НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2025 Том 25 № 1 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2025 Vol. 25 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

КОМПЬЮТЕРНЫЕ СИСТЕМЫ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

COMPUTER SCIENCE

doi: 10.17586/2226-1494-2025-25-1-42-52
УДК 004.056

Большие языковые модели в информационной безопасности и тестировании на проникновение: систематический обзор возможностей применения

Антон Александрович Конев[1]✉, Татьяна Игоревна Паюсова[2]

[1] Томский государственный университет систем управления и радиоэлектроники, Томск, 634050, Российская Федерация
[2] Тюменский государственный университет, Тюмень, 625003, Российская Федерация

[1] kaa@fb.tusur.ru✉, https://orcid.org/0000-0002-3222-9956
[2] t.i.payusova@utmn.ru, https://orcid.org/0000-0003-4923-1689

Аннотация

Введение. Развитие технологий искусственного интеллекта, в частности, больших языковых моделей (Large Learning Model, LLM), привело к изменениям во многих сферах жизни и деятельности человека. Информационная безопасность также претерпела существенные изменения. Тестирование на проникновение (пентест) позволяет оценить систему защиты на практике в «боевых» условиях. LLM могут вывести практический анализ защищенности на качественно новый уровень с точки зрения автоматизации и возможности генерации нестандартных шаблонов атаки. Представленный в работе систематический обзор направлен на определение уже известных способов применения LLM в кибербезопасности, а также на выявление «белых пятен» в развитии технологии. **Метод.** Отбор исследуемых научных работ осуществлялся в соответствии с многоступенчатым руководством PRISMA на основании анализа аннотаций и ключевых слов публикаций. Полученная выборка была дополнена с помощью метода «снежного кома» и ручного поиска статей. Суммарное количество публикаций составило 50 работ с января 2023 г. по март 2024 г. **Основные результаты.** В работе выполнен анализ способов применения LLM в области информационной безопасности (поддержка целеполагания и принятия решений, автоматизация пентеста, анализ защищенности моделей LLM и программного кода). Определены архитектуры LLM (GPT-4, GPT-3.5, Bard, LLaMA, LLaMA 2, BERT, Mixtral 8×7B Instruct, FLAN, Bloom) и программные решения на базе LLM (GAIL-PT, AutoAttacker, NetSecGame, Cyber Sentinel, Microsoft Counterfit, GARD project, GPTFUZZER, VuRLE), применяемые в области информационной безопасности. Установлены ограничения (конечное «время жизни» данных для обучения LLM, недостаточные когнитивные способности языковых моделей, отсутствие самостоятельного целеполагания и сложности при адаптации LLM к новым параметрам задачи). Выявлены потенциальные точки роста и развития технологии в контексте киберзащиты (исключение «галлюцинаций» моделей и обеспечение защиты LLM от джейлбрейков, осуществление интеграции известных разрозненных решений и программная автоматизация выполнения задач в области информационной безопасности с помощью LLM). **Обсуждение.** Полученные результаты могут быть полезны при разработке собственных теоретических и практических решений, обучающих и тренировочных наборов данных, программных комплексов и инструментов для проведения тестирования на проникновение. Исследование поможет в реализации новых подходов к построению LLM и повышению их когнитивных способностей, учитывающих аспекты работы с джейлбрейками и «галлюцинациями», а также для самостоятельного дальнейшего многостороннего изучения вопроса.

Ключевые слова

обработка естественного языка, компьютерная лингвистика, ChatGPT, пентест, искусственный интеллект, машинное обучение, наступательная безопасность, моделирование атак, автоматизация, Red Teaming, джейлбрейк

Благодарности

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках базовой части государственного задания ТУСУРа на 2023–2025 гг. (проект № FEWM-2023-0015).

Ссылка для цитирования: Конев А.А., Паюсова Т.И. Большие языковые модели в информационной безопасности и тестировании на проникновение: систематический обзор возможностей применения // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 1. С. 42–52. doi: 10.17586/2226-1494-2025-25-1-42-52

Large language models in information security and penetration testing: a systematic review of application possibilities

Anton A. Konev[1]✉, Tatyana I. Payusova[2]

[1] Tomsk State University of Control Systems and Radioelectronics (TUSUR), Tomsk, 634050, Russian Federation
[2] Tyumen State University, Tyumen, 625003, Russian Federation

[1] kaa@fb.tusur.ru✉, https://orcid.org/0000-0002-3222-9956
[2] t.i.payusova@utmn.ru, https://orcid.org/0000-0003-4923-1689

Abstract

The development of artificial intelligence technologies, in particular, large language models (LLM), has led to changes in many areas of human life and activity. Information security (IS) has also undergone significant changes. Penetration testing (pentest) allows evaluating the security system in practice in "combat" conditions. LLMs can take practical security analysis to a qualitatively new level in terms of automation and the ability to generate non-standard attack patterns. The presented systematic review is aimed at determining the already known ways of applying LLM in cybersecurity, as well as identifying "blank spots" in the development of technology. The selection of literature sources was carried out in accordance with the multi-stage PRISMA guidelines based on the analysis of abstracts and keywords of publications. The resulting sample was supplemented using the "snowball" method and manual search of articles. The total number of publications was 50 works from January 2023 to March 2024. The conducted research allowed to analyze the ways of using LLM in the field of information security (goal setting and decision-making support, pentest automation, security analysis of LLM models and program code), determine the LLM architectures (GPT-4, GPT-3.5, Bard, LLaMA, LLaMA 2, BERT, Mixtral 8×7B Instruct, FLAN, Bloom) and software solutions based on LLM used in the field of information security (GAIL-PT, AutoAttacker, NetSecGame, Cyber Sentinel, Microsoft Counterfit, GARD project, GPTFUZZER, VuRLE), to establish limitations (finite "lifetime" of data for LLM training, insufficient cognitive abilities of language models, lack of independent goal setting and difficulties in adapting LLM to new task parameters), identify potential growth points and development of technology in the context of cyber defense (elimination of "hallucinations" of models and ensuring protection of LLM from jailbreaks, implementation of integration of known disparate solutions and software automation of tasks in the field of information security using LLM). The presented results can be useful in developing theoretical and practical solutions, educational and training datasets, software packages and tools for penetration testing, new approaches to building LLM and improving their cognitive abilities, taking into account aspects of working with jailbreaks and "hallucinations", as well as for independent further multilateral study of the issue.

Keywords

natural language processing, computational linguistics, ChatGPT, artificial intelligence, machine learning, attack modeling, Red Teaming, jailbreaking

Acknowledgements

This research was funded by the Ministry of Science and Higher Education of the Russia, Government Order for 2023–2025, project no. FEWM-2023-0015 (TUSUR).

For citation: Konev A.A., Payusova T.I. Large language models in information security and penetration testing: a systematic review of application possibilities. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 1, pp. 42–52 (in Russian). doi: 10.17586/2226-1494-2025-25-1-42-52

Введение

30 ноября 2022 года компания OpenAI показала чат-бот ChatGPT на основе большой языковой модели (Large Learning Model, LLM) GPT-3.5. Данная демонстрация ознаменовала начало нового витка развития и распространения искусственного интеллекта. Только за первый квартал 2024 года на портале «Google Scholar» опубликованы 12 000 научных работ по теме LLM.

С открытием новых возможностей в области генерации текстового контента, поддержки принятия решений и поисковой оптимизации в свою очередь изменилась область информационной безопасности (ИБ). Первые активные действия по применению LLM осуществили злоумышленники для написания вредоносного кода, подготовки фишинговых писем, реализации «разведки по открытым источникам». Система защиты данных только начинает осваивать новые перспективы, устанавливает границы и практические способы применения больших языковых моделей.

Тестирование на проникновение (пентест) объединяет приемы «белых и черных шляп», позволяет моделировать атаки любой сложности и проводить многосторонний анализ защиты, с которым не сравнится «бумажный» подход, основанный на подготовке документов, подтверждающих в большей степени формальное соответствие системы защиты требованиям регуляторов. Усложнение ландшафта угроз, в том числе обусловленное развитием искусственного интеллекта и LLM, привело к ослаблению известных способов обеспечения и проверки безопасности. Рост успешных кибератак по данным Следственного комитета РФ[1], востребованность услуг по этичному взлому на сайтах вакансий и поиску работы[2], говорят о важности

[1] В СК сообщили о росте количества IT-преступлений в России с 2014 года более чем в 50 раз // Информационное агентство ТАСС. 2023. 22 мая. URL: https://tass.ru/proisshestviya/17812613 (дата обращения: 01.10.2024).

[2] Спрос на ИБ-кадры по-прежнему растет // Издательство ComNews. 2023. 1 дек. URL: https://www.comnews.ru/content/230484/2023-12-01/2023-w48/1008/spros-ib-kadry-prezhnemu-rastet?utm_source=rfinance (дата обращения: 01.10.2024).

изучения аспектов тестирования на проникновение и совершенствования известных подходов с помощью новых инструментов.

В работе [1] отмечено отсутствие качественных всеобъемлющих методологий, способных системно помочь при моделировании и изучении информационных атак. В [2] показана важность комплексного подхода к обеспечению ИБ, в том числе с использованием инновационных методов моделирования. Концептуализация и исследование атак в рамках тестирования на проникновение на основе LLM может объединить в одном контексте факторы и элементы разнообразной природы, описать на естественном языке сценарии любой сложности и структуры. Таким образом, актуальной представляется задача определения возможностей больших языковых моделей в ИБ, в том числе для проведения и автоматизации пентеста, а также выявления пробелов при применении LLM в киберзащите.

В настоящей работе представлен систематический обзор, с помощью которого предпринята попытка ответить на ряд вопросов о роли и применении LLM и методов компьютерной лингвистики для решения задач ИБ.

Вопрос 1. Какие существуют задачи в области ИБ, в которых является актуальным использование LLM и методов компьютерной лингвистики?

Вопрос 2. Выделяют ли авторы работ, представленных в исследовании, определенные архитектуры LLM и готовые программные продукты на базе LLM для решения задач ИБ и автоматизации тестирования на проникновение?

Вопрос 3. Существуют ли границы применения LLM, которые можно выделить при рассмотрении больших языковых моделей в контексте ИБ?

Вопрос 4. Возможно ли выделить аспекты применения LLM и методов компьютерной лингвистики для решения задач ИБ и автоматизации тестирования на проникновение, обладающие «белыми пятнами», восполнить которые могут дальнейшие более детальные исследования?

Полученные ответы позволят определить потенциальные сильные и слабые стороны в применении LLM для защиты информации, частично или полностью автоматизировать тестирование на проникновение, обозначить программные аспекты интеграции языковых моделей с элементами системы защиты, усовершенствовать подходы наступательной и оборонительной безопасности.

Материалы и методы

Выполнен отбор научных работ в соответствии с руководством PRISMA в поисковой системе https://scholar.google.com/[1].

Первоначально были отобраны публикации за последние пять лет (2019–2024 гг.) по запросу «*(“information security” OR pentest) (NLP OR LLM)*». Количество найденных источников составило $n = 16\,500$.

Первичный обзор аннотаций и ключевых слов найденных источников показал, что многие работы косвенно связаны с обработкой естественного языка, применением языковых моделей и методов компьютерной лингвистики в ИБ и при проведении тестирований на проникновение. Запрос был конкретизирован с точки зрения использования LLM: «*(“information security” OR pentest) LLM*» ($n = 4940$).

В силу относительной новизны темы и последнего проявленного интереса к большим языковым моделям, связанным с презентацией OpenAI ChatGPT 30 ноября 2022 г., поиск был сужен до наиболее актуальных работ, датированных 2023–2024 гг. Количество источников составило $n = 1490$.

На основе анализа аннотаций и ключевых слов и с учетом исключенных статей (дубликатов, источников в нерецензируемых журналах, работ, не имеющих полного текста) количество работ в выборке составило $n = 38$. Для расширения анализируемой выборки методом «снежного кома» были найдены и добавлены 8 источников, самая ранняя из найденных публикация датируется 2015 г. ($n = 46$). Путем ручного поиска в системе https://elibrary.ru/[2] были найдены и добавлены в выборку еще четыре публикации ($n = 50$).

Аналогичные обзоры

Аспекты применения LLM в области киберзащиты, в том числе при проведении тестирований на проникновение, нашли отражение в нескольких систематических обзорах. В частности, можно отметить работы [3–6].

Настоящая работа, в отличие от перечисленных исследований, выполнена в соответствии с методикой PRISMA и рассматривает публикации, напрямую связанные с применением LLM в области киберзащиты, в большинстве своем опубликованные в рецензируемых источниках за последний год (с января 2023 г. по март 2024 г.).

В работе [3] не были рассмотрены вопросы наступательной безопасности, сценарного моделирования при пентесте, связь LLM с текстовым контентом существующих баз знаний. В [4] практически не представлено описание известных наборов данных и программных решений на основе больших языковых моделей для решения задач кибербезопасности. В работе [5] поверхностно отмечена возможность применения LLM в контексте ИБ, в основном сосредотачиваясь на общих задачах. В [6] изучены только модели семейства GPT и не охвачено всестороннее применение LLM в ИБ.

Настоящее исследование существенно дополняет аналогичные обзоры. В работе приводятся направления дальнейших исследований LLM, связанные как с концептуальными задачами (повышение когнитивных способностей, осуществление самостоятельного целеполагания), так и с техническими вопросами (исключение «галлюцинаций» или их перенаправление в конструктивное русло, например, для генерации более сложных сценариев пентеста, повышение адаптивных

[1] Поисковая система Google Scholar. URL: https://scholar.google.com/ (дата обращения: 12.01.2025).

[2] Научная электронная библиотека Elibrary.ru. URL: https://www.elibrary.ru/ (дата обращения: 12.01.2025).

способностей с помощью автоматической контекстуализации запросов). Обозначены границы применения LLM и методов компьютерной лингвистики в контексте ИБ. Представлена подборка готовых программных решений на базе больших языковых моделей, а также наборов данных, подготовленных специально для решения задач ИБ.

Решение задач ИБ с помощью LLM

Обобщенные результаты систематического обзора представлены в табл. 1. Результаты отражают некоторые задачи в области поддержки целеполагания и принятия решений, автоматизации пентеста, анализа защищенности моделей LLM (AI Red Teaming) и программного кода, а также возможные способы их решения и дают ответ на вопрос 1. Настоящая работа выделяет пентест, как одну из наименее изученных тем с точки зрения автоматизации на базе больших языковых моделей, и при этом как наиболее перспективное направление в области наступательной безопасности с учетом стремительно меняющегося цифрового ландшафта.

Результаты проведенного исследования позволяют говорить о востребованности LLM при решении задач ИБ, о чем, в том числе, свидетельствуют созданные таксономии и «дорожные карты» применения больших языковых моделей в кибербезопасности [5, 6]. Можно отметить появление новых методов для повышения адаптационных способностей моделей [10], в частности, методе генерации с расширенным извлечением (Retrieval-Augmented Generation, RAG) [17], предполагающим автоматизированное включение в запрос ключевых слов/фраз/терминов/шаблонов, характери-

Таблица 1. Обзор направлений исследований
Table 1. Review of the research directions

Раздел и ссылки	Рассматриваемые задачи	Предложенные решения
Направление «Поддержка целеполагания и принятия решений»		
Общие подходы [7–9]	Стратегическое планирование	Повышение абстрактности вопросов, консолидация данных для достижения синергетического эффекта
	Декомпозиция целей, тактическое планирование	Повышение качества обучающих наборов данных, снижение двусмысленности формулировок
	Самостоятельное рассуждение LLM	Применение графового подхода, деревьев рассуждений для выстраивания логических связей
Пентест [10, 11]	Поиск оптимальной последовательности действий	Применение наиболее мощных архитектур LLM, например, LLaMA, BERT, FLAN, Bloom, Mixtral 8 × 7B Instruct
	Принятие решений	Повышение контекстуализации запросов
Сертификация и соревнования [12–15]	Принятие решений, осуществление рассуждений	Повышение когнитивных функций LLM, повышение качества обучающих наборов данных
Направление «Автоматизация пентеста»		
Сценарии атаки [16, 17]	Создание нестандартных сценариев	Реализация «мозгового штурма», участниками которого выступают LLM. Использование в сценариях полученных «галлюцинаций»
Уязвимости системы [18–23]	Выдвижение гипотез о наличии уязвимостей в системе	Сопоставление результатов сканирования и данных о системе с базами уязвимостей. Автоматизация генерации команд для эксплуатации уязвимостей
Социотехнические аспекты [24–26]	Проведение социотехнических атак, поддержка диалога, усложнение рассуждений	Проведение теста Тьюринга для проверки качества и правдоподобности модели, применение деревьев рассуждений для выстраивания более точных и логичных причинно-следственных действий
Интеграция с известными базами знаний [27–30]	Контекстуализация сценариев атаки с помощью известного текстового контента	Обращение к MITRE ATT&CK, базам CAPEC, CVE, CWE и NVD. Применение методов компьютерной лингвистики (TF-IDF, word2vec, bag-of-words) для обработки текстовых материалов
Пентест передовых технологий [31, 32]	Анализ защищенности квантовых технологий и смарт-контрактов на базе блокчейна	«Знакомство» LLM с описанием квантовых протоколов (BB84 и квантово-устойчивых криптографических алгоритмов от NIST). Добавление в наборы данных примеров смарт-контрактов
Программные решения для автоматизации пентеста [33–36]	Объединение разрозненных решений в единый продукт	Применение API, прокси, сериализации данных и других интеграционных решений для объединения модулей
	Автоматизация выполнения команд	Выполнение и обработка системных вызовов, осуществление межпроцессного взаимодействия

Таблица 1. Продолжение

Раздел и ссылки	Рассматриваемые задачи	Предложенные решения
Направление «Анализ защищенности моделей LLM»		
Защита LLM от базовых угроз [37–39]	Защита от Prompt Injection, Indirect Prompt Injection, «отравления» данных, кражи модели	Автоматическое написание подсказок, генерация безопасных ответов, например, с помощью Microsoft Counterfit и GARD project
Защита данных [40]	Защита от утечек данных через LLM	Зашумление данных, распределенное хранение, применение системы меток
Морально-этические аспекты [41, 42]	Исключение некорректных ответов на «чувствительные» темы	Исследование параметров ложного и «токсичного» контентов
	Исключение бессмысленных ответов и искажения фактов	Изменение алгоритмов сжатия, поддержка стабильности диалога
Джейлбрейки и промпт-инжиниринг [19, 34, 43–46]	Анализ угроз и перспектив применения джейлбрейков	Исследование параметров и возможностей джейлбрейков
	Автоматизация генерации шаблонов джейлбрейков	Изучение методик промпт-инжиниринга и структуры промптов
Направление «Анализ защищенности программного кода»		
Поиск уязвимостей [47–50]	Автоматизация генерации защищенного кода, поиск уязвимостей в программах	Повышение качества обучения модели, усложнение обучающей выборки, применение эталонных наборов данных
Форензика, поиск индикаторов компрометации [51, 52]	Анализ лог-файлов, классификация инцидентов и сигнатур вредоносных программ	Исследование кода/программы с помощью разных языковых моделей и сравнение результатов с точки зрения совпадающих индикаторов компрометации

зующих предметную область и уточняющих контекст. В работах [7–15] изучены когнитивные способности LLM [7–15], приведены способы усовершенствования процесса рассуждения с помощью генеалогического древа и общего графа [9].

В настоящей работе выполнен анализ работ, описывающих интеграцию LLM с базами CAPEC, CVE, NVD, MITRE ATT&CK, положениями NIST Cybersecurity Framework. В работах [27–30] представлены результаты применения методов компьютерной лингвистики (TF-IDF, Doc2Vec) для установления семантического сходства и сопоставления баз CVE, CAPEC, CWE. Инновационное направление AI Red Teaming работает в направлении повышения защиты LLM от подделки и кражи данных, генерации безопасных ответов, реализации методов автоматического анализа защищенности языковых моделей [37–46]. LLM находят применение для моделирования кибератак при тестировании защищенности квантовых технологий (протокола BB84 и квантово-устойчивых криптографических алгоритмов от NIST) [31], смарт-контрактов [32].

Архитектуры и готовые программные продукты на базе LLM

Для ответа на вопрос 2 работы, можно отметить, что применение LLM носит как высокоабстрактный концептуальный уровень применения, так и исключительно прикладной характер. Из примеров языковых моделей чаще всего упоминаются GPT-4 [11–14, 19–23, 28, 33, 41, 43, 46–48, 50, 52] и GPT-3.5 [23, 50, 52] от OpenAI, Google Bard [14, 41], LLaMA (LLaMA 2) [10, 21], BERT [15], Mixtral 8×7B Instruct (в работе [10] показано применение модели в рамках чат-бота Huggingchat), FLAN [10], Bloom [10]. Наиболее мощной архитектурой по результатам исследования представляется GPT-4, так как большинство разнообразных задач были решены именно на ее основе. LLaMA 2 и BERT отражены в работах как наиболее гибкие решения в силу их меньшей цензурированности и вследствие этого большей адаптивности и способности «встраиваться» в любой контекст. В работе [50] рассмотрены более редкие и экзотические модели, например, PLART, CodeT5, CodeGen, In-Coder, Codex-12B, GPT-Neo, ChatGPT-175B, но их применение носит скорее исследовательский интерес, чем практическую пользу, например, по сравнению с GPT-4.

В табл. 2 представлены примеры типов данных, используемые для обучения LLM. Типы данных систематизированы по частоте их упоминания в перечисленных работах.

Дополнительно, в качестве обучающих материалов упоминаются примеры текстовых описаний сетевых топологий и информационных инфраструктур [10, 35], ложный и «токсичный» контент [41], ранее собранные и накопленные результаты работы LLM [23], наборы данных Defects4J v1.2, Defects4J v2.0, QuixBugs, HumanEval-Java, Many-Bugs, образцы кода с площадки Stack Overflow [50].

В изученных работах представлен ряд готовых программных комплексов на основе LLM и наборов данных для обучения и тестирования качества моделей: программный комплекс GAIL-PT [19], система AutoAttacker [34] и NetSecGame [35] для автоматического взлома, диалоговая система Cyber Sentinel [36], Microsoft Counterfit и GARD project для проведения AI Red Teaming [38], среда GPTFUZZER [46] для автоматизации генерации шаблонов джейлбрейков, система

Таблица 2. Обзор типов данных и факты о них
Table 2. Overview of data types and facts about them

Данные для обучения	Примеры и факты	Ссылки
Описания атак и пентестерские отчеты	GPT-4 продемонстрировал на 228,6 % большее количество успешных шагов, чем GPT-3.5	[7, 10, 11, 16–19, 23, 34, 35]
Данные MITRE ATT&CK, CVE, CWE, NVD, NIST Cybersecurity Framework, баз IoC и др.	MITRE ATT&CK содержит более 200 техник, «разложенных» по 14 тактикам. База NVD содержит более 100 000 уязвимостей	[7, 11, 27–30, 36, 51, 52]
Код вредоносных программ и образцы полезной нагрузки	Алгоритм word2vec может быть применен для исследования признаков веб-шелла	[20, 22, 25, 26, 28, 33, 51, 52]
Примеры джейлбрейков	Процент взломов с помощью джейлбрейков может быть снижен при использовании LLM с 67,21 % до 19,34 %	[19, 34, 43–46]
Примеры фишинговых писем, спама	С помощью LLM возможно определение «подозрительных» писем и их параметров	[16, 20, 21, 25]
Фрагменты уязвимого программного кода	Более чем для 60 % примеров GPT-4 удалось правильно определить проблему с исходным кодом. GPT-4 продемонстрировал 73 % успешных взломов веб-сайтов. Точность обнаружения логических уязвимостей в смарт-контрактах – более 70 %	[32, 47, 48, 50]
Задания CTF-соревнований	GPT-4 превзошел 88,5 % игроков CTF	[12, 13, 49]
Примеры прохождений (writeup) уязвимых машин	Обращение авторов в исследованиях к машинам Chaos, SteamCloud, GoodGames на площадке HackTheBox	[18, 19]
Вопросы к сертификации Certified Ethical Hacking (CEH)	ChatGPT продемонстрировал уровень точности ответов 80,8 %, Bard – 82,6 %	[14]
Извлечение информации из нормативных правовых актов и регламентов	Для анализа GDPR применялись 107 вопросов и модель BERT. Точность ответов составила 91 %	[15]
Словари для брутфорса	Словарь RockYou содержит 14 млн паролей	[20]
Переписка из социальный сетей и других открытых источников	Для обучения LLM использовалось порядка 300 млрд слов в рамках социальной сети Reddit	[20]
Конфиденциальная информация	При тестировании моделей GPT-2 и GPT-3 имел место факт утечки финансовой и медицинской информации	[40]
Описания протоколов, спецификаций	Обучение LLM осуществлялось на основе протокола BB84 и криптографических алгоритмов NIST	[31]

для автоматического поиска и устранения уязвимостей VuRLE [49]. Представленные решения ответственны за различные грани применения LLM в области ИБ: пентест, джейлбрейки, поиск и устранение уязвимостей, поддержку принятий решений в области управления рисками и др. Каждый продукт показал свою эффективность при проведении экспериментов в рамках исследований: GAIL-PT позволил снизить временные издержки при работе LLM на 55 % по сравнению с популярным продуктом DeepExploit; AutoAttacker может обходить градиентную обфускацию при проведении атаки на LLM, что не представлено ни в одном другом решении; NetSecGame активно поддерживается и развивается разработчиками, в частности планируется создание более сложных сценариев атаки с помощью реализации мультиагентного подхода, в рамках которого каждый элемент отвечает за свой фрагмент сценария, и при их объединении получается кумулятивный эффект сценарных действий; фреймворк VuRLE позволил обнаружить 183 из 279 уязвимостей и исправить 101 из них, в отличие от продукта LASE, который выявил 58 «лазеек» и закрыл только 21 на основе анализа текстовой документации.

Ограничения и барьеры при применении LLM

Из ограничений и трудностей (табл. 3) при применении LLM (вопрос 3) можно выделить конечное «время жизни» данных для обучения, недостаточные когнитивные способности больших языковых моделей, отсутствие самостоятельного целеполагания и сложности при адаптации к новым условиям задачи [7–19, 21, 22, 24–30, 33, 34, 40, 41, 43, 47, 49, 50].

Актуализация набора данных для обучения LLM. Современные решения оперируют уже обученными на готовых наборах данных большими языковыми моделями. Сформированные обучающие выборки не предполагают обновление информации о ландшафте угроз, новых векторах атак и уязвимостях нулевого дня. Непрерывная актуализация наборов данных предполагает открытую архитектуру. Открытость, в свою очередь, связана с рисками безопасности: с возможными атаками на параметры модели, «отравлением» или умышленной порчей данных, аспектами обработки конфиденциальной информации при обучении языковых моделей [7, 8, 11, 27–30, 33, 34, 40, 41, 47, 49, 50].

Когнитивные способности языковых моделей. Несмотря на возможности, предоставляемые LLM,

Таблица 3. Возможные методы преодоления ограничений LLM
Table 3. Possible methods to overcome LLM limitations

Ограничения	Методы решения	Ссылки
Конечное «время жизни» данных для обучения LLM	Программная автоматизация процесса дообучения модели на актуальных наборах данных. Возможная автоматизация написания запросов к LLM с помощью другой специально обученной языковой модели. Программная реализация разведки и анализа поступающих данных. Достижение синергетического эффекта при объединении гетерогенных выборок	[7, 8, 11, 27–30, 33, 34, 40, 41, 47, 49, 50]
Недостаточные когнитивные способности языковых моделей	Применение обучения с подкреплением, повышение абстрактности и масштабности вопросов, задействование механизма рассуждений на основании теории графов, деревьев рассуждений для повышения мыслительной функции LLM, увеличение продолжительности тестирования моделей, организация нескольких LLM в единый кластер для усиления их способностей в формате «мозгового штурма»	[7–9, 12–17]
Отсутствие самостоятельного целеполагания	Повышение качества промптов с помощью уточнения формулировок, снижения двусмысленности, подведение LLM к написанию диалогов и вывода процесса рассуждений с помощью механизма джейлбрейкинга	[10, 11, 16, 21, 22]
Сложности при адаптации	Применение RAG, контекстуализация запросов и обеспечение адаптационного «люфта» с помощью джейлбрейков и дополнительного включения в обучающую выборку примеров описаний реальных систем	[10, 17–19, 24–26, 34, 43]

пока языковые модели не могут заменить мыслительный процесс, осуществляемый этичным взломщиком. Тест Тьюринга мог бы выступить инструментом для проверки качества работы LLM и разделения процесса имитации от настоящей когнитивной деятельности. Современные языковые модели опираются на готовые решения при выполнении атаки, в них отсутствует творческая составляющая и достаточная мыслительная способность [7–9, 12–17].

Отсутствие самостоятельного целеполагания у LLM. Взаимодействие с LLM выстраивается через пользователя и его открытые вопросы. Весь дальнейший процесс «общения» проходит в рамках запросов пользователей. Движение «за флажки» возможно, но на практике трудновыполнимо: LLM старается найти ответ именно на вопрос пользователя, при этом верное решение может находиться в другой плоскости. Представленные исследования отмечают, что выбор и формулирование открытых вопросов к LLM, требуют отдельной проработки и изучения [10, 11, 16, 21, 22].

Адаптация моделей. LLM могут эффективно работать только с системами, описанными в обучающей выборке. Изменение параметров информационной инфраструктуры требует дообучения или переучивания моделей. Повышение гибкости, изменчивости и адаптационных свойств больших языковых моделей, например, с помощью метода RAG, представляется актуальной задачей [10, 17–19, 24–26, 34, 43].

Возможные направления дальнейших исследований

Выполненное исследование научных работ [11, 17, 19–30, 33–36, 42–46] позволяет развернуто ответить на вопрос 4 и выделить ряд направлений, требующих дополнительного изучения и более глубокого раскрытия темы: «галлюцинации» моделей, джейлбрейки, интеграция разрозненных решений и программная автоматизация необходимых действий. В табл. 4 представлены возможные задачи, связанные с направлениями дальнейших исследований.

«Галлюцинации» моделей. Многие исследования указывают на проблему ложных или бессмысленных ответов от языковых моделей из-за ограниченной или противоречащей информации в обучающем наборе данных или чрезмерного сжатия данных для обучения. На определенном этапе при продолжительной серии однотипных вопросов к LLM, также может произойти «зацикливание» и компиляция ранее предоставленных ответов, в том числе с искажениями и общими формулировками [17, 19, 35, 42].

Джейлбрейки. С одной стороны, джейлбрейки позволяют обходить цензуру, нарушая нормы морали и этики, с другой — предлагать нестандартное видение вопроса, осуществлять проактивный поиск новых векторов атак и траекторий злоумышленника. Кроме этого, джейлбрейк помогает решить проблему отсутствующего контекста, учитывать предметную область и специфику анализируемой информационной системы [19, 34, 43–46].

Интеграция разрозненных решений. Многие исследования рассматривают частные аспекты проведения пентеста. Проведенный обзор показывает, что на данный момент отсутствует системное решение, охватывающее все направления пентеста и возможности их комплексной реализации с помощью LLM и методов компьютерной лингвистики. Потенциальным решением представляется рассмотрение пентеста на базе LLM через призму модели Kill Chain и/или MITRE ATT&CK [11, 19, 27–30].

Программная автоматизация. В основе большинства представленных решений для автоматизации вы-

Таблица 4. Основные области дальнейшего изучения и возможные задачи
Table 4. Major areas for further study and possible solution methods

Направления	Задачи	Ссылки
«Галлюцинации» моделей	Повышение качества обучающих наборов данных, тестирование алгоритмов преобразования данных, используемых при обучении. Возможное рассмотрение «галлюцинаций» как инструмента, способного привнести нестандартные сценарные элементы в работу LLM	[17, 19, 35, 42]
Джейлбрейки	Защита от джейлбрейкинга с помощью зашумления данных, распределенного хранения без возможности одноэтапной компрометации, меток конфиденциальности. Повышение контекстуализации задачи и выстраивание процесса рассуждений	[19, 34, 43–46]
Интеграция разрозненных решений	Систематизация материалов по проведению тестирования на проникновение с помощью LLM и методов компьютерной лингвистики, например, с помощью модели Kill Chain, пошагово раскладывающей любую атаку на ряд составляющих базовых элементов, от разведки до сокрытия следов	[11, 19, 27–30]
Программная автоматизация	Выбор наиболее подходящей архитектуры LLM для решения поставленной задачи. Автоматизация запуска команд, сгенерированных LLM	[11, 17, 19–26, 33–36]

полнения команд лежит среда Metasploit и оболочка Meterpreter. Наиболее опасные злоумышленники редко применяют стандартные решения на практике, поэтому ценность пентеста, выполненного известными инструментами, может снижаться. Дальнейшим направлением исследования представляется поиск альтернатив для автоматизированного выполнения команд, сгенерированных LLM [11, 17, 19–26, 33–36].

Заключение

Область информационной безопасности во многом претерпевает серьезные изменения, связанные с вызовами окружающей среды, новыми угрозами и векторами атаки, а также появлением инновационных технологий, требующих смены подходов к защите. Настоящее исследование позволило определить известные способы применения больших языковых моделей (Large Learning Model, LLM) в области информационной безопасности, а также обозначить ограничения, потенциальные точки роста и развития технологии в контексте киберзащиты. Полученные результаты работы могут быть основой дальнейших теоретических и практических исследований в информационной безопасности и тестировании на проникновение. Представленное исследование позволит усовершенствовать существующие решения, в том числе, связанные с автоматизацией пентеста, собрать целостную методику анализа защищенности, например, соединив этапы Kill Chain и материалы MITRE ATT&CK с возможностями LLM и методами компьютерной лингвистики.

Развитие темы исследования может быть связано с более глубоким изучением каждого этапа тестирования на проникновение через призму больших языковых моделей, формированием понимания, какие шаги требуют улучшений и могут быть усовершенствованы с помощью возможностей LLM. Требуется изучение аспектов актуализации обучающих данных, повышение когнитивных и адаптационных способностей языковых моделей, развитие функции самостоятельного целеполагания. Необходимо устранение «галлюцинаций», исследование возможностей джейлбрейков, «бесшовная» интеграция разрозненных решений для получения целостной методики системного анализа защищенности информационной инфраструктуры. Возможен выбор наиболее подходящих инструментов программной реализации комплексных решений на основе LLM.

Литература

1. Konev A., Shelupanov A., Kataev M., Ageeva V., Nabieva A. A survey on threat-modeling techniques: protected objects and classification of threats // Symmetry. 2022. V. 14. N 3. P. 549. https://doi.org/10.3390/sym14030549
2. Shelupanov A., Evsyutin O., Konev A., Kostyuchenko E., Kruchinin D., Nikiforov D. Information security methods–Modern research directions // Symmetry. 2019. V. 11. N 2. P. 150. https://doi.org/10.3390/sym11020150
3. Yao Y., Duan J., Xu K., Cai Y., Sun Z., Zhang Y. A survey on large language model (LLM) security and privacy: The Good, the Bad, and the Ugly // High-Confidence Computing. 2024. V. 4. N 2. P. 100211. https://doi.org/10.1016/j.hcc.2024.100211
4. da Silva G.J.C., Westphall C.B. A Survey of Large Language Models in Cybersecurity // arXiv. 2024. arXiv:2402.16968. https://doi.org/10.48550/arXiv.2402.16968

References

1. Konev A., Shelupanov A., Kataev M., Ageeva V., Nabieva A. A survey on threat-modeling techniques: protected objects and classification of threats. *Symmetry*, 2022, vol. 14, no. 3, pp. 549. https://doi.org/10.3390/sym14030549
2. Shelupanov A., Evsyutin O., Konev A., Kostyuchenko E., Kruchinin D., Nikiforov D. Information security methods–Modern research directions. *Symmetry*, 2019, vol. 11, no. 2, pp. 150. https://doi.org/10.3390/sym11020150
3. Yao Y., Duan J., Xu K., Cai Y., Sun Z., Zhang Y. A survey on large language model (LLM) security and privacy: The Good, the Bad, and the Ugly. *High-Confidence Computing*, 2024, vol. 4, no. 2, pp. 100211. https://doi.org/10.1016/j.hcc.2024.100211
4. da Silva G.J.C., Westphall C.B. A Survey of Large Language Models in Cybersecurity. *arXiv*, 2024, arXiv:2402.16968. https://doi.org/10.48550/arXiv.2402.16968

5. Wang L., Ma C., Feng X., Zhang Z., Yang H., Zhang J., Chen Z., Tang J., Chen X., Lin Y., Zhao W., Wei Z., Wen J. A survey on large language model based autonomous agents // Frontiers of Computer Science. 2024. V. 18. N 6. P. 186345. https://doi.org/10.1007/s11704-024-40231-1
6. Gupta M., Akiri C., Aryal K., Parker E., Praharaj L. From ChatGPT to ThreatGPT: Impact of generative ai in cybersecurity and privacy // IEEE Access. 2023. V. 11. P. 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381
7. Happe A., Cito J. Getting pwn'd by ai: Penetration testing with large language models // Proc. of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2023. P. 2082–2086. https://doi.org/10.1145/3611643.3613083
8. Chang Y., Wang X., Wang J., Wu Y., Yang L., Zhu K., Chen H., Yi X., Wang C., Wang Y., Ye W., Zhang Y., Chang Y., Yu P., Yang Q., Xie X. A survey on evaluation of Large Language Models // ACM Transactions on Intelligent Systems and Technology. 2024. V. 15. N 3. P. 1–45. https://doi.org/10.1145/3641289
9. Li Z., Cao Y., Xu X., Jiang J., Liu X., Teo Y., Lin S., Liu Y. LLMs for Relational Reasoning: How Far are We? // Proc. of the 1st International Workshop on Large Language Models for Code (LLM4Code'24). 2024. P. 119–126. https://doi.org/10.1145/3643795.3648387
10. Genevey-Metat C., Bachelot D., Gourmelen T., Quemat A., Satre P.-M., Scotto L., Di Perrotolo, Chaux M., Delesques P., Gesny O. Red Team LLM: towards an adaptive and robust automation solution // Proc. of the Conference on Artificial Intelligence for Defense. 2023. hal-04328468.
11. Franco M.F., Rodrigues B., Scheid E., Jacobs A., Killer C., Granville L., Stiller B. SecBot: a business-driven conversational agent for cybersecurity planning and management // Proc. of the 16th International Conference on Network and Service management (CNSM). 2020. P. 1–7. https://doi.org/10.23919/cnsm50824.2020.9269037
12. Chamberlain D., Casey E. Capture the Flag with ChatGPT: Security Testing with AI ChatBots // Proc. of the 19th International Conference on Cyber Warfare and Security (ICCWS). 2024. V. 19. N 1. P. 43–54. https://doi.org/10.34190/iccws.19.1.2171
13. Timmins J., Knight S., Lachine B. Offensive cyber security trainer for platform management systems // Proc. of the IEEE International Systems Conference (SysCon). 2021. P. 1–8. https://doi.org/10.1109/syscon48628.2021.9447060
14. Raman R., Calyam P., Achuthan K. ChatGPT or Bard: Who is a better Certified Ethical Hacker? // Computers & Security. 2024. V. 140. P. 103804. https://doi.org/10.1016/j.cose.2024.103804
15. Abualhaija S., Arora C., Sleimi A., Briand L. Automated question answering for improved understanding of compliance requirements: A multi-document study // Proc. of the IEEE 30th International Requirements Engineering Conference (RE). 2022. P. 39–50. https://doi.org/10.1109/re54965.2022.00011
16. Renaud K., Warkentin M., Westerman G. From ChatGPT to HackGPT: Meeting the cybersecurity threat of generative AI // MIT Sloan Management Review. 2023.
17. Yamin M.M., Hashmi E., Ullah M., Katt B. Applications of LLMs for generating cyber security exercise scenarios // IEEE Access. 2024. V. 12. P. 143806–143822. https://doi.org/10.1109/access.2024.3468914
18. Heim M.P., Starckjohann N., Torgersen M. The Convergence of AI and Cybersecurity: An Examination of ChatGPT's Role in Penetration Testing and its Ethical and Legal Implications. BS thesis. NTNU, 2023.
19. Chen J., Hu S., Zheng H., Xing C., Zhang G. GAIL-PT: An intelligent penetration testing framework with generative adversarial imitation learning // Computers & Security. 2023. V. 126. P. 103055. https://doi.org/10.1016/j.cose.2022.103055
20. Ананьев В.А., Человечкова А.В. ChatGPT в сфере информационной безопасности // Информационная безопасность цифровой экономики. 2023. С. 77–83.
21. Прохоров А.И. Киберполигон как современный инструмент обеспечения информационной безопасности // Информатизация в цифровой экономике. 2023. Т. 4. № 4. С. 363–378. https://doi.org/10.18334/ide.4.4.119301
22. Chen Y., Yao Y., Wang X., Xu D., Yue C., Liu X., Chen K., Tang H., Liu B. Bookworm game: Automatic discovery of LTE vulnerabilities through documentation analysis // Proc. of the IEEE Symposium on Security and Privacy (SP). 2021. P. 1197–1214. https://doi.org/10.1109/sp40001.2021.00104

5. Wang L., Ma C., Feng X., Zhang Z., Yang H., Zhang J., Chen Z., Tang J., Chen X., Lin Y., Zhao W., Wei Z., Wen J. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 2024, vol. 18, no. 6, pp. 186345. https://doi.org/10.1007/s11704-024-40231-1
6. Gupta M., Akiri C., Aryal K., Parker E., Praharaj L. From ChatGPT to ThreatGPT: Impact of generative ai in cybersecurity and privacy. *IEEE Access*, 2023, vol. 11, pp. 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381
7. Happe A., Cito J. Getting pwn'd by ai: Penetration testing with large language models. *Proc. of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2023, pp. 2082–2086. https://doi.org/10.1145/3611643.3613083
8. Chang Y., Wang X., Wang J., Wu Y., Yang L., Zhu K., Chen H., Yi X., Wang C., Wang Y., Ye W., Zhang Y., Chang Y., Yu P., Yang Q., Xie X. A survey on evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 2024, vol. 15, no. 3, pp. 1–45. https://doi.org/10.1145/3641289
9. Li Z., Cao Y., Xu X., Jiang J., Liu X., Teo Y., Lin S., Liu Y. LLMs for Relational Reasoning: How Far are We? *Proc. of the 1st International Workshop on Large Language Models for Code (LLM4Code '24)*, 2024, pp. 119–126. https://doi.org/10.1145/3643795.3648387
10. Genevey-Metat C., Bachelot D., Gourmelen T., Quemat A., Satre P.-M., Scotto L., Di Perrotolo, Chaux M., Delesques P., Gesny O. Red Team LLM: towards an adaptive and robust automation solution. *Proc. of the Conference on Artificial Intelligence for Defense*, 2023, hal-04328468.
11. Franco M.F., Rodrigues B., Scheid E., Jacobs A., Killer C., Granville L., Stiller B. SecBot: a business-driven conversational agent for cybersecurity planning and management. *Proc. of the 16th International Conference on Network and Service management (CNSM)*. 2020, pp. 1–7. https://doi.org/10.23919/cnsm50824.2020.9269037
12. Chamberlain D., Casey E. Capture the Flag with ChatGPT: Security Testing with AI ChatBots. *Proc. of the 19th International Conference on Cyber Warfare and Security (ICCWS)*, 2024, vol. 19, no 1, pp. 43–54. https://doi.org/10.34190/iccws.19.1.2171
13. Timmins J., Knight S., Lachine B. Offensive cyber security trainer for platform management systems. *Proc. of the IEEE International Systems Conference (SysCon)*, 2021, pp. 1–8. https://doi.org/10.1109/syscon48628.2021.9447060
14. Raman R., Calyam P., Achuthan K. ChatGPT or Bard: Who is a better Certified Ethical Hacker? *Computers & Security*, 2024, vol. 140, pp. 103804. https://doi.org/10.1016/j.cose.2024.103804
15. Abualhaija S., Arora C., Sleimi A., Briand L. Automated question answering for improved understanding of compliance requirements: A multi-document study. *Proc. of the IEEE 30th International Requirements Engineering Conference (RE)*, 2022, pp. 39–50. https://doi.org/10.1109/re54965.2022.00011
16. Renaud K., Warkentin M., Westerman G. From ChatGPT to HackGPT: Meeting the cybersecurity threat of generative AI. *MIT Sloan Management Review*, 2023.
17. Yamin M.M., Hashmi E., Ullah M., Katt B. Applications of LLMs for generating cyber security exercise scenarios. *IEEE Access*, 2024, vol. 12, pp. 143806–143822. https://doi.org/10.1109/access.2024.3468914
18. Heim M.P., Starckjohann N., Torgersen M. *The Convergence of AI and Cybersecurity: An Examination of ChatGPT's Role in Penetration Testing and its Ethical and Legal Implications*. BS thesis. NTNU, 2023.
19. Chen J., Hu S., Zheng H., Xing C., Zhang G. GAIL-PT: An intelligent penetration testing framework with generative adversarial imitation learning. *Computers & Security*, 2023, vol. 126, pp. 103055. https://doi.org/10.1016/j.cose.2022.103055
20. Ananyev V., Chelovechkova A. ChatGPT in information security. *Informacionnaja bezopasnost' cifrovoj ekonomiki*, 2023, pp. 77–83. (in Russian)
21. Prokhorov A.I. Cyberpolygon as a modern information security tool. *Informatization in the Digital Economy*, 2023, vol. 4, no. 4, pp. 363–378. (in Russian). https://doi.org/10.18334/ide.4.4.119301
22. Chen Y., Yao Y., Wang X., Xu D., Yue C., Liu X., Chen K., Tang H., Liu B. Bookworm game: Automatic discovery of LTE vulnerabilities through documentation analysis. *Proc. of the IEEE Symposium on Security and Privacy (SP)*, 2021, pp. 1197–1214. https://doi.org/10.1109/sp40001.2021.00104

23. Ren Z., Ju X., Chen X., Shen H. ProRLearn: boosting prompt tuning-based vulnerability detection by reinforcement learning // Automated Software Engineering. 2024. V. 31. N 2. P. 38. https://doi.org/10.1007/s10515-024-00438-9
24. Hoffmann J. Simulated penetration testing: from "Dijkstra" to "Turing Test++" // Proc. of the 25th International Conference on Automated Planning and Scheduling. 2015. V. 25. N 1. P. 364–372. https://doi.org/10.1609/icaps.v25i1.13684
25. Dube R. Large Language Models in Information Security Research: A January 2024 Survey: preprint. 2024.
26. Ai Z., Luktarhan N., Zhou A., Lv D. Webshell attack detection based on a deep super learner // Symmetry. 2020. V. 12. N 9. P. 1406. https://doi.org/10.3390/sym12091406
27. Esmradi A., Yip D.W., Chan C.F. A Comprehensive Survey of Attack Techniques, Implementation, and Mitigation Strategies in Large Language Models // Communications in Computer and Information Science. 2024. V. 2034. P. 76–95. https://doi.org/10.1007/978-981-97-1274-8_6
28. Gabrys R., Bilinski M., Fugate S., Silva D. Using natural language processing tools to infer adversary techniques and tactics under the Mitre ATT&CK framework // Proc. of the IEEE 14th Annual Computing and Communication Workshop and Conference. 2024. P. 541–547. https://doi.org/10.1109/CCWC60891.2024.10427746
29. Ebert C., Beck M. Generative Artificial Intelligence for Automotive Cybersecurity // ATZelectronics worldwide. 2024. V. 19. N 1. P. 50–54. https://doi.org/10.1007/s38314-023-1564-3
30. Kanakogi K., Washizaki H., Fukazawa Y., Ogata S., Okubo T., Kato T., Kanuka H., Hazeyama A., Yoshioka N. Tracing CVE Vulnerability Information to CAPEC Attack Patterns Using Natural Language Processing Techniques // Information. 2021. V. 12. N 8. P. 298. https://doi.org/10.3390/info12080298
31. Radanliev P., De Roure D., Santos O. Red Teaming Generative AI/NLP, the BB84 Quantum Cryptography Protocol and the NIST-Approved Quantum-Resistant Cryptographic Algorithms. SSRN Electronic Journal. September 17, 2023. URL: https://ssrn.com/abstract=4574446 (дата обращения: 01.10.2024).
32. Sun Y., Wu D., Xue Y., Liu H., Wang H., Xu Z., Xie X., Liu Y. GPTScan: Detecting Logic Vulnerabilities in Smart Contracts by Combining GPT with Program Analysis // Proc. of the IEEE/ACM 46th International Conference on Software Engineering (ICSE 2024). 2024. N 166. P. 1–13. https://doi.org/10.1145/3597503.3639117
33. Al-Hawawreh M., Aljuhani A., Jararweh Y. Chatgpt for cybersecurity: practical applications, challenges, and future directions // Cluster Computing. 2023. V. 26. N 6. P. 3421–3436. https://doi.org/10.1007/s10586-023-04124-5
34. Tsingenopoulos I., Preuveneers D., Joosen W. AutoAttacker: A reinforcement learning approach for black-box adversarial attacks // Proc. of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). 2019. P. 229–237. https://doi.org/10.1109/EuroSPW.2019.00032
35. Rigaki M., Lukas O., Catania C., Garcia S. Out of the cage: how stochastic parrots win in cyber security environments // Proc. of the 16th International Conference on Agents and Artificial Intelligence. 2024. V. 3. P. 774–781. https://doi.org/10.5220/0012391800003636
36. Liu B., Xiao B., Jiang X., Cen S., He X., Dou W. Adversarial attacks on large language model-based system and mitigating strategies: A case study on ChatGPT // Security and Communication Networks. 2023. V. 2023. N 1. P. 8691095. https://doi.org/10.1155/2023/8691095
37. Campbell M., Jovanovic M. Disinfecting AI: Mitigating Generative AI's Top Risks // Computer. 2024. V. 57. N 5. P. 111–116. https://doi.org/10.1109/MC.2024.3374433
38. Намиот Д.Е., Зубарева Е.В. О работе AI Red Team // International Journal of Open Information Technologies. 2023. Т. 11. № 10. P. 130–139.
39. Shi Z., Wang Y., Yin F., Chen X., Chang K., Hsieh C. Red teaming language model detectors with language models // Transactions of the Association for Computational Linguistics. 2024. V. 12. P. 174–189. https://doi.org/10.1162/tacl_a_00639
40. Alawida M., Mejri S., Mehmood A., Chikhaoui B., Abiodun O. A comprehensive study of ChatGPT: advancements, limitations, and ethical considerations in natural language processing and cybersecurity // Information. 2023. V. 14. N 8. P. 462. https://doi.org/10.3390/info14080462
41. Chen Y., Arunasalam A., Celik Z.B. Can large language models provide Security & Privacy advice? measuring the ability of LLms to refute misconceptions // Proc. of the 39th Annual Computer Security Applications Conference. 2023. P. 366–378. https://doi.org/10.1145/3627106.3627196

23. Ren Z., Ju X., Chen X., Shen H. ProRLearn: boosting prompt tuning-based vulnerability detection by reinforcement learning. *Automated Software Engineering*, 2024, vol. 31, no. 2, pp. 38. https://doi.org/10.1007/s10515-024-00438-9
24. Hoffmann J. Simulated penetration testing "Dijkstra" to "Turing Test++". *Proc. of the 25th International Conference on Automated Planning and Scheduling*, 2015, vol. 25, no. 1, pp. 364–372. https://doi.org/10.1609/icaps.v25i1.13684
25. Dube R. Large *Language Models in Information Security Research: A January 2024 Survey*. Preprint. 2024.
26. Ai Z., Luktarhan N., Zhou A., Lv D. Webshell attack detection based on a deep super learner. *Symmetry*, 2020, vol. 12, no. 9, pp. 1406. https://doi.org/10.3390/sym12091406
27. Esmradi A., Yip D.W., Chan C.F. A Comprehensive Survey of Attack Techniques, Implementation, and Mitigation Strategies in Large Language Models. *Communications in Computer and Information Science*, 2024, vol. 2034, pp. 76–95. https://doi.org/10.1007/978-981-97-1274-8_6
28. Gabrys R., Bilinski M., Fugate S., Silva D. Using natural language processing tools to infer adversary techniques and tactics under the Mitre ATT&CK framework. *Proc. of the IEEE 14th Annual Computing and Communication Workshop and Conference*, 2024, pp. 541–547. https://doi.org/10.1109/CCWC60891.2024.10427746
29. Ebert C., Beck M. Generative Artificial Intelligence for Automotive Cybersecurity. *ATZelectronics worldwide*, 2024, vol. 19, no. 1, pp. 50–54. https://doi.org/10.1007/s38314-023-1564-3
30. Hazeyama A., Yoshioka N. Tracing CVE Vulnerability Information to CAPEC Attack Patterns Using Natural Language Processing Techniques. *Information*, 2021, vol. 12, no. 8, pp. 298. https://doi.org/10.3390/info12080298
31. Radanliev P., De Roure D., Santos O. Red Teaming Generative AI/NLP, the BB84 Quantum Cryptography Protocol and the NIST-Approved Quantum-Resistant Cryptographic Algorithms. *SSRN Electronic Journal*. September 17, 2023. URL: https://ssrn.com/abstract=4574446 (accessed: 01.10.2024).
32. Sun Y., Wu D., Xue Y., Liu H., Wang H., Xu Z., Xie X., Liu Y. GPTScan: Detecting Logic Vulnerabilities in Smart Contracts by Combining GPT with Program Analysis. *Proc. of the IEEE/ACM 46th International Conference on Software Engineering (ICSE 2024)*, 2024, no. 166, pp. 1–13. https://doi.org/10.1145/3597503.3639117
33. Al-Hawawreh M., Aljuhani A., Jararweh Y. Chatgpt for cybersecurity: practical applications, challenges, and future directions. *Cluster Computing*, 2023, vol. 26, no. 6, pp. 3421–3436. https://doi.org/10.1007/s10586-023-04124-5
34. Tsingenopoulos I., Preuveneers D., Joosen W. AutoAttacker: A reinforcement learning approach for black-box adversarial attacks. *Proc. of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2019, pp. 229–237. https://doi.org/10.1109/EuroSPW.2019.00032
35. Rigaki M., Lukas O., Catania C., Garcia S. Out of the cage: how stochastic parrots win in cyber security environments. *Proc. of the 16th International Conference on Agents and Artificial Intelligence*, 2024, vol. 3. pp. 774–781. https://doi.org/10.5220/0012391800003636
36. Liu B., Xiao B., Jiang X., Cen S., He X., Dou W. Adversarial attacks on large language model-based system and mitigating strategies: A case study on ChatGPT. *Security and Communication Networks*, 2023, vol. 2023, no. 1, pp. 8691095. https://doi.org/10.1155/2023/8691095
37. Campbell M., Jovanovic M. Disinfecting AI: Mitigating Generative AI's Top Risks. *Computer*, 2024, vol. 57, no. 5, pp. 111–116. https://doi.org/10.1109/MC.2024.3374433
38. Namiot D., Zubareva E. About AI Red Team. *International Journal of Open Information Technologies*, 2023, vol. 11, no. 10, pp. 130–139. (in Russian)
39. Shi Z., Wang Y., Yin F., Chen X., Chang K., Hsieh C. Red teaming language model detectors with language models. *Transactions of the Association for Computational Linguistics*, 2024, vol. 12, pp. 174–189. https://doi.org/10.1162/tacl_a_00639
40. Alawida M., Mejri S., Mehmood A., Chikhaoui B., Abiodun O. A comprehensive study of ChatGPT: advancements, limitations, and ethical considerations in natural language processing and cybersecurity. *Information*, 2023, vol. 14, no. 8, pp. 462. https://doi.org/10.3390/info14080462
41. Chen Y., Arunasalam A., Celik Z.B. Can large language models provide Security & Privacy advice? measuring the ability of LLms to refute misconceptions. *Proc. of the 39th Annual Computer Security*

42. Bang Y., Cahyawijaya S., Lee N., Dai W., Su D., Wilie B., Lovenia H., Ji Z., Yu T., Chung W., Do Q.V., Xu Y, Fung P. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity // Proc. of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. 2023. V. 1. P. 675–718. https://doi.org/10.18653/v1/2023.ijcnlp-main.45
43. Xie Y., Yi J., Shao J., Curl J., Lyu L., Chen Q., Xie X., Wu F. Defending ChatGPT against jailbreak attack via self-reminders // Nature Machine Intelligence. 2023. V. 5. N 12. P. 1486–1496. https://doi.org/10.1038/s42256-023-00765-8
44. Deng G., Liu Y., Li Y., Wang K., Zhang Y., Li Z., Wang H., Zhang T., Liu Y. MASTERKEY: Automated jailbreaking of large language model Chatbots // Proc. of the Network and Distributed System Security Symposium. 2024. P. 1–16. https://doi.org/10.14722/ndss.2024.24188
45. Schulhoff S., Pinto J., Khan A., Bouchard L., Si C., Anati S., Tagliabue V., Kost A., Carnahan C., Boyd-Graber J. Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs Through a Global Prompt Hacking Competition // Proc. of the Conference on Empirical Methods in Natural Language Processing. 2023. P. 4945–4977. https://doi.org/10.18653/v1/2023.emnlp-main.302
46. Liang H., Li X., Xiao D., Liu J., Zhou Y., Wang A., Li J. Generative Pre-Trained Transformer-Based reinforcement learning for testing Web Application Firewalls // IEEE Transactions on Dependable and Secure Computing. 2024. V. 21. N 1. P. 309–324. https://doi.org/10.1109/TDSC.2023.3252523
47. Серобабов Д.С. Безопасность кода, созданного с помощью Chat GPT: анализ и исправление уязвимостей //Фундаментальные и прикладные исследования молодых учёных: сборник материалов VII Международной научно-практической конференции студентов, аспирантов и молодых учёных, приуроченной к 110-летию со дня рождения Т.В. Алексеевой, Омск, 20–21 апреля 2023 года. Омск: Сибирский государственный автомобильно-дорожный университет (СибАДИ). 2023. С. 615–620.
48. Gasiba T.E., Oguzhan K., Kessba I., Lechner U., Pinto-Albuquerque M. I'm Sorry Dave, I'm Afraid I Can't Fix Your Code: On ChatGPT, CyberSecurity, and Secure Coding // Proc. of the 4th International Computer Programming Education Conference (ICPEC 2023). 2023.
49. Lu G., Ju X., Chen X., Pei W., Cai Z. GRACE: Empowering LLM-based software vulnerability detection with graph structure and in-context learning // Journal of Systems and Software. 2024. V. 212. P. 112031. https://doi.org/10.1016/j.jss.2024.112031
50. Wang J., Huang Y., Chen C., Liu Z., Wang S.,Wang Q. Software testing with Large Language Models: Survey, Landscape, and Vision // IEEE Transactions on Software Engineering. 2024. V. 50. N 4. P. 911–936. https://doi.org/10.1109/TSE.2024.3368208
51. Chaudhary P.K. AI, ML, and Large Language Models in cybersecurity // International Research Journal of Modernization in Engineering Technology and Science. 2024. V. 6. N 2. P. 2229–2234. https://www.doi.org/10.56726/IRJMETS49546
52. Botacin M. GPThreats-3: Is Automatic Malware Generation a Threat? // Proc. of the IEEE Security and Privacy Workshops (SPW). 2023. P. 238–254. https://www.doi.org/10.1109/SPW59333.2023.00027

Applications Conference, 2023, pp. 366–378. https://doi.org/10.1145/3627106.3627196
42. Bang Y., Cahyawijaya S., Lee N., Dai W., Su D., Wilie B., Lovenia H., Ji Z., Yu T., Chung W., Do Q.V., Xu Y, Fung P. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *Proc. of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2023, vol. 1, pp. 675–718. https://doi.org/10.18653/v1/2023.ijcnlp-main.45
43. Xie Y., Yi J., Shao J., Curl J., Lyu L., Chen Q., Xie X., Wu F. Defending ChatGPT against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 2023, vol. 5, no. 12, pp. 1486–1496. https://doi.org/10.1038/s42256-023-00765-8
44. Deng G., Liu Y., Li Y., Wang K., Zhang Y., Li Z., Wang H., Zhang T., Liu Y. MASTERKEY: Automated jailbreaking of large language model Chatbots. *Proc. of the Network and Distributed System Security Symposium*, 2024, pp. 1–16. https://doi.org/10.14722/ndss.2024.24188
45. Schulhoff S., Pinto J., Khan A., Bouchard L., Si C., Anati S., Tagliabue V., Kost A., Carnahan C., Boyd-Graber J. Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs Through a Global Prompt Hacking Competition. *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 4945–4977. https://doi.org/10.18653/v1/2023.emnlp-main.302
46. Liang H., Li X., Xiao D., Liu J., Zhou Y., Wang A., Li J. Generative Pre-Trained Transformer-Based reinforcement learning for testing Web Application Firewalls. *IEEE Transactions on Dependable and Secure Computing*, 2024, vol. 21, no. 1, pp. 309–324. https://doi.org/10.1109/TDSC.2023.3252523
47. Serobabov D.C., Razumov S.Y. Security of code created using Chat GPT: vulnerability analysis and correction. *Fundamental'nye i prikladnye issledovanija molodyh uchjonyh*, 2023, pp. 615–620. (in Russian)
48. Gasiba T.E., Oguzhan K., Kessba I., Lechner U., Pinto-Albuquerque M. I'm Sorry Dave, I'm Afraid I Can't Fix Your Code: On ChatGPT, CyberSecurity, and Secure Coding. *Proc. of the 4th International Computer Programming Education Conference (ICPEC 2023)*. 2023.
49. Lu G., Ju X., Chen X., Pei W., Cai Z. GRACE: Empowering LLM-based software vulnerability detection with graph structure and in-context learning. *Journal of Systems and Software*, 2024, vol. 212, pp. 112031. https://doi.org/10.1016/j.jss.2024.112031
50. Wang J., Huang Y., Chen C., Liu Z., Wang S.,Wang Q. Software testing with Large Language Models: Survey, Landscape, and Vision. *IEEE Transactions on Software Engineering*, 2024, vol. 50, no. 4, pp. 911–936. https://doi.org/10.1109/TSE.2024.3368208
51. Chaudhary P.K. AI, ML, and Large Language Models in cybersecurity. *International Research Journal of Modernization in Engineering Technology and Science*, 2024, vol. 6, no. 2, pp. 2229–2234. https://www.doi.org/10.56726/IRJMETS49546
52. Botacin M. GPThreats-3: Is Automatic Malware Generation a Threat? *Proc. of the IEEE Security and Privacy Workshops (SPW)*, 2023, pp. 238–254. https://www.doi.org/10.1109/SPW59333.2023.00027

Авторы

Конев Антон Александрович — кандидат технических наук, доцент, заместитель директора Института системной интеграции и безопасности, доцент кафедры КИБЭВС, Томский государственный университет систем управления и радиоэлектроники, Томск, 634050, Российская Федерация, SC 23035057200, https://orcid.org/0000-0002-3222-9956, kaa@fb.tusur.ru

Паюсова Татьяна Игоревна — доцент, Тюменский государственный университет, Тюмень, 625003, Российская Федерация, SC 57188574761, https://orcid.org/0000-0003-4923-1689, t.i.payusova@utmn.ru

Authors

Anton A. Konev — PhD, Associate Professor, Deputy Director of the Institute of System Integration and Security, Associate Professor of the Department, Tomsk State University of Control Systems and Radioelectronics, Tomsk, 634050, Russian Federation, SC 23035057200, https://orcid.org/0000-0002-3222-9956, kaa@fb.tusur.ru

Tatyana I. Payusova — Associate Professor, Tyumen State University, Tyumen, 625003, Russian Federation, SC 57188574761, https://orcid.org/0000-0003-4923-1689, t.i.payusova@utmn.ru

Статья поступила в редакцию 07.10.2024
Одобрена после рецензирования 02.11.2024
Принята к печати 23.01.2025

Received 07.10.2024
Approved after reviewing 02.11.2024
Accepted 23.01.2025