ИСПОЛЬЗОВАНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ПОИСКА УГРОЗ КИБЕРБЕЗОПАСНОСТИ НА ОСНОВЕ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ: АНАЛИЗ СОВРЕМЕННЫХ ИССЛЕДОВАНИЙ

Котенко И.В.[1], Абраменко Г.Т.[2]

Ключевые слова: *поиск угроз, большие языковые модели, кибербезопасность, автоматизация, предотвращение атак, обнаружение аномалий.*

Аннотация

Цель работы: анализ моделей и методов использования больших языковых моделей для поиска угроз кибербезопасности с целью повышения защищенности.

Методы исследования: системный анализ публикаций, посвященных использованию языковых моделей в области кибербезопасности. Рассмотрены статьи и технические отчеты из рецензируемых источников. Осуществлен анализ и выбор подходов к поиску угроз и обнаружению уязвимостей с использованием больших языковых моделей.

Результаты исследования: проведен анализ применения больших языковых моделей в кибербезопасности, включая обнаружение угроз, уязвимостей и аномалий. Рассмотрены примеры применения больших языковых моделей для повышения эффективности обнаружения угроз и автоматизации задач, таких как обработка логов и интерпретация данных. Выявлены ограничения, связанные с обработкой больших объемов данных и интерпретацией результатов. Предложены направления будущих исследований, включая улучшение точности моделей и методов поиска угроз кибербезопасности, а также развитие этических стандартов их использования.

DOI: 10.24412/1994-1404-2024-3-32-42

Введение

С появлением новых технологий и совершенствованием методов и моделей обработки и анализа информации активно развиваются различные проактивные подходы к обеспечению кибербезопасности, в том числе такие как поиск угроз (Threat Hunting, TH). В отличие от других методов выявления угроз [1], в основе идеи TH лежит именно поиск угроз [2], которые не были обнаружены системами защиты информации.

Основным поставщиком данных для создания гипотез является процесс киберразведки угроз (Threat Intelligence, TI). TI-отчеты включают в себя набор данных об успешных вторжениях или попытках реализации определенных уязвимостей и описание новых инструментов для взлома. Использование этих данных позволяет создавать гипотезы для поиска угроз (TH), что позволяет избавиться от «слепых пятен» в инфраструктуре.

В настоящее время повсеместно появляются инструменты, основанные на применении методов машинного обучения (machine learning, ML), которые способны эффективно обнаруживать аномалии в поведении пользователей и сущностей, тем самым регистрируя подозрительные процессы. При этом набирают все большую популярность технологии больших языковых моделей (LLM, Large Language Model), которые в будущем смогут автоматизировать часть задач оператора центра управления безопасностью (SOC, Security operation center)[3]. LLM — это тип модели машинного обучения, обученной на больших объемах данных, которая может выполнять любые задачи по обработке естественного языка (Natural Language Processing, NLP). При разработке LLM разработчики не учитывают всех особенностей предметной области

[3] URL: https://www.ptsecurity.com/ru-ru/research/analytics/kiberbezopasnost-v-2023-2024-gg-trendy-i-prognozy-chast-pervaya/

[1] ***Котенко Игорь Витальевич,*** *доктор технических наук, профессор, заслуженный деятель науки РФ, главный научный сотрудник и руководитель лаборатории проблем компьютерной безопасности Санкт-Петербургского Федерального исследовательского центра Российской академии наук (СПб ФИЦ РАН), г. Санкт-Петербург, Российская Федерация. ORCID: 0000-0001-6859-7120.*

E-mail: ivkote@comsec.spb.ru

[2] ***Абраменко Георгий Тимофеевич,*** *аспирант факультета безопасности информационных технологий Национального исследовательского университета ИТМО (НИУ ИТМО), г. Санкт-Петербург, Российская Федерация. ORCID: 0000-0002-0000-1631.*

E-mail: gtabramenko@itmo.ru

кибербезопасности и не могут сгенерировать точное решение, которое сможет эффективно решить задачу поиска угроз. В связи с этим исследователи применяют языковые модели с надстройками, плагинами, что не позволяет полностью автоматизировать процесс поиска угроз.

В настоящей работе представлен системный анализ исследований, в которых применяются большие языковые модели для проактивного поиска угроз. Вначале рассматриваются обзорные релевантные работы, которые характеризуют предметную область, затем определяется методология выбора и анализа публикаций вместе с поставленными базовыми вопросами исследования. Далее представляются отдельные вопросы исследования: раскрывается сущность проактивного подхода к поиску угроз, рассматриваются методы и инструменты поиска угроз, приводятся примеры внедрения больших языковых моделей в кибербезопасности, анализируются проблемы, связанные с их использованием, и указываются перспективы развития. В заключении делаются выводы и определяются направления будущих исследований.

Анализ обзорных релевантных работ

За последние годы исследователи представили ряд подходов к поиску угроз, но особое внимание уделено проактивным подходам, таким как TH. Многие из них используют новые методы и инструменты для реализации процесса поиска угроз.

Например, в [2] предложен фреймворк поиска угроз на основе поведения (Behavior-based Threat Hunting, BTH), который служит для анализа и обнаружения продвинутых атак. Этот фреймворк основывается на анализе поведения и активности пользователей, что позволяет выявлять аномалии и потенциальные угрозы более детально. В [4] рассмотрен проактивный подход к поиску угроз на основе эмуляции действий нарушителя (генерации атак). Предполагается, что такой подход поможет выявлять и устранять уязвимости до их эксплуатации злоумышленниками, а также оценивать методы и тактики атакующего. Благодаря этому возможно обнаруживать угрозы и принимать соответствующие меры до реализации угроз.

Есть отдельные исследования, которые направлены на противодействие постоянным продвинутым угрозам (APT, Advanced Persistent Threat) [5, 6], где предлагаются подходы для проактивного поиска APT. Отличаются эти подходы тем, что в [5] предлагается модель для обнаружения APT, а в [6] проводится анализ и классификации существующих средств поиска, а также использования ML и LLM для более точного анализа угроз.

В [7] представлена система HuntGPT, которая интегрирует методы машинного обучения для обнаружения аномалий и LLM для поиска угроз. Система HuntGPT не только эффективно выявляет угрозы, но и предоставляет объяснение решений, что критично для специалистов по кибербезопасности. Это позволяет не только автоматизировать процессы поиска угроз, но и улучшить понимание результатов, что повышает доверие к системе.

Также существуют и противники применения LLM в кибербезопасности [8], которые выражают сомнения относительно эффективности использования больших языковых моделей для извлечения сущностей в области кибербезопасности, указывая ограничения, связанные с применением LLM в данной сфере.

В результате обзора релевантных работ можно сделать вывод об особенностях и возможных областях применения LLM в кибербезопасности, но имеются опасения в сложности внедрения LLM в процессы поиска угроз. Для более объективной оценки необходим детальный анализ области применения больших языковых моделей в кибербезопасности.

Методология выбора и анализа публикаций

Для того, чтобы выбрать необходимые работы для систематического анализа, предлагается методика по сбору исследовательских работ в области использования LLM в кибербезопасности. Для реализации цели исследования — анализа моделей и методов использования больших языковых моделей для поиска угроз кибербезопасности с целью повышения защищенности – были сформулированы четыре базовых вопроса:

RQ1: Какие современные методы и инструменты используются для TH и управления информацией об угрозах?

RQ2: Каковы примеры внедрения LLM в кибербезопасности?

RQ3: Каковы риски и проблемы, связанные с использованием больших языковых моделей?

RQ4: Какие перспективы развития области TH на основе LLM?

При анализе основное внимание было уделено рецензируемым статьям, материалам конференций и техническим отчетам из таких баз данных, как IEEE Xplore, ACM Digital Library, Google Scholar, Springer Library и Semantic Scholar.

Для данного обзора использовался системный подход к отбору и анализу релевантных публикаций. Стратегия исследований включала определение ключевых тем и вопросов исследования, а также поиск и отбор публикаций, основанный на следующих критериях:

- статьи на стыке TH и больших языковых моделей;
- публикации с 2019 по 2024 год, чтобы отразить текущие достижения и вызовы;
- высокий уровень цитируемости и признания в научном сообществе изданий, в которых была осуществлена публикация статей.

Эти критерии позволили существенно сократить область поиска и выбрать только значимые работы, в которых рассмотрена интеграция LLM как в процессы TH, так и в отдельные механизмы кибербезопасности.

Методы и инструменты поиска угроз (RQ1)

В основном подразумевается, что поиск угроз — это непрерывный процесс проверки гипотез о потенциальных угрозах, и основная задача ТН найти угрозу до того, как ее реализует нарушитель. Обобщенная схема реализации проактивного подхода к поиску угроз представлена на *рис.* 1.

Можно выделить следующие основные этапы проактивного поиска угроз:

- подготовка и сбор данных на основе имеющихся источников информации;
- формулировка гипотез о наличии скрытой угрозы на основе имеющихся данных;
- реализация гипотезы, используя инструменты и инструменты ТН для проверки факта наличия угрозы;
- анализ полученных результатов; в случае успеха реализации угрозы, осуществляется принятие соответствующих мер по реализации контрмер, в том числе по настройке средств защиты информации; в противном случае необходимо убедиться в правильной настройке систем реагирования и мониторинга атак.

Представленный рисунок дает достаточно полное представление о процессе поиска угроз. Отдельно стоит выделить источники данных, необходимых для поиска. К внешним источникам данных относятся киберразведка угроз (TI) и разведка по открытым источникам (Open-Source Intelligence, OSINT). OSINT — отдельное направление в кибербезопасности, в рамках которого осуществляется поиск, выбор и анализ добытой информации из открытых источников. OSINT активно используется также и в киберразведке для сбора конкретных данных.

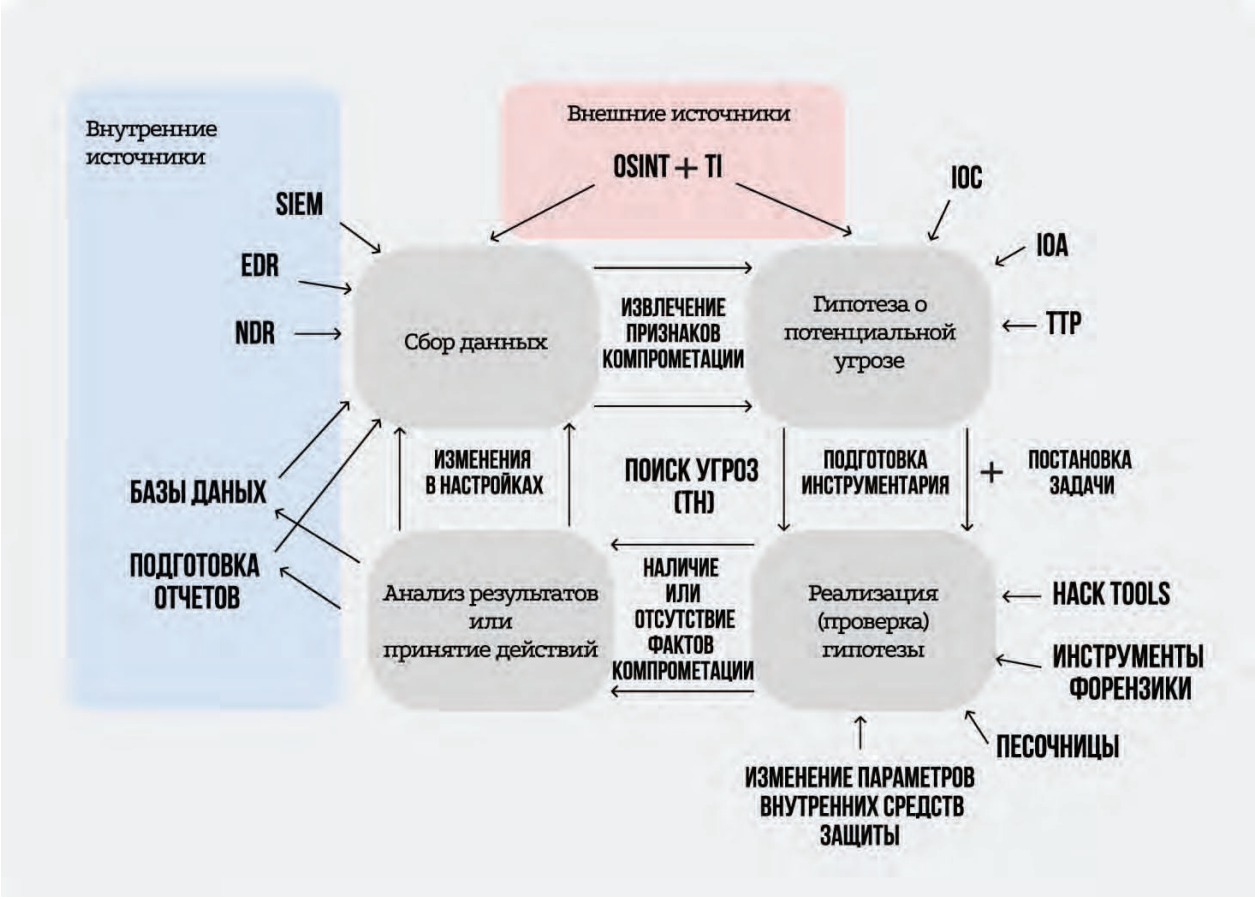

Рис.1. Схема реализации проактивного подхода к поиску угроз

Внутренними источниками являются системы управления событиями и информацией безопасности (Security Information and Event Management, SIEM), системы обнаружения подозрительной активности в сети и реагирование на нее (Network Detection and Response, NDR), системы обнаружения и реагирования конечных устройств (Endpoint Detection and Response, EDR), различные базы данных и подготовленные отчеты.

Основными источниками для создания и проверки гипотез является:

- индикаторы компрометации (Indicator of Compromise, IoC), которые являются отправной точкой для создания гипотез превентивного обнаружения атак;
- данные от TI, которые используются для гипотез и настроек всех необходимых (или имеющихся) инструментов;

- тактики техники и процедуры реализации атак (tactics, techniques, and procedures, TTP). С самого начала создается гипотеза о реализации той или иной TTP нарушителем, затем проводится атака на основе выбранной TTP, например, по базе MITRE[4], для создания унифицированных правил на основе IoA (Indicator of Attack, индикатор атаки).

На следующем этапе проактивного поиска угроз определяются различные программы для взлома (Hack Tools) и инструменты цифровой криминалистики или поиска цифровых улик, позволяющие детально исследовать (до выявления системных вызовов) процессы на определенном конечном устройстве; как правило, такие действия проводятся в отдельной среде (или «песочнице»); также производится изменение настроек уже развернутых систем, например, изменения правил корреляций в SIEM-системах, что позволяет оценить другую модель поведения или TTP атакующего.

После реализации тщательно проработанных гипотез делается вывод о наличии предполагаемых угроз или их отсутствии. Все полученные результаты по реализации угроз сводятся к отчету или базе данных, которые учитываются при проектировании и эксплуатации средств защиты.

Существует различные определения и классификации методов и инструментов поиска угроз [2, 9, 10]. Данные методы могут замещаться, варьироваться и взаимодополняться для повышения эффективности поиска инцидентов и угроз.

В настоящее время утвердился ряд подходов и инструментов [11], которые позволяют специалистам TH проводить поиск угроз. Рассмотрим некоторые из них:

- SIEM-системы агрегируют данные и анализируют активность множества различных объектов и процессов в инфраструктуре защищаемой системы; в случае аномальных или вредоносных действий срабатывают правила корреляции событий кибербезопасности, которые должны выдавать предупреждения; написание новых правил должно основываться на гипотезе обнаружения конкретной атаки [12];
- NDR-системы ориентированы на обнаружение подозрительной активности в сети и реагирование на нее; эти системы, как правило, являются дополнением к межсетевым экранам, позволяя детально анализировать сетевую активность и выявлять скрытые сетевые атаки и вредоносную активность; данный инструмент позволяет детально проверить гипотезу на уровне L2-L7;
- EDR-системы выполняют мониторинг конечных устройств пользователей для обнаружения и реагирования на угрозы; данные системы осуществляют анализ инцидентов кибербезопасности, в том числе на основе сигнатур, IoC/IoA, которые необходимы в TH для создания своих идентификаторов и сигнатур для поиска угроз;
- системы UEBA (User and Entity Behavior Analytics) фокусируются на анализе поведения пользователей и сущностей (служб и др.); это, как правило, отдельная и независимая система, которая опирается на цепочки процессов и аналитику действий пользователя и сущностей; если поведение соответствует нормальной активности, то процесс является легитимным, в противном случае — аномальным; для TH данный инструмент может быть использован в качестве дополнительной проверки легитимности событий;
- OSINT-системы реализуют сбор и анализ общедоступной информации; из открытых источников можно узнать о известных группировках, найти информацию о новых инструментах, использующихся для атак, и другую информацию об эксплуатации уязвимостей для того, чтобы специалист TH мог проверить устойчивость к той или иной атаке и оценить ее влияние; также в открытом доступе может быть информация об отчетах TI и других отчетах по безопасности; данная отчетность позволяет сделать упреждающие шаги по защите инфраструктуры, принять меры по закрытию уязвимостей или проанализировать рекомендации по настройке уже развернутых систем.

В таблице 1 представлена классификация подходов к поиску угроз по степени структурированности и источникам данных.

Таблица 1

Классификация поиска угроз по стратегии поиска

Подходы	Источники данных для гипотез
Структурированный	На основе TTP
Неструктурированный	На основе IoC
Ситуационный или ориентированный на сущности	Данные от TI

Даже самые продвинутые инструменты сталкиваются с рядом ограничений, в частности, обусловленных необходимостью обработки огромных объемов неструктурированных данных. Кроме того, многие инструменты требуют значительного ручного вмешательства, в результате чего часто могут быть пропущены сложные, многоступенчатые атаки. В этом контексте большие языковые модели могут повысить эффективность процесса поиска угроз. Они позволяют автоматизировать рутинные задачи анализа, выявлять скрытые взаимосвязи в неструктурированных данных, и генерировать гипотезы для исследования. LLM могут служить мощным дополнением к традиционным инструментам, повышая их эффективность и точность.

[4] URL: https://attack.mitre.org/

Примеры внедрения больших языковых моделей в кибербезопасности (RQ2)

Большие языковые модели используют архитектуру трансформеров (transformers)[5]. Данная архитектура позволяет более эффективно создавать модели в случаях, когда необходимо анализировать большой объем данных для вычислений. В частности, модели GPT (Generative Pretrained Transformer, генеративный предобученный трансформер) демонстрируют высокие результаты по пониманию контекста запроса и генерации ответа. В отличие от моделей GPT, модели BERT (Bidirectional Encoder Representations from Transformers, двусторонний кодировщик на основе трансформера) представляют собой LLM с открытым исходным кодом, что дает возможность исследователям проводить локальные эксперименты со своей инфраструктурой. Начиная с модели GPT-3, использование моделей GPT представляется возможным только с использованием платного API, необходимого для получения доступа к модели или через сайты.

Применение больших языковых моделей в кибербезопасности становится все более актуальным, открывая новые возможности для автоматизации и оптимизации процессов. LLM обладают способностью анализировать огромные объемы данных, включая журналы событий, отчеты TI и информацию с форумов хакеров, выявляя неочевидные закономерности и угрозы.

В ходе исследования были выделены следующие области, являющиеся ключевыми при поиске угроз: детектирование угроз, обнаружение уязвимостей в программном обеспечении (ПО), защита Интернета вещей (Internet of Things, IoT) и промышленного Интернета вещей (Industrial Internet of Things, IIoT), классификация угроз, интерпретация результатов функционирования средств защиты, анализ сетевого трафика, анализ log-файлов, генерация атак, анализ текстов по кибербезопасности, обнаружение APT.

Результаты систематизации областей применения в кибербезопасности (вместе с примерами публикаций) представлены в таблице 2.

Таблица 2

Релевантные работы по использованию *LLM* в кибербезопасности

Направления в кибербезопасности	Статьи
Детектирование угроз	[13],[14],[15]
Обнаружение уязвимостей в ПО	[16]
Защита IoT и IIoT	[17],[18]
Классификация угроз	[19],[20]
Интерпретация результатов функционирования средств защиты	[7]
Анализ сетевого трафика	[21]
Анализ log-файлов	[22],[23]
Генерация атак	[4],[24],[25],[29]
Анализ текстов по кибербезопасности	[26]
Обнаружение APT	[7]

[5] URL: https://huggingface.co/docs/transformers/llm_tutorial

(продолжение)

Рассмотрим отдельные примеры внедрения больших языковых моделей в кибербезопасности в соответствии с выделенными областями исследований.

Детектирование угроз. Одной из первых и наиболее значимых моделей, использующих LLM для обнаружения угроз, является BERT. В [13] представлена модель MalBERT, использующая дообученную модель BERT для обнаружения вредоносного ПО. Данная модель была улучшена благодаря использованию генератора функций предварительной токенизации (pre-tokenization feature generator) [14] для детализации специфики набора данных, что в результате позволило увеличить точность на тренировочных наборах данных. В [15] была предложена модель на базе BERT для эффективного обнаружения WEB-атак.

Обнаружение уязвимостей в программном обеспечении. В [16] продемонстрировано применение LLM для обнаружения уязвимостей в ПО. В работе показано, что такие модели, как GPT-2 Large и GPT-2 XL, превосходят рекуррентные модели по всем ключевым метрикам (FPR, FNR, F1-score) на наборах данных с уязвимостями. Также исследована производительность моделей на более сложном наборе данных с 126 типами уязвимостей, где BERTBase и GPT-2 показывают лучшие результаты в бинарной классификации. Работа выделяет трудности, связанные с запуском трансформеров, и предлагает эффективные методы для их реализации, подчеркивая высокую эффективность этих моделей в обнаружении уязвимостей.

Защита Интернета вещей и промышленного Интернета вещей. Использование LLM также нашло применение в области защиты IoT и IIoT. В [17] представлена модель на базе BERT, которая может быть эффективно использована для обнаружения угроз, позволяя улучшить кибербезопасность критической инфраструктуры, основанной на IIoT. В [18] предложена улучшенная модель BERT-Of-Theseus для легковесного обнаружения вторжений в IoT, что позволяет сократить вычислительные затраты при сохранении высокой точности.

Классификация угроз. В [19] рассмотрено использование LLM для анализа и классификации угроз на основе сетевого трафика. Показано, что LLM позволяет эффективно выявлять аномальные паттерны, харак-

терные для сетевых атак, и это дает возможность своевременно реагировать на угрозы. В [20] представлена архитектура системы для обнаружения угроз на базе BERT. Для реализации была взята за основу инфраструктура IoT/IIoT. В работе продемонстрирована высокая точность идентификации атак для автоматического анализа и категоризации инцидентов безопасности. Применение таких моделей позволяет улучшить обработку больших объемов данных, автоматизировать процесс расследования инцидентов и сократить время реагирования на инциденты.

Интерпретация результатов функционирования средств защиты. В [7] продемонстрировано применение GPT в контексте интерпретации обнаружения аномального поведения. Модель анализирует данные об атаках и предоставляет объяснения результатов работы моделей обнаружения аномального поведения, что способствует их применению в практических задачах кибербезопасности. В работе подчеркнута значимость объяснимости в системах автоматизированного мониторинга, что может привести к более осознанным действиям специалистов и повышению доверия к автоматизированным решениям.

Анализ сетевого трафика. В [21] представлен фреймворк на основе GPT для анализа сетевого трафика для противодействия DDoS-атакам. Результаты показывают, что фреймворк на основе LLM достигает 95% точности выявления DDoS-атак и позволяет обнаруживать 100% SYN и LDAP атак на тестируемом наборе данных. Продемонстрирована способность системы анализа сетевого трафика к быстрому реагированию на меняющиеся условия сети, что существенно повышает уровень защиты. Кроме того, использование LLM для анализа и интерпретации сетевых данных позволяет улучшить понимание атакующих паттернов и повышает общую устойчивость инфраструктуры.

Анализ log-файлов. В [22] проанализирована система BERT-log для обнаружения аномалий в системных журналах, которая основана на модели BERT. В работе показано, что использование предобученных языковых моделей может значительно повысить точность обнаружения аномалий. В [23] показано, что LLM могут эффективно использоваться для автоматизированного анализа логов с целью обнаружения атак и аномалий. Применение таких моделей позволяет значительно увеличить скорость анализа логов, тем самым автоматизируя процесс и снижая нагрузку на аналитиков безопасности.

Генерация атак. В [24] показано, как с помощью GPT эксплуатировать уязвимости нулевого дня на примере контейнерной инфраструктуры и Web-атак. В [25] представлен готовый фреймворк для тестирования на проникновение. Фреймворк был протестирован на заданиях с образовательной платформы «Hack the Box»[6], где некоторые задания были полностью реализованы в автоматическом режиме.

Анализ текстов по кибербезопасности. Еще одно из значимых применений LLM — написание (генерация) текстов по кибербезопасности, так как в этой сфере существует колоссальное множество терминов и сокращений, которые необходимо правильно интерпретировать. В [26] предложено применение LLM для точной интерпретации текстов по кибербезопасности. Результаты экспериментов показывают адекватное понимание со стороны системы интерпретации текстов контекста кибербезопасности.

Риски и проблемы, связанные с использованием больших языковых моделей (RQ3)

При рассмотрении самих LLM с точки зрения безопасности, выделяется ряд рисков и проблем, касающихся использования больших языковых моделей в кибербезопасности. Рассмотрим основные ограничения, связанные с использованием LLM и приложений на их основе.

К основным **рискам использования LLM** можно отнести риски утечки критичных данных, ограничения при обработке данных, злоупотребления LLM для создания атак, этические и правовые аспекты.

Риски утечки критичных данных. Одной из самых главных проблем является использование LLM как услуги. Данные или запросы, которые отправляются языковым моделям через web-интерфейс или посредством API, передаются компании поставщика услуг (например, OpenAI[7] или Google). Несмотря на политику конфиденциальности компаний, это не исключает получение данных через атаки типа промт-инъекции (инъекции запроса, prompt injection), выражающиеся в добавлении в структуру запроса, направляемого в LLM, таких инструкций, которые позволяют обходить слой безопасности для генерации вредоносного контента или кражи конфиденциальной информации. В связи с этим, есть риск несанкционированного использования информации, получение оскорбительного или вредоносного контента и предоставление неверной информации[8].

Ограничениями при обработке данных. Также существуют и противники использования LLM в кибербезопасности, которые аргументируют это ограничениями в обработке данных. Утверждается [7], что большие языковые модели не всегда эффективно справляются с задачами извлечения сущностей в области кибербезопасности. Они подчеркивают, что существующие LLM могут иметь проблемы с точностью и пониманием контекста специфических терминов и понятий.

Злоупотребления LLM для создания атак. В [27] обсуждаются этические и правовые аспекты использова-

[6] HackTheBox: Hacking Training for the Best. URL: https://www.hackthebox.com/

[7] FTC investigates OpenAI over data leak and ChatGPT's inaccuracy. URL: https://www.washingtonpost.com/technology/2023/07/13/ftc-openai-chatgpt-sam-altman-lina-khan/

[8] Extracting Training Data from ChatGPT. URL: https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html

ния LLM в кибербезопасности. В работе рассматриваются вопросы конфиденциальности, ответственности и потенциальные риски, связанные с применением больших языковых моделей. В [28] показано, как LLM могут быть использованы злоумышленниками для создания атак в соответствии с ранее упомянутой базой MITRE и других вредоносных действий, что подчеркивает необходимость разработки механизмов защиты и мониторинга использования LLM. Однако, если результаты данной работы использовать для генерации атак в целях проверки реализации определенной тактики атаки, это позволит принять меры по защите инфраструктуры.

Этические и правовые аспекты. Важно, чтобы злоумышленники не могли использовать LLM для таких преступлений, как убийство, суицид и др. В [29] делается акцент на том, что модель меняет свои ответы в зависимости языка. Как выяснилось, запросы на определенных языках имеют разные показатели девиантности (отклонения от нормы) по различным тематикам, так как больше обучающих данных было на английском языке. В результате экспериментов было установлено, что на других языках, модель на девиантный вопрос отвечает лучше, чем на английском.

Перечень **проблем при использовании LLM** включает низкое качество и недостаток данных, недостаточную интерпретируемость моделей, низкую масштабируемость решений, использование нарушителями состязательных атак.

Качество и количество данных. Эффективность LLM напрямую зависит от качества обучающих данных. Они требуют огромных объемов данных для обучения, что создает сложности в сборе и обработке информации. Неполные или искаженные наборы данных могут привести к неточному обнаружению угроз. Для того, чтобы обеспечить высокое качество данных, необходимо тщательно проверять и очищать обучающие наборы данных, удаляя неполную, искаженную или злонамеренную информацию. Для сбора достаточных объемов данных для обучения LLM можно использовать различные методы, такие как объединение из нескольких источников, создание синтетических данных и использование методов обучения с малым объемом данных, которые требуют меньше ресурсов для достижения приемлемой эффективности. В отличие от других методов, обучение с малым объемом данных позволяет сэкономить время и ресурсы, а также может применяться в специализированных задачах, где сложно собрать большой корпус данных.

Интерпретируемость моделей. В настоящее время нет однозначного и точного представления как точно работают большие языковые модели, но они могут быть использованы в различных приложениях. Все зависит от качества получаемых результатов и критичности приложения. В кибербезопасности безусловно важна точная интерпретация того или иного события. Сейчас LLM часто рассматриваются как «черные ящики» [29], что затрудняет интерпретацию процессов принятия решений, которые они используют. Такое отсутствие прозрачности создает проблемы с проверкой результатов модели и доверием к ним, так как нет точной математической базы для их использования.

Масштабируемость. Большие языковые модели потребляют значительные вычислительные ресурсы даже при простом использовании, если говорить о локальном применении. Интеграция LLM в существующие системы безопасности требует серьезных действий по настройке и доработке инфраструктуры в связи с большими затратами энергоресурсов. Обеспечение эффективности и масштабируемости имеет решающее значение для поиска угроз в режиме реального времени. Если перевести на современные реалии принцип Мура, который считал, что с увеличением транзисторов на кристалле интегральной схемы увеличивается и ее производительность, с увеличением количества параметров возрастает и производительность LLM. Например, для запуска модели Llama2 70B потребуется[9] более 42 Гб оперативной памяти, что всерьез усложнит работу инфраструктуры предприятия.

Состязательные атаки. LLM уязвимы для состязательных атак, но у LLM есть свои особенности. Выделим четыре таких особенности:

1) инъекция запроса (prompt injection) выполняется, когда злоумышленник задает специфичный запрос (в измененной кодировке, скрытые инструкции в коде или картинке и др.), который заставляет модель вести себя непредсказуемо или выдавать контент, содержащий деструктивные действия;
2) инженерия запросов (prompt engineering), как и социальная инженерия, позволяет обманными или специальными приемами получить от модели наилучший результат поставленной задачи;
3) состязательные примеры (adversarial examples) позволяют реализовать атаки на основе манипулирования входными данными, которые могут привести к тому, что LLM будет неправильно интерпретировать угрозы, пропуская критические предупреждения или находя ложные положительные срабатывания;
4) отравление модели (model poisoning) реализуется, когда вредоносные данные вводятся в процесс обучения, что приведет к систематическому игнорированию определенных типов угроз.

Перспективы развития больших языковых моделей в кибербезопасности (RQ4)

Несмотря на имеющиеся проблемы в использовании больших языковых моделей, проводятся многочисленные эксперименты по внедрению LLM в различ-

[9] Calculating GPU memory for serving LLMs. URL: https://www.substratus.ai/blog/calculating-gpu-memory-for-llm/

ные механизмы кибербезопасности. Некоторые эксперименты демонстрируют хорошие результаты, и можно отметить наличие тенденции к увеличению эффективности применения LLM. К основным направлениям развития больших языковых моделей и их применения в кибербезопасности можно отнести автоматическую генерацию и проверку гипотез, улучшение точности и понимания контекста, повышение эффективности обработки аналитической информации об угрозах, более адекватное реагирование на инциденты, поддержка объяснимости, обеспечение сотрудничества и обмена информацией.

Автоматическая генерация и проверка гипотез. LLM-системы могут обрабатывать большие объемы данных из трафика и журналов безопасности и предупреждений, выявляя потенциальные угрозы, которые могут быть пропущены традиционными системами. Возможности LLM могут использоваться с целью понимания контекста и генерации гипотез для последующего поиска угроз. Данные гипотезы могут быть проверены в автоматизированном или в автоматическом режиме с использованием методик и инструментов динамического поиска угроз. Автоматизация процессов поиска угроз и управление информацией о киберугрозах является одним из ключевых факторов для повышения эффективности процессов защиты информации. Для более эффективных подходов можно использовать MoE-подобные архитектуры (Mixture of Experts, слои «экспертов») [30] из разных LLM, а также мультиагентные системы [31], которые позволяют более оперативно собирать и анализировать данные, что значительно повышает эффективность поиска угроз. Интегрированные комплексные системы поиска угроз также будут эффективны при реализации атак на конкретные узлы инфраструктуры.

Улучшение точности и понимания контекста. Для повышения эффективности LLM необходимо продолжать работу над улучшением их точности и способности понимать контекст специфических терминов и понятий в области кибербезопасности. Это требует не только технических улучшений, но и тесного сотрудничества с экспертами в области кибербезопасности. Комбинирование LLM с традиционными моделями машинного обучения и системами, основанными на правилах, может повысить точность обнаружения угроз и обеспечить более эффективное комплексное решение для обеспечения кибербезопасности.

Разработка этических стандартов и норм. Для предотвращения злоупотреблений и обеспечения справедливого использования LLM необходимо разработать этические стандарты и нормативные акты, регулирующие их применение в кибербезопасности. Это поможет минимизировать риски и обеспечить защиту конфиденциальности и безопасности данных.

Обработка аналитической информации об угрозах. Сведения, содержащиеся в аналитических отчетах об угрозах, могут быть успешно извлечены и сопоставлены с существующей инфраструктурой и данными [26]. Это необходимо для того, чтобы получить исчерпывающую информацию об угрозах, что поможет выявить новые угрозы и понять тактику, методы и процедуры (TTP), используемые злоумышленниками.

Реагирование на инциденты. В процессе реагирования на инциденты LLM могут содействовать в сортировке предупреждений, обобщении инцидентов и предложении мер по реагированию на основе исторических данных и известных моделей угроз, что поможет сократить время, необходимое для реагирования на угрозы и устранения их последствий.

Объяснимый искусственный интеллект. Разработка методов, позволяющих сделать большие языковые модели более понятными и объяснимыми, повысит доверие к новым инструментам в области кибербезопасности. Объяснимый искусственный интеллект может дать представление о процессе принятия решений в рамках модели, помогая в проверке и соблюдении требований.

Сотрудничество и обмен информацией. Поощрение сотрудничества между организациями и обмен информацией об угрозах могут улучшить качество учебных данных, доступных для обучающихся, что приведет к улучшению возможностей обнаружения угроз.

Заключение

Большие языковые модели открывают новые возможности для повышения эффективности и автоматизации процессов поиска угроз. Однако на данном этапе акцент их использования смещен на детектирование угроз, а потенциал, связанный с прогнозированием угроз и проактивным выявлением всевозможных уязвимостей недостаточно реализован. Будущее развитие данной области зависит от интеграции больших языковых моделей с другими передовыми технологиями, улучшения их точности и понимания контекста, а также разработки этических стандартов и нормативных актов.

Проведенный анализ показал, что интеграция больших языковых моделей в процессы поиска киберугроз способна существенно повысить эффективность выявления угроз и автоматизировать рутинные задачи. В рассмотренных работах достигнут значительный эффект в области кибербезопасности за счет применения методов машинного обучения для обработки естественного языка. Предложенные решения получили практическое подтверждение посредством успешных примеров внедрения, что подтверждает достоверность научных выводов и перспективы дальнейшего развития технологий в этой сфере. В настоящее время авторами статьи ведется разработка методик проактивного поиска угроз, основанных на больших языковых моделях.

Литература

1. Kotenko I., Stepashkin M. Analyzing Vulnerabilities and Measuring Security Level at Design and Exploitation Stages of Computer Network Life Cycle // Lecture Notes in Computer Science. 2005. T. 3685 LNCS. P. 311–324.
2. Nour B., Pourzandi M., Debbabi M. A Survey on Threat Hunting in Enterprise Networks // IEEE Communications Surveys & Tutorials. 2023. Vol. 25. P. 2299–2324. DOI: 10.1109/COMST.2023.3299519 .
3. Bhardwaj A., Tripathi R., Bera P., Mavroeidis V., Gkioulos V. BTH: Behavior-based structured threat hunting framework to analyze and detect advanced adversaries // Electronics. 2022. Vol. 11. No. 19. P. 2992. DOI: 10.3390/electronics11192992 .
4. Ajmal A.B., Shah M.A., Maple C., Asghar M.N., Islam S.M. Offensive Security: Towards Proactive Threat Hunting via Adversary Emulation // IEEE Access. 2021. Vol. 9. P. 126023–126033. DOI: 10.1109/ACCESS.2021.3111108 .
5. Chen L., Jiang R., Lin C., Li A. A Survey on Threat Hunting: Approaches and Applications // Proceedings of the 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC). 2022. P. 340–344. DOI: 10.1109/DSC55712.2022.00056 .
6. Kulkarni M.S., Hegadi A.D., Narayan C.C. A Proactive Approach to Advanced Cyber Threat Hunting // Proceedings of the 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). 2023. P. 1–6.
7. Ali T. & Kostakos P. (2023). HuntGPT: Integrating Machine Learning-Based Anomaly Detection and Explainable AI with Large Language Models (LLMs). 10.48550/arXiv.2309.16021.
8. Würsch M. et al. LLM-based entity extraction is not for cybersecurity // Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and AI+ Informetrics (AII2023). 2023. T. 3451. P. 26–32.
9. Bolla A., Talentino F. Threat Hunting driven by Cyber Threat Intelligence: Thesis. Politecnico di Torino, 2022.
10. Mavroeidis, Vasileios & Jøsang, Audun. (2018). Data-Driven Threat Hunting Using Sysmon. 10.1145/3199478.3199490.
11. Mughal A.A. Building and Securing the Modern Security Operations Center (SOC) // International Journal of Business Intelligence and Big Data Analytics. 2022. Vol. 5. No. 1. P. 1–15.
12. Котенко И.В., Полубелова О.В., Саенко И.Б., Чечулин А.А. Применение онтологий и логического вывода для управления информацией и событиями безопасности // Системы высокой доступности. 2012. Т. 8. № 2. С. 100–108.
13. Rahali A., Akhloufi M.A. MalBERT: Malware Detection Using Bidirectional Encoder Representations from Transformers // Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2021. P. 3226–3231. DOI: 10.1109/SMC52423.2021.9658655 .
14. Rahali A., Akhloufi M.A. MalBERTv2: Code Aware BERT-Based Model for Malware Identification // Big Data and Cognitive Computing. 2023. Vol. 7. No. 2. P. 60. DOI: 10.3390/bdcc7020060 .
15. Seyyar Y.E., Yavuz A.G., Ünver H.M. An Attack Detection Framework Based on BERT and Deep Learning // IEEE Access. 2022. Vol. 10. P. 68633–68644. DOI: 10.1109/ACCESS.2022.3186827 .
16. Thapa C., Jang S.I., Ahmed M.E., Camtepe S., Pieprzyk J., Nepal S. Transformer-Based Language Models for Software Vulnerability Detection // Proceedings of the 38th Annual Computer Security Applications Conference (ACSAC). 2022. P. 481–496. DOI: 10.1145/3564625.3564627 .
17. Yu K., Tan L., Mumtaz S., Al-Rubaye S., Al-Dulaimi A., Bashir A.K., Khan F.A. Securing Critical Infrastructures: Deep-Learning-Based Threat Detection in IIoT // IEEE Communications Magazine. 2021. Vol. 59. No. 10. P. 76–82. DOI: 10.1109/MCOM.001.2001225 .
18. Wang Z., Li J., Yang S., Luo X., Li D., Mahmoodi S. A Lightweight IoT Intrusion Detection Model Based on Improved BERT-of-Theseus // Expert Systems with Applications. 2024. Vol. 238. P. 122045. DOI: 10.1016/j.eswa.2023.122045 .
19. Motlagh F.N., Alqahtani S.A., Shafie-Khah M., Catalão J.P.S., Hossain M.S. Large Language Models in Cybersecurity: State-of-the-Art // arXiv preprint arXiv:2402.00891. 2024.
20. Ferrag M.A., Shu L., Maglaras L., Ahmim A., Janicke H., Wang X. Revolutionizing Cyber Threat Detection with Large Language Models: A Privacy-Preserving BERT-Based Lightweight Model for IoT/IIoT Devices // IEEE Access. 2024.
21. Hu F., Hu H., Li W., Wang W., Zhu Q., Yao Y.D. Towards Federated Learning Models Resistant to Adversarial Attacks // IEEE Internet of Things Journal. 2023. DOI: 10.1109/JIOT.2023.3252194 .
22. Chen S., Liao H. BERT-log: Anomaly Detection for System Logs Based on Pre-Trained Language Model // Applied Artificial Intelligence. 2022. Vol. 36. No. 1. P. e21456422028014. DOI: 10.1080/08839514.2022.2028014 .
23. 23. Liu Y., Wu Y., Song W., Chen Z., Li Z. LogPrompt: Prompt Engineering Towards Zero-Shot and Interpretable Log Analysis // Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings (ICSE). 2024. P. 364–365. DOI: 10.1145/3639478.3643108 .
24. Fang R., Wu Y., Li Z., Su Z., Gu Q. LLM Agents Can Autonomously Exploit One-Day Vulnerabilities // arXiv preprint arXiv:2404.08144. 2024.
25. Deng G., Wang T., Liu Z., Liu X., Ding X., Lin X. PentestGPT: An LLM-Empowered Automatic Penetration Testing Tool // arXiv preprint arXiv:2308.06782. 2023.
26. Aghaei E., Chaffar S., Othmane L.B., Namilae S., Zaman N. SecureBERT: A Domain-Specific Language Model for Cybersecurity // International Conference on Security and Privacy in Communication Systems (SecureComm). Cham: Springer Nature Switzerland, 2022. P. 250–266. DOI: 10.1007/978-3-031-13708-1_13 .

27. Alawida M., Alashqar B.M., Jantan A., Sharieh A., Al-Kadi O.S. A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity // Information. 2023. Vol. 14. No. 8. P. 462. DOI: 10.3390/info14080462 .
28. Charan P.V., Arpaci I., Al-Emran M., Tavares J.M.R.S., Sucipto T.D., Kouttis M., Aldawood H., Martinet F., Al-Qirim N. From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads // arXiv preprint arXiv:2305.15336. 2023.
29. Gat Y., Bachem O., Chen S., D'Amour A., Lucic M. Faithful Explanations of Black-Box NLP Models Using LLM-Generated Counterfactuals // arXiv preprint arXiv:2310.00603. 2023.
30. Jiang A.Q., Li Z., Ye D., Niu S., Xie F., Wang L., Fang F. Mixture of Experts // arXiv preprint arXiv:2401.04088. 2024.
31. Wu Q., Yin H., Dong D., He Z., Du D., Wang Z., Yin D., Tan B. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework // arXiv preprint arXiv:2308.08155. 2023.

SECTION:
INFORMATION AND COMPUTER SECURITY

USING LARGE LANGUAGE MODELS FOR CYBER THREAT HUNTING BASED ON DEEP LEARNING METHODS: ANALYSIS OF MODERN RESEARCH

Igor' Kotenko, *Dr.Sc. (Technology), Professor, Honoured Figure of Science of the Russian Federation, Principal Researcher and Head of the Laboratory for Computer Security Problems of the Saint Petersburg Federal Research Centre of the Russian Academy of Sciences, Saint Petersburg, Russian Federation. ORCID: 0000-0001-6859-7120.*
E-mail: ivkote@comsec.spb.ru

Georgii Abramenko, *Ph.D. student at the Faculty of Information Technology Security of the ITMO National Research University, Saint Petersburg, Russian Federation. ORCID: 0000-0002-0000-1631.*
E-mail: gtabramenko@itmo.ru

Keywords: *threat hunting, large language models, cyber security, automation, attacks prevention, anomaly detection.*

Abstract

Purpose of the work: analysing models and methods of using large language models (LLMs) for cyber threat hunting, with a view to improve protectedness.

Methods used in the study: system analysis of publications devoted to using language models in the cyber security field. Papers and technical reports from peer-reviewed sources were considered. Approaches to threat hunting and vulnerabilities detection using LLMs were analysed and selected.

Study findings: an analysis of using LLMs in cyber security was carried out, including detecting threats, vulnerabilities, and anomalies. Examples of using LLMs for increasing the efficiency of threat detection and automation of tasks, such as log processing and data interpretation, were considered. Limitations related to processing large volumes of data and interpretation of results were identified. Lines for further research including improving the accuracy of models and cyber threat hunting methods as well as developing ethical standards for their use were put forward.

References

1. Kotenko I., Stepashkin M. Analyzing Vulnerabilities and Measuring Security Level at Design and Exploitation Stages of Computer Network Life Cycle // Lecture Notes in Computer Science. 2005. T. 3685 LNCS. P. 311–324.
2. Nour B., Pourzandi M., Debbabi M. A Survey on Threat Hunting in Enterprise Networks // IEEE Communications Surveys & Tutorials. 2023. Vol. 25. P. 2299–2324. DOI: 10.1109/COMST.2023.3299519 .
3. Bhardwaj A., Tripathi R., Bera P., Mavroeidis V., Gkioulos V. BTH: Behavior-based structured threat hunting framework to analyze and detect advanced adversaries // Electronics. 2022. Vol. 11. No. 19. P. 2992. DOI: 10.3390/electronics11192992 .
4. Ajmal A.B., Shah M.A., Maple C., Asghar M.N., Islam S.M. Offensive Security: Towards Proactive Threat Hunting via Adversary Emulation // IEEE Access. 2021. Vol. 9. P. 126023–126033. DOI: 10.1109/ACCESS.2021.3111108 .

5. Chen L., Jiang R., Lin C., Li A. A Survey on Threat Hunting: Approaches and Applications // Proceedings of the 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC). 2022. P. 340–344. DOI: 10.1109/DSC55712.2022.00056 .
6. Kulkarni M.S., Hegadi A.D., Narayan C.C. A Proactive Approach to Advanced Cyber Threat Hunting // Proceedings of the 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). 2023. P. 1–6.
7. Ali T. & Kostakos P. (2023). HuntGPT: Integrating Machine Learning-Based Anomaly Detection and Explainable AI with Large Language Models (LLMs). 10.48550/arXiv.2309.16021.
8. Würsch M. et al. LLM-based entity extraction is not for cybersecurity // Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and AI+ Informetrics (AII2023). 2023. T. 3451. P. 26–32.
9. Bolla A., Talentino F. Threat Hunting driven by Cyber Threat Intelligence: Thesis. Politecnico di Torino, 2022.
10. Mavroeidis, Vasileios & Jøsang, Audun. (2018). Data-Driven Threat Hunting Using Sysmon. 10.1145/3199478.3199490.
11. Mughal A.A. Building and Securing the Modern Security Operations Center (SOC) // International Journal of Business Intelligence and Big Data Analytics. 2022. Vol. 5. No. 1. P. 1–15.
12. Kotenko I.V., Polubelova O.V., Saenko I.B., Chechulin A.A. Primenenie ontologii i logicheskogo vyvoda dlia upravleniia informatsiei i sobytiiami bezopasnosti // Sistemy vysokoi dostupnosti. 2012. T. 8. No. 2. Pp. 100–108.
13. Rahali A., Akhloufi M.A. MalBERT: Malware Detection Using Bidirectional Encoder Representations from Transformers // Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2021. P. 3226–3231. DOI: 10.1109/SMC52423.2021.9658655 .
14. 14. Rahali A., Akhloufi M.A. MalBERTv2: Code Aware BERT-Based Model for Malware Identification // Big Data and Cognitive Computing. 2023. Vol. 7. No. 2. P. 60. DOI: 10.3390/bdcc7020060 .
15. 15. Seyyar Y.E., Yavuz A.G., Ünver H.M. An Attack Detection Framework Based on BERT and Deep Learning // IEEE Access. 2022. Vol. 10. P. 68633–68644. DOI: 10.1109/ACCESS.2022.3186827 .
16. Thapa C., Jang S.I., Ahmed M.E., Camtepe S., Pieprzyk J., Nepal S. Transformer-Based Language Models for Software Vulnerability Detection // Proceedings of the 38th Annual Computer Security Applications Conference (ACSAC). 2022. P. 481–496. DOI: 10.1145/3564625.3564627 .
17. Yu K., Tan L., Mumtaz S., Al-Rubaye S., Al-Dulaimi A., Bashir A.K., Khan F.A. Securing Critical Infrastructures: Deep-Learning-Based Threat Detection in IIoT // IEEE Communications Magazine. 2021. Vol. 59. No. 10. P. 76–82. DOI: 10.1109/MCOM.001.2001225 .
18. Wang Z., Li J., Yang S., Luo X., Li D., Mahmoodi S. A Lightweight IoT Intrusion Detection Model Based on Improved BERT-of-Theseus // Expert Systems with Applications. 2024. Vol. 238. P. 122045. DOI: 10.1016/j.eswa.2023.122045 .
19. Motlagh F.N., Alqahtani S.A., Shafie-Khah M., Catalão J.P.S., Hossain M.S. Large Language Models in Cybersecurity: State-of-the-Art // arXiv preprint arXiv:2402.00891. 2024.
20. Ferrag M.A., Shu L., Maglaras L., Ahmim A., Janicke H., Wang X. Revolutionizing Cyber Threat Detection with Large Language Models: A Privacy-Preserving BERT-Based Lightweight Model for IoT/IIoT Devices // IEEE Access. 2024.
21. Hu F., Hu H., Li W., Wang W., Zhu Q., Yao Y.D. Towards Federated Learning Models Resistant to Adversarial Attacks // IEEE Internet of Things Journal. 2023. DOI: 10.1109/JIOT.2023.3252194 .
22. Chen S., Liao H. BERT-log: Anomaly Detection for System Logs Based on Pre-Trained Language Model // Applied Artificial Intelligence. 2022. Vol. 36. No. 1. P. e21456422028014. DOI: 10.1080/08839514.2022.2028014 .
23. Liu Y., Wu Y., Song W., Chen Z., Li Z. LogPrompt: Prompt Engineering Towards Zero-Shot and Interpretable Log Analysis // Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings (ICSE). 2024. P. 364–365. DOI: 10.1145/3639478.3643108 .
24. Fang R., Wu Y., Li Z., Su Z., Gu Q. LLM Agents Can Autonomously Exploit One-Day Vulnerabilities // arXiv preprint arXiv:2404.08144. 2024.
25. Deng G., Wang T., Liu Z., Liu X., Ding X., Lin X. PentestGPT: An LLM-Empowered Automatic Penetration Testing Tool // arXiv preprint arXiv:2308.06782. 2023.
26. Aghaei E., Chaffar S., Othmane L.B., Namilae S., Zaman N. SecureBERT: A Domain-Specific Language Model for Cybersecurity // International Conference on Security and Privacy in Communication Systems (SecureComm). Cham: Springer Nature Switzerland, 2022. P. 250–266. DOI: 10.1007/978-3-031-13708-1_13 .
27. Alawida M., Alashqar B.M., Jantan A., Sharieh A., Al-Kadi O.S. A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity // Information. 2023. Vol. 14. No. 8. P. 462. DOI: 10.3390/info14080462 .
28. Charan P.V., Arpaci I., Al-Emran M., Tavares J.M.R.S., Sucipto T.D., Kouttis M., Aldawood H., Martinet F., Al-Qirim N. From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads // arXiv preprint arXiv:2305.15336. 2023.
29. Gat Y., Bachem O., Chen S., D'Amour A., Lucic M. Faithful Explanations of Black-Box NLP Models Using LLM-Generated Counterfactuals // arXiv preprint arXiv:2310.00603. 2023.
30. Jiang A.Q., Li Z., Ye D., Niu S., Xie F., Wang L., Fang F. Mixture of Experts // arXiv preprint arXiv:2401.04088. 2024.
31. Wu Q., Yin H., Dong D., He Z., Du D., Wang Z., Yin D., Tan B. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework // arXiv preprint arXiv:2308.08155. 2023.