Рубрика 2. Методы и системы защиты информации, информационная безопасность

УДК *004.056.5:004.8*

В. С. Греков [1]

[1]*ФГАОУ ВО «РГУ нефти и газа (НИУ) имени И.М. Губкина», г. Москва, Российская Федерация*

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ В АВТОМАТИЗАЦИИ ТЕСТИРОВАНИЯ НА ПРОНИКНОВЕНИЕ: СОВРЕМЕННОЕ СОСТОЯНИЕ, ОГРАНИЧЕНИЯ И ПЕРСПЕКТИВЫ ПРИМЕНЕНИЯ В ПРОФЕССИОНАЛЬНОЙ ПОДГОТОВКЕ СПЕЦИАЛИСТОВ ПО ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Аннотация. *Статья посвящена анализу современного состояния исследований в области применения больших языковых моделей (LLM) для автоматизации тестирования на проникновение (пентеста). Рассматриваются причины стремительного роста интереса к данному направлению, обусловленные усложнением ИТ-инфраструктур, высокой стоимостью традиционных методов аудита защищённости и развитием концепции DevSecOps. На основе систематизации обзорных и прикладных исследований 2023–2025 годов выявлены четыре ключевых направления применения LLM в кибербезопасности: поддержка принятия решений и стратегического планирования атак, автоматизация этапов пентеста, обеспечение защищённости самих языковых моделей (AI Red Teaming), а также интеллектуальный анализ исходного кода. Проведён критический анализ архитектур мультиагентных систем с использованием подходов Retrieval-Augmented Generation (RAG), систем тестирования и оценки эффективности, а также коммерческих решений. Установлено, что даже наиболее продвинутые системы пентеста на базе LLM находятся на промежуточных уровнях автономности (3–4 из 5), не способны самостоятельно формулировать цели и оценивать последствия своих решений. Обосновывается тезис о том, что на конец 2025 года LLM выступают прежде всего инструментом аналитика, а не автономным пентестером. Рассматриваются педагогические аспекты интеграции LLM-инструментов в профессиональную подготовку специалистов по информационной безопасности, обосновывается необходимость формирования у обучающихся критического мышления в отношении возможностей и ограничений данных технологий. Отдельное внимание уделяется требованиям к воспроизводимости исследований и педагогической интерпретации полученных результатов.*

Ключевые слова: *большие языковые модели, тестирование на проникновение, кибербезопасность, автоматизация, мультиагентные системы, RAG, профессиональная подготовка.*

Введение

Цифровая трансформация экономики и общества сопровождается не только расширением функциональных возможностей информационных технологий, но и качественным усложнением угроз информационной безопасности. Современные корпоративные ИТ-инфраструктуры — это распределённые гетерогенные системы, включающие облачные платформы, микросервисные архитектуры, контейнеризированные приложения, конвейеры непрерывной интеграции и доставки (CI/CD), а также устройства Интернета вещей. Указанная сложность формирует принципиально новые требования к компетенциям специалистов по информационной безопасности, в том числе реализующих тестирование на проникновение (пентест) — методологически структурированную практику имитации атак с целью выявления уязвимостей до их эксплуатации злоумышленниками.

Традиционный пентест, выполняемый квалифицированными специалистами в ручном режиме, характеризуется рядом системных ограничений: высокой стоимостью, значительными временными затратами, трудностью масштабирования, зависимостью результатов от субъективного опыта конкретного исполнителя. В условиях непрерывного расширения атакующей поверхности организаций и ужесточения требований к регулярности и полноте аудита защищённости формируется устойчивый запрос на автоматизированные и воспроизводимые средства анализа.

Начиная с 2023 года фиксируется интенсивный рост интереса исследовательского и практического сообщества к применению больших языковых моделей (Large Language Models, LLM) в задачах кибербезопасности, включая автоматизацию пентеста. Указанный интерес обусловлен характеристиками LLM: способностью к семантическому пониманию технической документации, генерации исполняемого кода, планированию многошаговых действий, интерпретации результатов сканирования в контексте известных таксономий уязвимостей. LLM открывают возможность формирования семантического слоя поверх существующих

инструментов пентеста — гибкого интерфейса между специалистом и конвейером тестирования с относительно низкими операционными издержками.

Обобщённая схема факторов, определяющих рост интереса к применению LLM в задачах пентеста, представлена на рис. 1.

Рис. 1. Факторы роста интереса к применению LLM в автоматизации тестирования на проникновение

Интенсивность развития данного направления порождает серьёзные методологические проблемы. Исследовательское поле остаётся фрагментированным. Большинство публикуемых результатов получены на синтетических стендах, а реальная автономность существующих систем существенно уступает декларируемой. Указанное обстоятельство формирует риск формирования у специалистов и обучающихся завышенных ожиданий относительно функциональных возможностей LLM в задачах информационной безопасности.

Настоящая статья решает две взаимосвязанные задачи. Первая: систематизация актуального состояния исследований в области применения LLM для автоматизации пентеста. Вторая: анализ педагогических аспектов подготовки специалистов по информационной безопасности в условиях, когда LLM-инструменты приобретают статус элемента профессионального арсенала. Как отмечает А.М. Новиков, «профессиональное образование должно опережать производство» [1, с. 47]. Указанное требование определяет обоснованность включения тематики LLM в кибербезопасности в образовательные программы подготовки специалистов соответствующего профиля.

Материалы и методы

В основу исследования положен нарративный обзор научных публикаций, посвящённых применению LLM в задачах кибербезопасности и пентеста, опубликованных в период с января 2023 по март 2025 года. Для формирования корпуса источников использовались базы данных arXiv, IEEE Xplore, ACM Digital Library, а также репозитории открытого программного обеспечения (GitHub). Критерии включения: наличие эмпирических результатов или систематического обзора; фокус на применении LLM именно в задачах тестирования на проникновение или смежных задачах кибербезопасности; доступность полного текста.

Методологическую основу составили: метод обзорно-аналитического сопоставления литературы, сравнительный анализ архитектурных решений, критический анализ методологии экспериментов, а также контент-анализ коммерческих продуктов в данной области. При анализе педагогических аспектов применялись методы теоретического анализа и синтеза, опирающиеся на труды отечественных и зарубежных исследователей в области профессионального образования и дидактики высшей школы.

Теоретическую базу педагогической части исследования составляют концепция профессионального образования А.М. Новикова [1], теория деятельностного подхода А.Н. Леонтьева

[2], концепция проблемного обучения И.Я. Лернера [3], а также современные исследования в области подготовки специалистов по информационной безопасности [4; 5].

Результаты

1. Обзор исследовательского поля: от глубокого обучения к LLM

Применение методов искусственного интеллекта в задачах кибербезопасности имеет достаточно длительную историю. По данным обзорного исследования [6], в период с 2015 по 2023 год в данной области доминировали алгоритмы обучения с подкреплением (Reinforcement Learning) и глубокого обучения (Deep Learning). Эти подходы демонстрировали определённые успехи в узкоспециализированных задачах — обнаружении аномалий, классификации вредоносного программного обеспечения, анализе сетевого трафика, — однако оставались ограниченными в части семантического понимания контекста и способности к многошаговому планированию.

Появление больших языковых моделей, прежде всего семейства GPT (OpenAI), Claude (Anthropic), Llama (Meta) и их аналогов, ознаменовало качественный сдвиг в подходах к автоматизации задач кибербезопасности. Как показывает анализ корпуса работ начиная с января 2023 года [7], исследовательское поле вокруг применения LLM в информационной безопасности стало многомерным и быстро формирующимся направлением. Наибольший интерес вызывают четыре направления:

– поддержка принятия решений и стратегического планирования атак;
– автоматизация этапов тестирования на проникновение;
– обеспечение защищённости самих языковых моделей (AI Red Teaming);
– интеллектуальный анализ исходного кода.

Указанные направления применения LLM в задачах кибербезопасности систематизированы на рис. 2.

Рис. 2. Основные направления применения LLM в задачах кибербезопасности

Принципиальное отличие LLM от предшествующих подходов к автоматизации пентеста состоит в функциональной роли этих моделей. LLM применяются не как самостоятельные классификаторы или агенты, а как семантический слой поверх существующих инструментов: модели обеспечивают планирование, установление соответствия между результатами сканирования и таксономиями уязвимостей (MITRE ATT&CK, CVE, CWE), формулирование исполнимых шагов тестирования, а также реализуют функцию интерфейса между специалистом и конвейером тестирования [6].

Начиная с 2024 года, несмотря на рост числа прикладных публикаций и разработку коммерческих решений в данной области, исследовательское поле сохраняет фрагментированный и методологически неоднородный характер [7]. Указанное обстоятельство имеет

существенное значение как для практиков, так и для педагогов: отсутствие устоявшейся методологической базы обусловливает необходимость критического анализа публикуемых результатов.

2. Мультиагентные системы с RAG: возможности и ограничения

Мультиагентные системы на основе подхода Retrieval-Augmented Generation (RAG) составляют наиболее интенсивно разрабатываемое направление в области LLM-пентеста. Архитектурная суть подхода состоит в расширении языковой модели внешней базой знаний, включающей документацию по уязвимостям, базы эксплойтов и технические руководства. В режиме реального времени из указанной базы извлекаются релевантные фрагменты для формирования контекстно-обогащённых запросов к модели. Применение RAG-архитектуры обеспечивает преодоление ограничений, связанных с устареванием обучающих данных, и поддержание актуальности генерируемых рекомендаций.

Обобщённая логика построения мультиагентной RAG-системы для задач пентеста представлена на рис. 3.

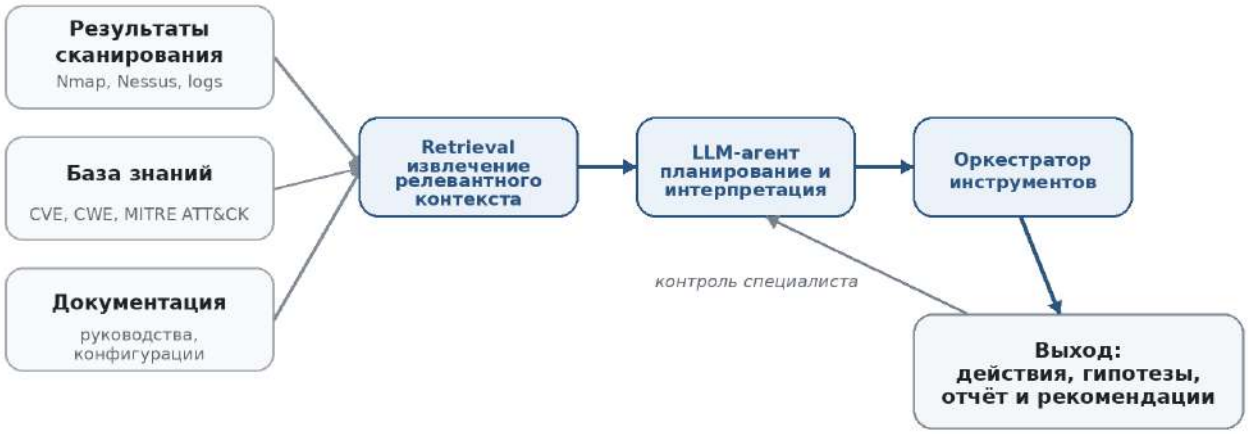

Рис. 3. Обобщённая логика мультиагентной RAG-системы в тестировании на проникновение

Анализ отобранных для обзорного сопоставления публикаций [8–18] позволяет выделить следующие преимущества мультиагентных RAG-систем в контексте пентеста:

– автоматизация рутинных этапов тестирования (сбор информации, сканирование портов, идентификация сервисов);

– способность к контекстному связыванию результатов различных инструментов;

– генерация структурированных отчётов с рекомендациями по устранению уязвимостей;

– снижение когнитивной нагрузки на специалиста при работе с большими объёмами данных.

Вместе с тем критический анализ выявляет ряд существенных ограничений, которые нередко остаются в тени в публикациях, ориентированных на демонстрацию достижений.

Первое ограничение — отсутствие системного подхода к проектированию архитектуры. Большинство опубликованных работ представляют конкретный прототип, решающий одну конкретную задачу в специфических условиях. Обобщаемость предложенных архитектурных решений на другие классы задач или иные операционные среды, как правило, не обосновывается.

Второе ограничение — проблема утечки данных в обучающий стек (data contamination). Выбранные для тестирования задачи практически наверняка входили в обучающие данные для моделей: синтетические примеры с соревнований по кибербезопасности (CTF — Capture The Flag) и уязвимых виртуальных машин (типа Metasploitable, HackTheBox, TryHackMe), решения к которым публично доступны в интернете. Это означает, что модель может демонстрировать высокую «производительность» не за счёт реального понимания задачи, а за счёт воспроизведения паттернов из обучающих данных.

Третье ограничение — отставание от актуального состояния моделей. Исследования, как правило, отстают от SoTA (State of the Art) моделей на год-полтора. Учитывая темпы развития LLM, это означает, что к моменту публикации результаты могут быть уже неактуальны.

Четвёртое ограничение — отсутствие воспроизводимости. Большинство экспериментов не могут быть воспроизведены независимыми исследователями: не публикуются промпты, конфигурации агентов, версии используемых моделей и инструментов. Это фундаментальная проблема для научного сообщества, поскольку, как подчёркивает В.В. Краевский, «воспроизводимость результатов является необходимым условием научного знания» [19, с. 112].

3. Системы тестирования и оценки эффективности (benchmarks и evals)

Параллельно с разработкой прикладных систем ведётся работа по созданию инструментов для их объективной оценки. Анализ работ [15–18] показывает, что данное направление находится в начальной стадии развития и также страдает от ряда системных проблем.

С одной стороны, наличие специализированных бенчмарков для LLM-пентеста является необходимым условием научного прогресса в данной области: без стандартизированных метрик невозможно корректно сравнивать различные системы и отслеживать динамику развития. С другой стороны, существующие инструменты оценки имеют существенные ограничения.

Во-первых, отсутствует системный подход к проектированию архитектуры систем тестирования и оценки. Каждая исследовательская группа создаёт собственный бенчмарк, что делает невозможным прямое сравнение результатов разных работ.

Во-вторых, с учётом стремительного роста эффективности моделей существующие инструменты не позволяют в полной мере оценивать качество выполнения задач, степень галлюцинаций и совокупный скор каждой модели. Бенчмарки, разработанные для моделей 2023 года, оказываются «насыщенными» (saturated) для моделей 2025 года — то есть все системы показывают близкие к максимальным результаты, что лишает бенчмарк дифференцирующей силы.

В-третьих, проблема утечки данных в обучающий стек актуальна и для систем оценки: если задания бенчмарка публично доступны, они с высокой вероятностью входят в обучающие данные тестируемых моделей.

Исследование [5], посвящённое применению LLM в образовании по пентесту, предлагает интересный подход: использование LLM не только как инструмента тестирования, но и как средства мониторинга прогресса студентов, поддержки совместной работы и предоставления обратной связи в больших учебных группах. Авторы отмечают, что «образование в области кибербезопасности является сложным, и понимание возможностей LLM для поддержки образования никогда не было более важным» [5]. Это наблюдение имеет прямое отношение к педагогической проблематике настоящей статьи.

4. Оценка автономности: где реально находятся LLM-системы пентеста

Одним из наиболее методологически значимых вкладов в данную область является работа [19], в которой введена терминологическая ясность касательно автономности и различных её уровней в контексте тестирования на проникновение. Авторы предлагают пятиуровневую шкалу автономности:

– Уровень 1: Полностью ручное тестирование; LLM используется только как справочный инструмент.

– Уровень 2: LLM предлагает следующий шаг, специалист принимает решение и выполняет действие.

– Уровень 3: LLM планирует и выполняет отдельные этапы, специалист осуществляет общий надзор.

– Уровень 4: LLM выполняет большинство задач автономно, специалист вмешивается при необходимости.

– Уровень 5: Полностью автономное тестирование без участия человека.

Сводная интерпретация уровней автономности LLM-систем пентеста приведена на рис. 4.

Рис. 4. Шкала автономности LLM-систем тестирования на проникновение

Качественное и количественное сравнение инструментов автоматического пентеста, проведённое в данной работе, приводит к справедливому выводу: даже самые продвинутые системы пентеста на базе LLM находятся на промежуточных уровнях 3–4. Модели не могут самостоятельно формулировать цели и оценивать последствия своих решений — это принципиальное ограничение, обусловленное самой природой языковых моделей как систем, оптимизированных для предсказания следующего токена, а не для целеполагания и стратегического планирования.

Данный вывод имеет важное педагогическое следствие: подготовка специалистов по информационной безопасности не может строиться на предположении о полной автоматизации пентеста в обозримом будущем. Напротив, как указывает И.Я. Лернер, «проблемное обучение формирует у учащихся способность к самостоятельному решению нестандартных задач» [3, с. 89] — именно эта способность остаётся незаменимой в условиях, когда автоматизированные системы достигают своих пределов.

5. Коммерческие решения: возможности и риски

Параллельно с академическими исследованиями активно развивается рынок коммерческих LLM-решений для задач кибербезопасности [20–23]. Анализ доступных продуктов позволяет выделить их ключевые преимущества и недостатки.

Преимущества коммерческих решений: хорошая масштабируемость — возможность одновременного тестирования большого числа систем; способность выявлять уязвимости бизнес-логики, которые традиционные сканеры пропускают; частичное покрытие программных уязвимостей; экономическая эффективность — дешевле, чем использование лицензированного ПО автоматизированных сканеров уязвимостей не на базе LLM, и значительно дешевле, чем полноценный операционный центр безопасности (SOC).

Недостатки коммерческих решений: непрозрачность («кот в мешке») — отсутствуют инструменты для независимой оценки эффективности, гарантии надёжности и устойчивости платформы; неясность в отношении использования конфиденциальных данных клиентов, передаваемых в систему в процессе тестирования; исследование подключённых инструментов и баз данных «вслепую» потенциально влечёт за собой избыточные временные и финансовые затраты; отсутствие стандартизированных метрик для сравнения с конкурирующими решениями.

Эти ограничения особенно важны с точки зрения профессиональной подготовки: будущие специалисты должны уметь критически оценивать коммерческие инструменты, не принимая на веру маркетинговые заявления производителей. Как подчёркивает С.Я. Батышев, «профессиональная компетентность специалиста включает не только умение применять инструменты, но и способность критически оценивать их возможности и ограничения» [20, с. 156].

6. Наиболее перспективный открытый проект: Cybersecurity AI (CAI)

Рубрика 2. Методы и системы защиты информации, информационная безопасность

Среди рассмотренных решений особого внимания заслуживает проект Cybersecurity AI (CAI) [24] — открытый инструмент, ориентированный на активное использование в академической среде. Его ключевые преимущества:

– открытость исходного кода — возможность независимой верификации алгоритмов и архитектурных решений;

– пополняемая научная база — исследования эффективности публикуются и доступны для критического анализа;

– практическое тестирование — инструмент проходит апробацию на различных платформах по кибербезопасности и в «полевых» условиях;

– прозрачность и документация — подробная документация позволяет воспроизводить эксперименты.

Вместе с тем CAI не лишён ограничений: при заявленной полной автоматизации инструмент всё же требует участия специалиста; наиболее продвинутая модель предоставляется по подписке и не находится в публичном доступе; по-прежнему отсутствует системный подход к тестированию производительности.

С педагогической точки зрения CAI представляет значительный интерес именно как образовательный инструмент: его открытость позволяет использовать его в учебном процессе для демонстрации принципов работы LLM-агентов в задачах безопасности, а также для формирования у студентов навыков критического анализа автоматизированных систем.

Таблица 1 – Сводная характеристика материалов презентации по прикладным LLM-решениям в пентесте

Направление / материал	Потенциал применения	Ключевые ограничения	Значение для подготовки специалистов
Мультиагентные RAG-системы	Автоматизация рутинных этапов, связывание результатов инструментов, генерация отчётов.	Фрагментарность архитектур, риск data contamination, отставание от SoTA, слабая воспроизводимость.	Использовать как объект критического разбора архитектуры и качества эксперимента.
Benchmarks и evals	Формирование базы для сравнения LLM-систем и отслеживания динамики качества.	Несопоставимость методик, насыщение тестов, утечка заданий в обучающие данные.	Обучать студентов постановке корректной оценки и анализу метрик.
Оценка автономности	Разделение уровней участия человека и автоматизации отдельных этапов.	Даже продвинутые решения фактически находятся на уровнях 3–4 из 5.	Показывать границы автоматизации и необходимость профессионального надзора.
Коммерческие решения	Масштабируемость, частичное выявление уязвимостей бизнес-логики, снижение стоимости проверок.	Непрозрачность, риски конфиденциальности, отсутствие независимых метрик.	Формировать навык критической оценки продуктов и маркетинговых заявлений.
Cybersecurity AI (CAI)	Open-source, документация, академическая ориентация, возможность воспроизводимой апробации.	Требует участия специалиста; продвинутая модель доступна по подписке; тестирование пока незрелое.	Подходит для учебных стендов и демонстрации принципов LLM-агентов.

Обсуждение

7. Реальная роль LLM в автоматизации пентеста: итоговая оценка

Систематизация рассмотренных исследований позволяет сформулировать взвешенную оценку реальной роли LLM в автоматизации тестирования на проникновение по состоянию на конец 2025 года.

Применение LLM в пентесте остаётся экспериментальным направлением. Методологическая база только формируется: отсутствуют общепринятые стандарты проектирования систем, метрики оценки и протоколы воспроизводимости экспериментов. Большинство решений — прототипы для учебных стендов и CTF-сценариев, демонстрирующие частичные успехи в контролируемых условиях.

При переносе в реальные корпоративные или облачные среды проявляются ключевые ограничения. Нестабильность поведения моделей, слабая воспроизводимость результатов и низкая автономность делают существующие системы непригодными для самостоятельного использования в производственных условиях без постоянного надзора квалифицированного специалиста.

Отсутствуют открытые эталоны, формализованные метрики и зрелые механизмы верификации выводов моделей. Это означает, что любые заявления о «высокой эффективности» той или иной системы должны восприниматься с критической осторожностью.

На конец 2025 года LLM — это прежде всего инструмент аналитика, а не автономный пентестер. Конкретные задачи, в которых LLM демонстрируют реальную ценность:

– сбор и агрегация разрозненных данных из множества источников;
– анализ технической документации и конфигурационных файлов;
– генерация инсайтов и подсказок для специалистов уровня L1–L2;
– автоматизация составления отчётов;
– поддержка принятия решений при классификации уязвимостей.

Сводное место LLM в контуре пентеста отражено на рис. 5.

Рис. 5. Итоговая роль LLM в контуре тестирования на проникновение

Интеграция LLM в CI/CD-цепочки пентеста возможна для узких задач (например, автоматическая проверка конфигураций на соответствие политикам безопасности), но до полной автоматизации полного цикла тестирования на проникновение ещё далеко.

8. Педагогические аспекты подготовки специалистов по ИБ в условиях развития LLM

Описанное состояние исследовательского поля ставит перед системой профессионального образования ряд принципиальных вопросов. Как готовить специалистов по информационной безопасности в условиях, когда LLM-инструменты становятся частью профессионального арсенала, но при этом их возможности существенно ограничены, а методологическая база нестабильна?

8.1. Формирование критического мышления как ключевая образовательная задача

Первоочередной педагогической задачей является формирование у обучающихся критического мышления в отношении возможностей и ограничений LLM-инструментов. Это требование вытекает непосредственно из анализа состояния исследовательского поля: специалист, некритически доверяющий выводам LLM-системы, рискует пропустить реальные уязвимости или, напротив, потратить ресурсы на ложные срабатывания.

Как отмечает И.Я. Лернер, «проблемное обучение не только вооружает учащихся знаниями, но и формирует у них опыт творческой деятельности» [3, с. 94]. Применительно к подготовке специалистов по ИБ это означает необходимость включения в учебный процесс задач, требующих не просто применения LLM-инструментов, но и критической оценки их результатов, выявления ошибок и галлюцинаций, сопоставления автоматически сгенерированных рекомендаций с собственным профессиональным суждением.

8.2. Деятельностный подход и симуляция реальных условий

Деятельностный подход, разработанный в трудах А.Н. Леонтьева [2] и получивший развитие в концепции контекстного обучения А.А. Вербицкого [21], предполагает организацию учебного процесса таким образом, чтобы студенты осваивали профессиональные компетенции в условиях, максимально приближённых к реальной профессиональной деятельности.

В контексте подготовки специалистов по ИБ это означает необходимость создания учебных стендов, имитирующих реальные корпоративные инфраструктуры, и организации практических занятий, в ходе которых студенты используют LLM-инструменты в связке с традиционными средствами пентеста. При этом принципиально важно, чтобы учебные задачи не ограничивались CTF-сценариями (которые, как было показано выше, могут входить в обучающие данные моделей), но включали нестандартные ситуации, требующие самостоятельного профессионального суждения.

Исследование [5] подтверждает педагогическую ценность такого подхода: авторы показывают, что LLM могут эффективно использоваться для мониторинга прогресса студентов и предоставления персонализированной обратной связи в больших учебных группах, что

особенно актуально в условиях дефицита квалифицированных преподавателей по кибербезопасности.

8.3. Проблема актуальности учебного контента

Стремительное развитие LLM-технологий создаёт серьёзную проблему для системы профессионального образования: учебные программы и материалы устаревают быстрее, чем успевают обновляться. Как подчёркивает А.М. Новиков, «содержание профессионального образования должно отражать не только сегодняшние, но и завтрашние требования производства» [1, с. 203].

Решение этой проблемы требует перехода от статичных учебных программ к динамичным, предусматривающим регулярное обновление контента в соответствии с актуальным состоянием технологий. При этом важно не просто включать в программу описание новейших инструментов, но и формировать у студентов метакомпетенции — способность самостоятельно осваивать новые технологии, критически оценивать их возможности и интегрировать в профессиональную деятельность.

8.4. Этические аспекты подготовки специалистов

Применение LLM в задачах кибербезопасности поднимает ряд этических вопросов, которые должны быть отражены в образовательных программах. Снижение порога входа в профессию пентестера за счёт LLM-инструментов означает, что те же инструменты могут использоваться злоумышленниками с недостаточной квалификацией. Это требует особого внимания к формированию у студентов профессиональной этики и понимания правовых рамок деятельности специалиста по информационной безопасности.

Как отмечает В.А. Сластёнин, «профессиональная подготовка специалиста неотделима от его нравственного воспитания» [22, с. 78]. Применительно к специалистам по ИБ это означает необходимость включения в образовательные программы не только технических дисциплин, но и курсов по профессиональной этике, правовому регулированию в сфере информационной безопасности и ответственному раскрытию уязвимостей.

8.5. Требования к компетенциям преподавателей

Эффективная подготовка специалистов по ИБ в условиях развития LLM-технологий предъявляет повышенные требования к компетенциям преподавателей. Педагог, не имеющий практического опыта работы с LLM-инструментами в задачах кибербезопасности, не способен сформировать у студентов адекватное понимание их возможностей и ограничений.

Это требование согласуется с положениями ФГОС ВО по направлению «Информационная безопасность», предусматривающими регулярное повышение квалификации преподавателей с привлечением специалистов-практиков из отрасли. Как указывает Э.Ф. Зеер, «профессиональное развитие педагога является необходимым условием качества профессионального образования» [23, с. 134].

8.6. Структура учебного модуля по LLM в кибербезопасности

На основе проведённого анализа можно предложить следующую структуру учебного модуля, посвящённого применению LLM в задачах кибербезопасности, для включения в образовательные программы подготовки специалистов по ИБ.

Тема 1. Основы больших языковых моделей (теоретический блок): архитектура трансформеров и принципы работы LLM; возможности и ограничения LLM — галлюцинации, data contamination, проблема актуальности знаний; обзор ключевых моделей: GPT-4o, Claude, Llama, специализированные модели для кибербезопасности.

Тема 2. LLM как инструмент аналитика безопасности (практический блок): применение LLM для анализа технической документации и конфигураций; генерация и интерпретация результатов сканирования; автоматизация составления отчётов о результатах тестирования; практическая работа — использование LLM для анализа результатов сканирования Nmap/Nessus.

Тема 3. Мультиагентные системы и RAG в пентесте (теоретико-практический блок): архитектура мультиагентных систем; принципы RAG и их применение в задачах

кибербезопасности; критический анализ существующих систем (PentestGPT, CAI и др.); практическая работа — развёртывание и тестирование открытого LLM-агента на учебном стенде.

Тема 4. Оценка эффективности и ограничения LLM-систем (аналитический блок): методология оценки LLM-систем в задачах безопасности; проблема воспроизводимости и data contamination; уровни автономности — где реально находятся существующие системы; практическая работа — сравнительный анализ результатов LLM-агента и ручного тестирования на одном стенде.

Тема 5. Этические и правовые аспекты (гуманитарный блок): правовое регулирование тестирования на проникновение в России и за рубежом; этика ответственного раскрытия уязвимостей; риски злоупотребления LLM-инструментами; профессиональные стандарты и сертификации в области пентеста.

Предложенная структура реализует принцип единства теории и практики, сформулированный С.Я. Батышевым как основополагающий для профессионального образования [20, с. 89], и обеспечивает формирование у студентов как технических компетенций, так и критического мышления, необходимого для работы с быстро развивающимися технологиями.

Заключение

Проведённый анализ позволяет сформулировать следующие основные выводы.

Применение LLM в автоматизации пентеста — перспективное, но методологически незрелое направление. Исследовательское поле сохраняет фрагментированный характер, большинство результатов получены на синтетических стендах, а проблемы воспроизводимости и контаминации обучающих данных (data contamination) существенно снижают достоверность публикуемых данных.

Реальная автономность LLM-систем пентеста не превышает уровней 3–4 по пятибалльной шкале. Модели не обладают способностью к самостоятельному целеполаганию и оценке последствий принимаемых решений. По состоянию на конец 2025 года LLM функционируют преимущественно в качестве аналитического инструмента — для агрегации данных, анализа технической документации, генерации аналитических гипотез и поддержки принятия решений специалистами уровня L1–L2.

Коммерческие решения обладают практическими преимуществами, однако сопряжены с существенными рисками в части прозрачности, верифицируемости результатов и обеспечения конфиденциальности данных клиентов.

Открытый проект CAI обладает наибольшим потенциалом с точки зрения академического применения: архитектура проекта обеспечивает прозрачность, воспроизводимость и возможность независимой верификации результатов.

Система профессионального образования в области информационной безопасности нуждается в адаптации к условиям LLM-эпохи. Задача подготовки специалистов включает формирование как технических компетенций работы с LLM-инструментами, так и навыков критического анализа их возможностей и ограничений. Базовыми педагогическими принципами при этом выступают деятельностный подход, проблемное обучение, симуляция реальных профессиональных условий, регулярное обновление содержания учебных программ.

Интеграция LLM в CI/CD-цепочки пентеста реализуема для решения узкоспециализированных задач. Полная автоматизация полного цикла тестирования на проникновение в обозримой перспективе недостижима. Квалифицированный специалист по информационной безопасности, способный к самостоятельному профессиональному суждению, сохраняет статус незаменимого участника процесса обеспечения безопасности информационных систем.

Перспективными направлениями дальнейших исследований являются: разработка стандартизированных методологий оценки LLM-систем применительно к задачам пентеста; создание открытых бенчмарков, устойчивых к проблеме контаминации обучающих данных; исследование педагогической эффективности различных подходов к интеграции LLM-инструментов в образовательные программы по кибербезопасности; разработка этических стандартов применения LLM в задачах наступательной безопасности.

Библиографический список

1. Новиков, А. М. Профессиональное образование России: перспективы развития / А. М. Новиков. – Москва: ИЦП НПО РАО, 1997. – 254 с.
2. Леонтьев, А. Н. Деятельность. Сознание. Личность / А. Н. Леонтьев. – Москва: Политиздат, 1975. – 304 с.
3. Лернер, И. Я. Проблемное обучение / И. Я. Лернер. – Москва: Знание, 1974. – 64 с.
4. Тихомиров, В. П. Цифровая трансформация образования: вызовы и возможности / В. П. Тихомиров, Н. В. Тихомирова // Открытое образование. – 2019. – № 4. – С. 4–14.
5. Alaryani, M. Towards Supporting Penetration Testing Education with Large Language Models: an Evaluation and Comparison / M. Alaryani [et al.] // arXiv preprint arXiv:2501.17539. – 2025.
6. Happe, A. Getting pwn'd by AI: Penetration Testing with Large Language Models / A. Happe, J. Cito // Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. – 2023. – P. 2082–2086.
7. Deng, G. PentestGPT: An LLM-empowered Automatic Penetration Testing Tool / G. Deng [et al.] // arXiv preprint arXiv:2308.06782. – 2023.
8. Xu, Z. AutoAttacker: A Large Language Model Guided System to Implement Automatic Cyberattacks / Z. Xu [et al.] // arXiv preprint arXiv:2403.01038. – 2024.
9. Fang, R. LLM Agents can Autonomously Exploit One-day Vulnerabilities / R. Fang [et al.] // arXiv preprint arXiv:2404.08144. – 2024.
10. Shao, R. Empirical Analysis of Large Language Models in Automated Vulnerability Discovery / R. Shao [et al.] // arXiv preprint arXiv:2312.02337. – 2023.
11. Bhatt, M. Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models / M. Bhatt [et al.] // arXiv preprint arXiv:2312.04724. – 2023.
12. Wan, Y. Cybersecurity Issues and Challenges: In Brief / Y. Wan [et al.] // IEEE Access. – 2024. – Vol. 12. – P. 1–15.
13. Moskal, S. LLM in the Shell: Generative AI-Powered Pentesting / S. Moskal [et al.] // arXiv preprint arXiv:2312.09552. – 2023.
14. Charan, P. V. S. From Text to MITRE Techniques: Examining the Layered Approach to Explainability With a Large Language Model (LLM) / P. V. S. Charan [et al.] // arXiv preprint arXiv:2308.09197. – 2023.
15. Tihanyi, N. CyberMetric: A Benchmark Dataset for Evaluating Large Language Models Knowledge in Cybersecurity / N. Tihanyi [et al.] // arXiv preprint arXiv:2402.07688. – 2024.
16. Peng, B. PenHeal: A Two-Stage LLM Framework for Automated Pentesting and Optimal Remediation / B. Peng [et al.] // arXiv preprint arXiv:2407.17788. – 2024.
17. Wan, Y. HackMentor: Fine-Tuning Large Language Models for Cybersecurity / Y. Wan [et al.] // arXiv preprint arXiv:2312.01234. – 2023.
18. Yang, J. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? / J. Yang [et al.] // arXiv preprint arXiv:2310.06770. – 2023.
19. Краевский, В. В. Методология педагогики: новый этап / В. В. Краевский. – Москва: Академия, 2006. – 400 с.
20. Батышев, С. Я. Профессиональная педагогика / С. Я. Батышев; под ред. С. Я. Батышева, А. М. Новикова. – 3-е изд. – Москва: ЭГВЕС, 2010. – 456 с.
21. Вербицкий, А. А. Активное обучение в высшей школе: контекстный подход / А. А. Вербицкий. – Москва: Высшая школа, 1991. – 207 с.
22. Сластёнин, В. А. Педагогика / В. А. Сластёнин. – Москва: Академия, 2002. – 576 с.
23. Зеер, Э. Ф. Психология профессионального образования / Э. Ф. Зеер. – Москва: Академия, 2009. – 384 с.
24. Alias Robotics. Cybersecurity AI (CAI): An Open, Bug Bounty-Ready Agentic Framework // arXiv preprint arXiv:2504.06017. – 2025.

References

1. Novikov, A. M. (1997). Professional education in Russia: Development prospects. ICP NPO RAO. (In Russian)
2. Leont'ev, A. N. (1975). Activity. Consciousness. Personality. Politizdat. (In Russian)
3. Lerner, I. Ya. (1974). Problem-based learning. Znanie. (In Russian)
4. Tikhomirov, V. P., & Tikhomirova, N. V. (2019). Digital transformation of education: Challenges and opportunities. Open Education, (4), 4–14. (In Russian)
5. Alaryani, M., et al. (2025). Towards supporting penetration testing education with large language models: An evaluation and comparison. arXiv preprint arXiv:2501.17539.
6. Happe, A., & Cito, J. (2023). Getting pwn'd by AI: Penetration testing with large language models. In Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (pp. 2082–2086).
7. Deng, G., et al. (2023). PentestGPT: An LLM-empowered automatic penetration testing tool. arXiv preprint arXiv:2308.06782.
8. Xu, Z., et al. (2024). AutoAttacker: A large language model guided system to implement automatic cyber-attacks. arXiv preprint arXiv:2403.01038.
9. Fang, R., et al. (2024). LLM agents can autonomously exploit one-day vulnerabilities. arXiv preprint arXiv:2404.08144.
10. Shao, R., et al. (2023). Empirical analysis of large language models in automated vulnerability discovery. arXiv preprint arXiv:2312.02337.
11. Bhatt, M., et al. (2023). Purple Llama CyberSecEval: A secure coding benchmark for language models. arXiv preprint arXiv:2312.04724.
12. Wan, Y., et al. (2024). Cybersecurity issues and challenges: In brief. IEEE Access, 12, 1–15.
13. Moskal, S., et al. (2023). LLM in the shell: Generative AI-powered pentesting. arXiv preprint arXiv:2312.09552.
14. Charan, P. V. S., et al. (2023). From text to MITRE techniques: Examining the layered approach to explainability with a large language model (LLM). arXiv preprint arXiv:2308.09197.
15. Tihanyi, N., et al. (2024). CyberMetric: A benchmark dataset for evaluating large language models knowledge in cybersecurity. arXiv preprint arXiv:2402.07688.
16. Peng, B., et al. (2024). PenHeal: A two-stage LLM framework for automated pentesting and optimal remediation. arXiv preprint arXiv:2407.17788.
17. Wan, Y., et al. (2023). HackMentor: Fine-tuning large language models for cybersecurity. arXiv preprint arXiv:2312.01234.
18. Yang, J., et al. (2023). SWE-bench: Can language models resolve real-world GitHub issues? arXiv preprint arXiv:2310.06770.
19. Kraevskiy, V. V. (2006). Methodology of pedagogy: A new stage. Akademiya. (In Russian)
20. Batyshev, S. Ya. (2010). Professional pedagogy (S. Ya. Batyshev & A. M. Novikov, Eds.; 3rd ed.). EGVES. (In Russian)
21. Verbitskiy, A. A. (1991). Active learning in higher education: A contextual approach. Vysshaya Shkola. (In Russian)
22. Slastenin, V. A. (2002). Pedagogy. Akademiya. (In Russian)
23. Zeer, E. F. (2009). Psychology of professional education. Akademiya. (In Russian)
24. Alias Robotics. (2025). Cybersecurity AI (CAI): An open, bug bounty-ready agentic framework. arXiv preprint arXiv:2504.06017.

Информация об авторе

Греков Владимир Сергеевич — аспирант 1 курса кафедры безопасности информационных технологий РГУ нефти и газа (НИУ) имени И.М. Губкина. Область научных интересов: применение методов искусственного интеллекта в задачах кибербезопасности, автоматизация тестирования на проникновение, большие языковые модели. E-mail: grekov.v@gubkin.ru

LARGE LANGUAGE MODELS IN PENETRATION TESTING AUTOMATION: CURRENT STATE, LIMITATIONS AND PROSPECTS FOR APPLICATION IN PROFESSIONAL TRAINING OF INFORMATION SECURITY SPECIALISTS

Grekov V. S.[1]

[1]National University of Oil and Gas «Gubkin University», Moscow, Russian Federation

Abstract. *The article analyzes the current state of research on the application of large language models (LLMs) for penetration testing automation. The reasons for the rapid growth of interest in this area are examined, driven by the increasing complexity of IT infrastructures, the high cost of traditional security audit methods, and the development of the DevSecOps concept. Based on the systematization of review and applied research from 2023–2025, four key areas of LLM application in cybersecurity are identified: decision support and strategic attack planning, automation of penetration testing stages, security of language models themselves (AI Red Teaming), and intelligent source code analysis. A critical analysis is conducted of multi-agent system architectures using Retrieval-Augmented Generation (RAG) approaches, testing and evaluation systems, and commercial solutions. It is established that even the most advanced LLM-based penetration testing systems operate at intermediate autonomy levels (3–4 out of 5), unable to independently formulate goals and evaluate the consequences of their decisions. The thesis is substantiated that as of the end of 2025, LLMs serve primarily as an analyst tool rather than an autonomous penetration tester. The pedagogical aspects of integrating LLM tools into professional training of information security specialists are examined, and the necessity of developing critical thinking among students regarding the capabilities and limitations of these technologies is justified.*

Keywords: *large language models, penetration testing, cybersecurity, automation, multi-agent systems, RAG, professional training.*

Information about the author

Grekov Vladimir Sergeevich — 1st year postgraduate student, Department of Information Technology Security, Gubkin Russian State University of Oil and Gas (National Research University). Research interests: application of artificial intelligence methods in cybersecurity, penetration testing automation, large language models. E-mail: grekov.v@gubkin.ru