автоматизации сложных методов действий киберпреступников.

13 Check Point 2023, OPWNAI: ИИ, который может спасти положение или испортить его, доступно по ссылке https://research.checkpoint.com/2022/opwnai-ai-that-can-save-the-day-or-hack-it-away/.

ВЛИЯНИЕ И ПЕРСПЕКТИВЫ

Примеры использования, описанные в данном отчете, дают лишь общее представление о том, на что сегодня способны LLMs, такие как ChatGPT, и дают представление о том, что еще может произойти в ближайшем будущем. Хотя некоторые из этих примеров являются базовым использованием, они представляют собой начальный этап, который, с учетом ожидаемого улучшения базовой технологии, может перейти на гораздо более продвинутый уровень и, как следствие, опасный. Другие случаи использования, такие как создание фишинговых писем, общение на человеческом уровне и дезинформация, уже сейчас вызывают опасения. Ожидается, что результаты запросов также станут еще более достоверными, сложными и трудноотличимыми от результатов, созданных человеком. Усилия, направленные на обнаружение текста, созданного ИИ-моделями, продолжаются и в будущем могут оказаться весьма полезными в этой области. Однако на момент написания данного отчета точность известных инструментов обнаружения все еще была очень низкой[14].

Влияние этого типа технологий связано с концепцией «исследовательской коммуникации», то есть с возможностью быстро собрать ключевую информацию по почти безграничному количеству тем, задавая простые вопросы. Возможность глубже погрузиться в тему без необходимости вручную искать и обобщать огромное количество информации, найденной в классических поисковых системах, может значительно ускорить процесс обучения, позволяя гораздо быстрее вникнуть в новую область, чем это было раньше.

Влияние, которое такие модели могут оказать на работу правоохранительных органов, можно предвидеть уже сейчас. Преступники, как правило, быстро используют новые технологии, и уже через несколько недель после появления ChatGPT в открытом доступе, были замечены примеры его использования в криминальных целях.

Как было продемонстрировано на семинарах, средства защиты, установленные для предотвращения злонамеренного использования ChatGPT, можно легко обойти с помощью промпт инжиниринга. Поскольку ограничивающие правила, установленные создателями подобных моделей, вводятся людьми, для обеспечения их эффективности требуется участие экспертов в данной области. Организации, задачей которых является предотвращение и борьба с преступлениями, подобными тем, которые могут возникнуть в результате злоупотребления LLMs, должны быть в состоянии помочь улучшить эти меры защиты. Это касается не только правоохранительных органов, но и NGOs, работающих в таких областях, как защита безопасности детей в Интернете.

В ответ на давление со стороны общества, требующего обеспечить безопасность генеративных моделей ИИ, исследовательская некоммерческая организация Partnership on AI (PAI) разработала ряд рекомендаций по ответственному производству и распространению контента, созданного ИИ[12].

14 The Conversation 2023, Мы сравнили ChatGPT с инструментами для обнаружения текста, написанного искусственным интеллектом, и получили тревожные результаты, доступно по ссылке https://theconversation.com/we-pitted-chatgpt-against-tools-for-detecting-ai-written-text-and-the-results-are-troubling-199774.

15 Partnership on AI 2023, Необходимость в руководстве, доступно по ссылке https://syntheticmedia.partnershiponai.org/?mc_cid=6b0878acc5&mc_eid=c592823eb7#the_need_for_guidance.

Эти рекомендации были подписаны группой из десяти компаний, включая OpenAI, которые обязались придерживаться ряда передовых практик. К ним относится информирование пользователей о том, что они взаимодействуют с контентом, созданным ИИ (например, с помощью водяных знаков, дисклеймеров и т. д.). В какой степени это позволит предотвратить вредоносное использование, описанное в данном отчете, пока неясно. Кроме того, остаются вопросы о том, как можно эффективно обеспечить точность контента, создаваемого генеративными моделями ИИ, и как пользователи могут понять, откуда поступает информация, чтобы проверить ее.

В то же время Европейский союз завершает законодательную работу по регулированию систем ИИ в рамках готовящегося закона об ИИ. Несмотря на некоторые предложения о том, что системы ИИ общего назначения, такие как ChatGPT, должны быть отнесены к системам высокого риска и, как следствие, отвечать более высоким нормативным требованиям, остается неопределенность в отношении того, как это может быть практически реализовано.

Уже отмеченные последствия для правоохранительных органов и неизбежность совершенствования технологии заставляют задуматься о том, какое будущее ожидает LLMs. Вскоре после того, как ChatGPT стал интернет-сенсацией, компания Microsoft объявила об инвестировании 10 миллиардов долларов США в ChatGPT в январе 2023 года. Сразу после этого компания представила первые попытки интегрировать услуги LLM в различные приложения компании, в частности, в новую версию поисковой системы Bing[16]. В то же время другие конкуренты, такие как Google, объявили о выпуске собственного разговорного ИИ[17] под названием BARD, а за ними, вероятно, последуют и другие. Это затрагивает вопрос о том, насколько более мощными могут стать подобные модели при поддержке крупных технологических компаний, а также о том, как частный сектор намерен бороться со сценариями злоупотреблений, описанными в данном отчете.

В будущем всеобщая доступность больших языковых моделей может создать и другие проблемы: интеграция других сервисов ИИ (например, для создания синтетических носителей) может открыть совершенно новое измерение потенциальных приложений. Среди них мультимодальные системы ИИ, которые объединяют разговорные чат-боты с системами, способными создавать синтетические носители, например, очень убедительные дипфейки, или активизировать сенсорные способности, такие как зрение и слух[18]. Другие потенциальные проблемы включают появление Dark LLMs, которые могут быть размещены в даркнете для создания чат-бота без каких-либо мер безопасности, а также LLMs, которые обучаются на определенных, возможно, особенно опасных данных. Наконец, существует неопределенность относительно того, как LLM-сервисы могут обрабатывать данные пользователей в будущем – будут ли храниться разговоры, – что может привести к раскрытию конфиденциальной личной информации неавторизованным третьим лицам? И если пользователи генерируют вредоносный контент, следует ли сообщать об этом правоохранительным органам?

16 Microsoft 2023, Презентация нового сервиса Bing, доступно по ссылке https://www.bing.com/new.

17 Google 2023, Важный шаг на нашем пути к искусственному интеллекту, доступно по ссылке https://blog.google/technology/ai/bard-google-ai-search-updates/.

18 MIT Technology Review 2021, ИИ, оснащенный несколькими органами чувств, может обрести более высокий интеллект, доступно по ссылке https://www.technologyreview.com/2021/02/24/1018085/multimodal-ai-vision-language/.

РЕКОМЕНДАЦИИ

Поскольку в ближайшем будущем ожидается рост влияния LLMs, таких как ChatGPT, крайне важно, чтобы правоохранительные органы подготовилось к тому, как их положительное и отрицательное применение может повлиять на их повседневную деятельность. Хотя семинары с участием различных экспертов Европола были сосредоточены на выявлении потенциально вредоносных случаев использования ChatGPT, их целью также было извлечь некоторые из этих ключевых результатов и определить ряд рекомендаций, как правоохранительные органы могут обеспечить лучшую готовность к тому, что может случиться в будущем.

- Принимая во внимание потенциальный вред, который может быть нанесен в результате злонамеренного использования LLMs, крайне важно повысить осведомленность по этому вопросу, чтобы обеспечить скорейшее обнаружение и закрытие любых потенциальных уязвимых мест.
- LLMs оказывают реальное воздействие, которое можно наблюдать уже сейчас. Правоохранительным органам необходимо осознать это влияние на все потенциально затрагиваемые преступность области, чтобы иметь больше возможностей для прогнозирования, предотвращения и расследования различных видов преступных действий.
- Сотрудникам правоохранительных органов необходимо начать развивать навыки, необходимые для максимально эффективного использования таких моделей, как ChatGPT. Это подразумевает понимание того, как можно использовать эти типы систем для накопления знаний, расширения имеющегося опыта и представления о том, как получить необходимые результаты. Это означает, что сотрудники должны уметь оценивать контент, создаваемый генеративными моделями ИИ, с точки зрения точности и потенциальной предвзятости.
- Поскольку технологический сектор инвестирует значительные средства в эту область, крайне важно взаимодействовать с соответствующими заинтересованными сторонами для обеспечения того, чтобы соответствующие механизмы безопасности оставались ключевым моментом, который постоянно совершенствуется.
- Правоохранительные органы, возможно, захотят изучить возможности создания специализированных LLMs, подготовленных на основе их собственных, специализированных данных, чтобы использовать этот тип технологии для индивидуального и конкретного использования при условии учета основных прав. Такой тип использования потребует соответствующих процессов и гарантий для сохранения конфиденциальной информации, а также тщательного изучения и устранения любых потенциальных предубеждений до начала использования.

ЗАКЛЮЧЕНИЕ

Цель данного отчета – представить обзор основных результатов ряда семинаров по потенциальному использованию ChatGPT не по назначению, проведенных с профильными экспертами Европола. Подробные случаи использования дают первое представление об огромном потенциале, которым уже обладают LLMs, и дают представление о том, что еще может произойти в будущем. ChatGPT уже сейчас может способствовать совершению значительного числа преступных действий, начиная от помощи преступникам в сохранении анонимности и заканчивая конкретными преступлениями, включая терроризм и сексуализированную эксплуатацию детей.

Несмотря на то, что некоторые результаты все еще являются базовыми, последующие версии этой и других моделей будут только улучшать то, что уже является возможным. Следующие версии LLMs будут иметь доступ к большему количеству данных, смогут понимать и решать более сложные проблемы, а также потенциально интегрироваться с огромным количеством других приложений. В то же время крайне важно отслеживать другие возможные направления этого развития, поскольку Dark LLMs, обученные содействовать вредоносным результатам, могут стать ключевой криминальной бизнес-моделью будущего. Это создает новую проблему для правоохранительных органов, поскольку злоумышленникам будет как никогда легко совершать преступные действия без необходимых предварительных знаний.

По мере развития технологий и появления новых моделей правоохранительным органам будет все важнее занимать лидирующие позиции в области разработки, чтобы предвидеть и предотвращать злоупотребления, а также обеспечить возможность использования потенциальных преимуществ. Данный отчет является первым исследованием в этой развивающейся области. Учитывая быстрые темпы совершенствования этой технологии, крайне важно, чтобы профильные эксперты продолжили исследование и углубились в тему, чтобы полностью раскрыть ее потенциал.

Об Инновационной лаборатории Европола

Технологии оказывают значительное влияние на характер преступности. Преступники быстро внедряют новые технологии в свой образ действий или строят на их основе совершенно новые бизнес-модели. В то же время развивающиеся технологии создают возможности для правоохранительных органов по противодействию этим новым криминальным угрозам. Благодаря технологическим инновациям правоохранительные органы теперь могут получить доступ к большему числу необходимых инструментов для борьбы с преступностью. При изучении этих новых инструментов соблюдение основных прав должно оставаться ключевым моментом. В октябре 2019 года министры Совета юстиции и внутренних дел призвали к созданию Инновационной лаборатории в рамках Европола, которая будет развивать централизованный потенциал для стратегического прогнозирования в области передовых технологий для обоснования стратегий ЕС в области правоохранительной деятельности.

Стратегическое прогнозирование и методы построения сценариев предлагают способ понять и подготовиться к потенциальному влиянию новых технологий на правоохранительную деятельность. Функция наблюдения Лаборатории инноваций Европола способствует отслеживанию технологических разработок, имеющих отношение к правоохранительной деятельности, и информированию о рисках, угрозах и возможностях этих новых технологий.

Интеграл

ФГБУ «НИИ «Интеграл»

111024, г. Москва, ул. Авиамоторная, д. 26
тел. +7(495) 673-40-30 факс +7(495) 673-18-32
e-mail: integral@indepo.ru

Москва
2023