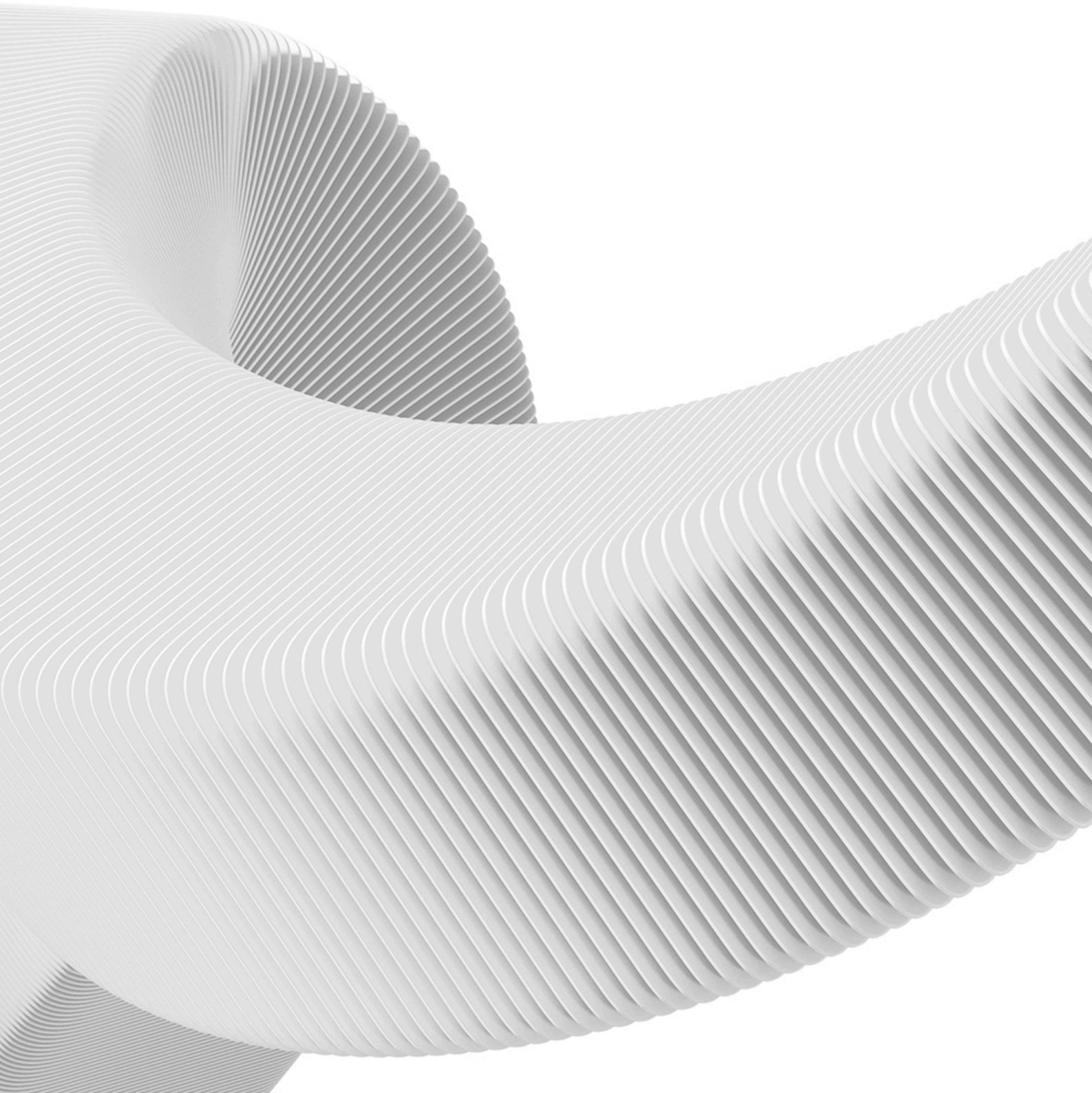
ИИ в 2026

Угроза снаружи и внутри



ОГЛАВЛЕНИЕ

ОБ ИССЛЕДОВАНИИ

Исследование рассказывает об эволюции и развитии угроз, связанных с применением искусственного интеллекта в кибератаках, и является продолжением серии публикаций о технологии, в которой мы рассматривали применение ИИ в киберзащите и разбирали ключевые тренды в ее развитии. Этот отчет основан на собственной экспертизе компании Positive Technologies, результатах расследований, а также на данных авторитетных источников.

По нашим оценкам, большинство кибератак не предается огласке из-за репутационных рисков. В связи с этим подсчитать точное число угроз не представляется возможным даже для организаций, занимающихся расследованием инцидентов и анализом действий хакерских группировок. Наше исследование проводится с целью обратить внимание компаний, специалистов по защите информации и всех людей, интересующихся современным состоянием информационной безопасности, на наиболее актуальные угрозы и методы защиты от них, появляющиеся и развивающиеся под влиянием искусственного интеллекта.

В рамках отчета каждая массовая атака, в ходе которой злоумышленники проводят, например, фишинговую рассылку на разные адреса, рассматривается как одна атака, а не множество разных. Термины, которые мы используем в исследовании, приведены в словаре на сайте Positive Technologies.

РЕЗЮМЕ

- Развитие и внедрение технологий искусственного интеллекта является одним из ключевых факторов изменения киберландшафта, который влияет на все основные применяемые киберпреступниками методы атак.
- Полностью автономные кибератаки с использованием ИИ пока остаются лишь гипотетической возможностью. Но в тоже время масштабы внедрения технологии в киберпреступную активность продолжают расти.
- Сегодня наибольшую выгоду от применения технологии получают профессиональные APT-группировки и подготовленные злоумышленники, способные применить сильные стороны ИИ для автоматизации и ускорения подготовки отдельных шагов атаки.
- Основой наступательного ИИ-арсенала являются не разработки злоумышленников, а легальные модели и инструменты с отсутствующими или легко обходимыми ограничениями безопасности.
- ИИ ускоряет поиск уязвимостей, сокращает время между раскрытием недостатка безопасности и появлением для него эксплойта.
- Генерация фишингового контента, дипфейков и фрагментов вредоносного кода с помощью ИИ становится массовой практикой.
- Киберпреступники начали эксперименты с внедрением ИИ-модулей в ВПО. Встречаются как образцы, загружающие модель для распознавания интерфейса непосредственно на устройство жертвы, так и обращающиеся к ИИ-сервису удаленно для генерации команд.
- Угрозы, исходящие от применения искуственного интеллекта, сегодня связаны не только с наступательным ИИ, но и с небезопасным внедрением технологии в бизнес-процессы. Shadow AI, AI-driven разработка и агентские системы формируют новую высокорисковую поверхность атаки.
- Shadow AI — использование сотрудниками неконтролируемых ИИ-инструментов может приводить к сложным в обнаружении инцидентам с утечками данных.
- AI-driven разработка масштабирует распространение небезопасного кода, Рост синтаксического качества генерации сопровождается сохраняющимися проблемами в задачах безопасности.

- Инфраструктура ИИ-инструментов становится целью кибератак для получения первоначального доступа, похищения данных и несанкционированного использования вычислительных ресурсов и тарифов использования моделей.

- ИИ-агенты формируют новый класс рисков безопасности, поскольку способны самостоятельно инициировать действия, взаимодействовать с внутренними и внешними сервисами. Ошибки в работе агентов уже стали причиной ряда инцидентов безопасности, а киберпреступники начали проводить атаки на цепочки поставок агентов.

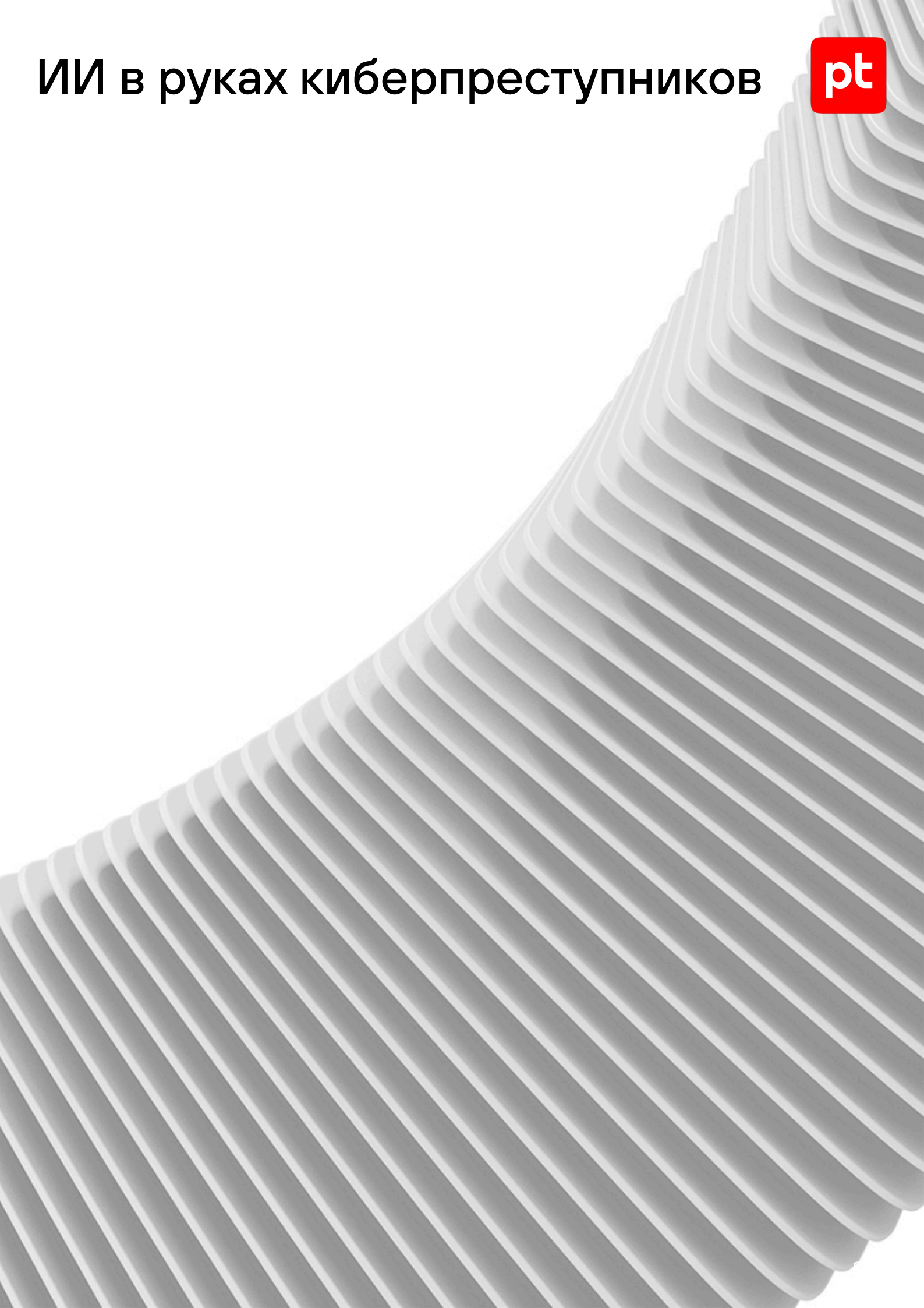
- Большинство успешных атак с применением ИИ становятся возможными из-за классических недостатков безопасности: слабых учетных данных, открытых сервисов, уязвимого, устаревшего ПО и игнорирования принципов безопасной разработки.

ВВЕДЕНИЕ

Искусственный интеллект является одним из ключевых драйверов изменения киберландшафта в 2026 году, он воздействует на все три основных метода кибератак: социальную инженерию, эксплуатацию уязвимостей и применение вредоносного программного обеспечения. Потенциал применения ИИ для автоматизации, масштабирования и усложнения атак остается огромным, но при этом реальный эффект остается пока ограниченным. Влияние ИИ на кибератаки усиливается не революционными, а постепенными, эволюционными темпами и будет продолжаться в обозримом будущем.

Вторая часть аналитики посвящена масштабной угрозе, которая активно проявляется из-за повсеместного, недостаточно зрелого внедрения ИИ без обеспечения защищенности как самой технологии, так и инфраструктуры. Угроза уже перестала быть гипотетической: инциденты, связанные с ИИ, внедренным в процессы компаний, приводят к нарушениям в работе и утечкам конфиденциальных данных; в ближайшие годы число таких нарушений будет только возрастать, а вопросы безопасности новых решений выйдут на первый план.

ИИ в руках киберпреступников

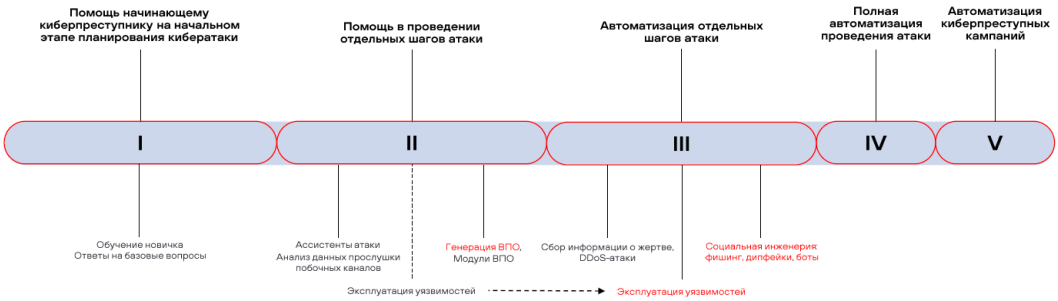


УРОВНИ ПРИМЕНЕНИЯ ИИ В КИБЕРАТАКАХ

В исследовании, посвященном применению ИИ в кибератаках, мы разбирали пять уровней внедрения технологии в процесс подготовки и реализации атаки. С 2024 года основные области и принципы использования искусственного интеллекта киберпреступниками не претерпели кардинальных изменений. Основные перемены касаются значительного развития и масштабирования применения доказавших свою эффективность техник: к примеру, генерация кода для ВПО и фишингового контента становятся массовыми явлениями, а в отдельных атаках задачи поиска и эксплуатации уязвимостей ложатся на автоматизированные решения с ИИ.

Несмотря на рост масштабов эксплуатации технологии для отдельных киберпреступных действий, полная автоматизация атаки остается недоступной даже для наиболее продвинутых моделей. Исследования показывают, что несмотря на постоянный рост возможностей, лишь в отдельных случаях, в условиях, приближенных к настоящим, ИИ-инструменты могут пройти дальше этапа эксплуатации уязвимостей веб-ресурсов компании. Результаты экспериментов показывают, что для реализации атак пока недостаточно только моделей и агентов, необходим человек, управляющий ими и проводящий не автоматизируемые этапы атаки.

Рисунок 1. Уровни применения ИИ в кибератаках. Красным отмечены методы с массовым применением технологии



На теневых площадках начинает развиваться применение ИИ для обработки похищенных данных, то есть в сегменте действий после атаки. Тем не менее эксплуатация ИИ обсуждается теневым сообществом не так активно: для киберпреступников внедрение ИИ остается экспериментальным, и большая часть обсуждений все еще посвящена поиску актуальных нецензурированных LLM либо моделей, ограничения которых легко обойти. Реально применяющие возможности ИИ решения и инструменты встречаются нечасто, намного реже мошеннических вариантов, в которых ИИ выступает лишь яркой, привлекающей внимание вывеской.

КТО ПРИМЕНЯЕТ ИИ В КИБЕРАТАКАХ

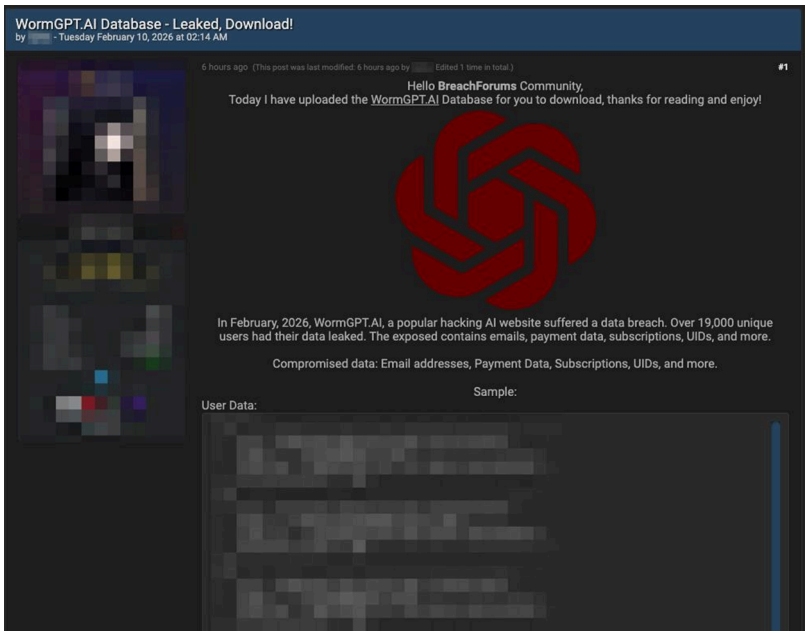
Одно из ключевых опасений, высказываемых в связи с наступательным потенциалом ИИ, — резкое повышение возможностей множества низкоквалифицированных киберпреступников. Сегодня мы наблюдаем, что наибольшую пользу от злонамеренной эксплуатации технологии получают не скрипт-кидди, а APT-группировки. Именно кибершпионы идут на острие внедрения ИИ, в их атаках исследователи обнаруживают не только сгенерированный код или фишинговый контент, но и полноценные вредоносные модули, решения и выстроенные с применением технологии процессы. Наступательные «успехи» ИИ реализуются подготовленными преступниками, которые понимают принципы и нюансы реализуемых ими действий, и не только эксплуатируют сильные черты искусственного интеллекта, но и справляются со слабыми сторонами технологии.

В ближайшие годы такое положение вещей сохранится, ИИ не даст возможность новичку сразу проводить сложные атаки. Но при этом способность внедрить искусственный интеллект в процесс подготовки и реализации атаки станет одним из новых мерил квалификации киберпреступника или группировки.

КАКОЙ ИИ ИСПОЛЬЗУЮТ В КИБЕРАТАКАХ

В 2023–2024 годах на теневых площадках появился целый ряд киберпреступных LLM, якобы обученных непосредственно для помощи в подготовке и проведении кибератак. На поверку они оказывались либо мошенничеством, либо, в лучшем случае, легальными моделями с встроенным автоматизированным jailbreak, позволяющим получить ответ на любые вопросы, в обход установленных ограничений. Яркий пример такой LLM — WormGPT. Ее появление в 2023 году активно обсуждалось на теневых площадках и широко освещалось в СМИ. Мы уже рассказывали, что реальные отзывы на этот инструмент говорили о его низкой эффективности, проект быстро закрылся, а будущие многочисленные клоны и новые версии уже не смогли привлечь столько же внимания. В феврале 2026 года произошла утечка, позволяющая глубже взглянуть на внутреннее устройство проекта. Анализ данных утечки показывает, что большинства заявленных инструментов и возможностей, якобы доступных WormGPT, например Nmap и DNS-сканирование, фактически не существует, возможности проекта ограничиваются только текстовыми ответами. Оказалось, что это языковая модель Mistral-7B с RAG-датасетом из выжимок новостей о кибератаках с профильных сайтов и системным промптом, который сообщает модели, что она ничем не ограничена и не участвует в злонамеренных действиях. Интересно, что среди более чем 4000 попавших в утечку учетных записей лишь 200 были подписаны на один из планов.

Рисунок 2. Продажа утечки WormGPT



На теневых площадках продолжают появляться предложения по продаже доступа к неограниченным LLM. Мы ожидаем, что и в будущем они окончательно не исчезнут, независимо от провалов мошеннических проектов; подстегивать интерес к ним может недоверие к легальным LLM и опасение деанонимизации. Важно отметить, что страх перед раскрытием разработчиками легальных моделей не является беспочвенным. OpenAI и другие компании регулярно сообщают об использовании их моделей конкретными акторами. Впрочем, как показывает опыт утечки WormGPT, создаваемые киберпреступниками для киберпреступников LLM не являются гарантированно безопасными для пользователей.

Рисунок 3. Предостережение от использования моделей из-за потенциальной деанонимизации

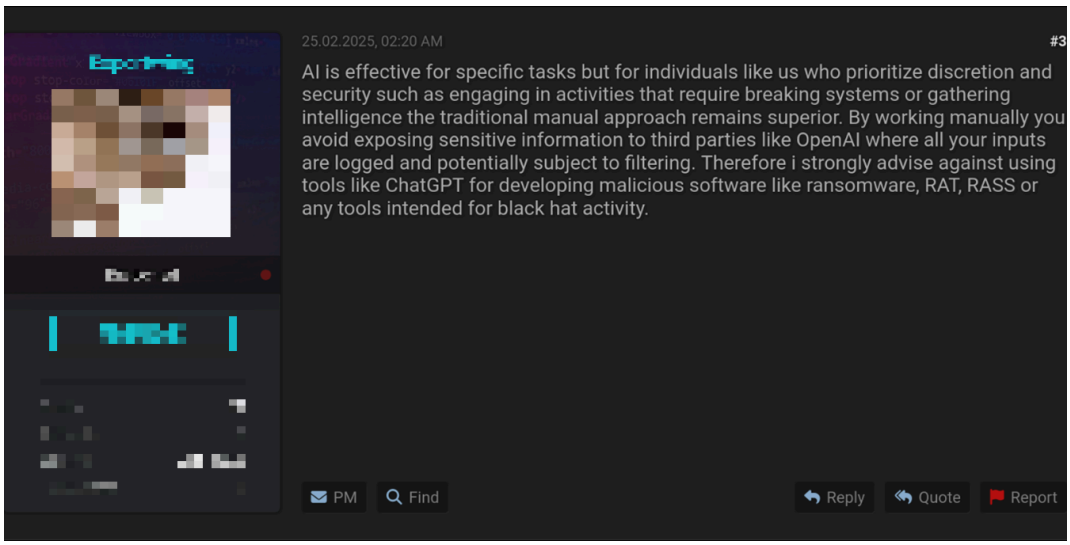
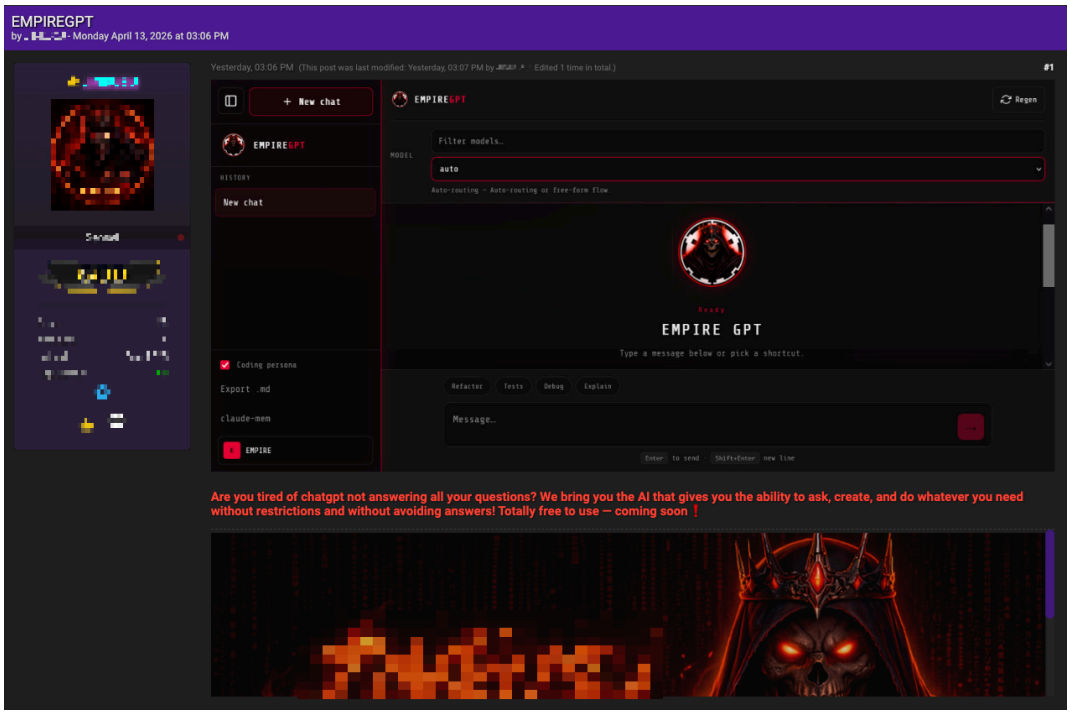


Рисунок 4. Реклама GPT без ограничений в феврале 2026

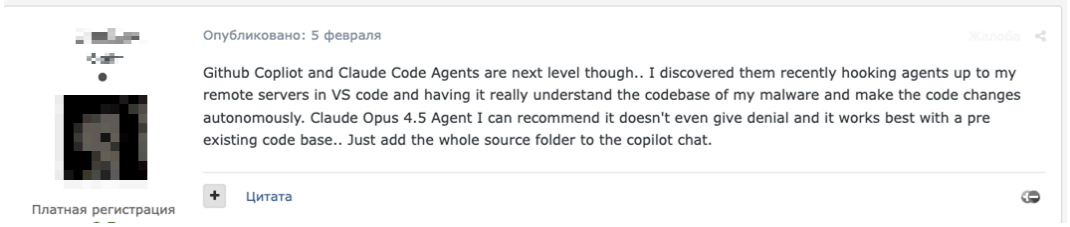
Рисунок 5. Реклама GPT без ограничений в апреле 2026



Постепенно предложения с собственными вредоносными LLM окончательно отойдут на второй план, уступив место формирующейся сегодня категории услуг jailbreak as a service, объединяющей методы обхода ограничений легальных моделей. Коммерческие модели обладают уровнем качества ответов, которого киберпреступникам невозможно добиться без сопоставимых с передовыми компаниями-разработчиками ресурсов; при этом обойти ограничения не слишком сложно, несмотря на усилия разработчиков. Как мы и прогнозировали, прогресс в развитии ИИ-инструментов позволяет прогрессировать и преступному применению технологии.

В ближайшие годы легальные модели останутся фундаментом киберпреступного ИИ-арсенала. Измениться это может только в том случае, если радикально упростятся и подешевеют методы качественного обучения собственных моделей, из-за чего злоумышленники перейдут на собственные решения. Далее в исследовании будет приведено множество различных примеров использования злоумышленниками именно легальных моделей, таких как ChatGPT или Claude Code.

Рисунок 6. Обсуждение на теневой площадке моделей, применяемых для написания вредоносного кода



ТЕПЛОВАЯ МАТРИЦА MITRE ATT&CK: ИИ СЕГОДНЯ, ЗАВТРА, В БУДУЩЕМ

Для оценки реальных текущих возможностей и перспектив применения ИИ в кибератаках мы анализируем матрицу MITRE ATT&CK (у нее также есть и русскоязычная версия).

Матрица MITRE ATT&CK — это база знаний, поддерживаемая корпорацией MITRE и разработанная на основе анализа реальных APT-атак. Матрица описывает тактики и техники, которыми злоумышленники пользуются в атаках на корпоративную инфраструктуру.

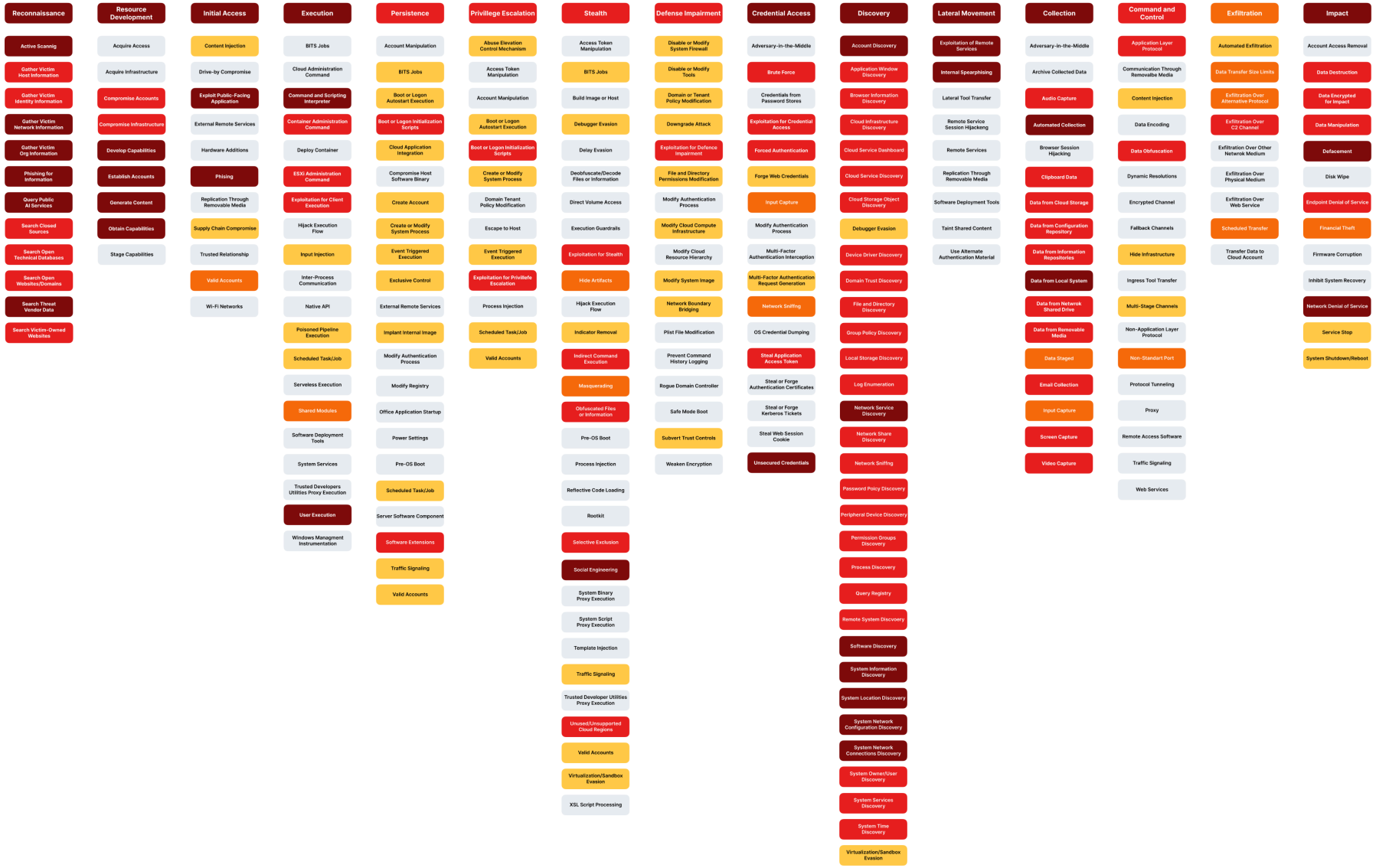
Тактики — столбцы матрицы — описывают цель киберпреступников и делят атаку на этапы. Например, тактика «Первоначальный доступ» (TA0001: Initial Access) описывает действия, с помощью которых злоумышленник пытается проникнуть в сеть и получить плацдарм для следующих шагов.

Техники — элементы столбцов — описывают конкретные действия, которые киберпреступники реализуют для достижения цели. Например, тактика «Первоначальный доступ» включает технику «Распространение через съемные носители» (T1091: Replication Through Removable Media). В этой технике описывается, как злоумышленники могут проникать в отключенные от сети системы с помощью зараженных носителей, например флеш-накопителей USB.

Мы обновили тепловую карту из предыдущего исследования, сохранив прежнюю методологию для оценки, как скоро киберпреступники смогут применить ИИ для решения задач каждой тактики, техники и подтехники. Применение ИИ в кибератаках сохраняет потенциально очень широкое поле: согласно нашему анализу в 100% тактик MITRE ATT&CK и больше чем в половине техник (62%) в будущем может найтись место для применения технологий искусственного интеллекта.

Рисунок 7. Тепловая матрица MITRE ATT&CK

Полная версия с поддтехниками доступна по ссылке



Все техники и подтехники мы разделяем на пять уровней по потенциальному времени появления в них технологий ИИ.

Бордовый — уже известны случаи применения. На момент написания исследования киберпреступники хотя бы раз применяли ИИ в реальных атаках для каждой десятой (10%) техники. За последние два года доля таких техник выросла вдвое : в 2024 году их было только 5%. Кроме того, явно характеризует эволюционное развитие применения ИИ в кибератаках и изменение самой матрицы, дополняющейся связанными непосредственно с ИИ техниками, например в марте 2026 года появилась T1683 Generate Content, в описании которой существенное внимание отведено созданию контента именно с помощью ИИ.

Красный — может применяться в ближайшее время. Каждая четвертая (26%) техника относится к «красной» категории, в которую входят техники, для которых уже доказана применимость ИИ, опубликованы подтверждающие исследования и известны proof of concept. Примером техники, в которой применимость ИИ была недавно доказана, может послужить T1567 Exfiltration Over Web Service.

Оранжевый — может применяться в обозримом будущем. Для 6% техник перед интеграцией ИИ киберпреступникам придется решить ряд серьезных задач. Несмотря на сложности, киберпреступники продолжают развивать арсенал инструментов, в том числе интегрируя в них все больше технологий искусственного интеллекта, что подтверждается постепенным переходом техник из этой категории в «красную» и «бордовую». Нет сомнений, что в будущем тенденция сохранится, и постепенно все техники, в которых внедрение ИИ будет приносить пользу, будут как минимум опробованы злоумышленниками в кибератаках.

Желтый — теоретически применение возможно, но в обозримом будущем практически недостижимо. К «желтой» категории относится 19% техник. В их использование можно интегрировать ИИ, но пока что это остается недостижимым в рамках кибератаки. Важно учитывать, что технологии ИИ не стоят на месте. Исследователи делают различные прогнозы и предположения о сроках достижения новых вех развития ИИ и открывающихся возможностях. Если технологии ИИ сделают качественный шаг вперед, киберпреступники обязательно попытаются усложнить, автоматизировать и масштабировать атаки.

Серый — применение ИИ не оправдано или не принесет существенной пользы. Технологии искусственного интеллекта могут применяться в кибератаках самыми разными способами, тем не менее есть множество вариантов атак, в которых ИИ не нужен или вовсе не применим на текущем уровне развития.

ЭКСПЛУАТАЦИЯ УЯЗВИМОСТЕЙ

За последние два года возможности ИИ для поиска и эксплуатации уязвимостей значительно выросли, в 2024 году мы рассказывали в первую очередь об исследовательских и экспериментальных результатах. Сегодня же различные ИИ-инструменты стали массово использовать при тестировании на проникновение легальные специалисты, участники CTF-соревнований и багхантеры. Кроме того, известны случаи применения технологии киберпреступниками в реальных атаках для поиска и эксплуатации недостатков защиты и уязвимостей.

В задачах поиска известных уязвимостей актуальные ИИ-решения показывают неоднородные результаты для различных типов уязвимостей, наибольших успехов добиваясь в поиске недостатков в API и веб-приложениях. Масштабы возможного применения ИИ для поиска уязвимостей и недостатков безопасности показывает переполнение багбаунти платформы в последние полгода ИИ-отчетами. Далеко не все из этих отчетов содержат действительно полезные сведения о киберугрозах, что создает проблему для команд, вынужденных изучать резко возросшие объемы ИИ-мусора. Так, в начале 2026 года было объявлено о завершении багбаунти программы curl именно из-за вала некачественных ИИ-отчетов. Интересно, что Standoff Bug Bounty пока не столкнулся с такой проблемой, хотя большинство багхантеров экспериментируют или уже активно применяют ИИ в работе.

Как мы и предполагали, важную роль в пополнении инструментария киберпреступников играют не только модели общего назначения, но и открытые наступательные фреймворки с ИИ, например HexStrike AI и CyberStrikeAI. Платформа HexStrike AI для управления более чем 150 ИИ-агентами для сканирования, эксплуатации уязвимостей и закрепления активно обсуждается киберпреступниками на теневых площадках и получает положительные отзывы.

Поиск уязвимостей с помощью ИИ применяется не только белыми хакерами, но и реальными киберпреступниками для проведения атак, в том числе на государственные цели. В конце 2025 — начале 2026 года киберпреступники, убеждая Claude Code и GPT-4.1 в том, что проводят пентест, смогли провести кибератаку на целый ряд правительственных учреждений Мексики и похитить сотни миллионов записей персональных и других конфиденциальных данных. Этот случай — интересный пример широкого применения ИИ на всем протяжении подготовки кибератаки: для сбора информации, поиска уязвимостей и разработки для них эксплойтов. Благодаря применению моделей киберпреступники смогли значительно ускорить все процессы, и нет сомнений, что это только начало. Важно отметить, что эксплуатируемые уязвимости могли быть устранены стандартными практиками защиты, такими как своевременное обновление ПО, использование надежных учетных данных, сегментация сети и реализация обнаружения угроз на конечных устройствах.

На теневых площадках продолжают появляться предложения по продаже доступа к неограниченным LLM. Мы ожидаем, что и в будущем они окончательно не исчезнут, независимо от провалов мошеннических проектов; подстегивать интерес к ним может недоверие к легальным LLM и опасение деанонимизации. Важно отметить, что страх перед раскрытием разработчиками легальных моделей не является беспочвенным. OpenAI и другие компании регулярно сообщают об использовании их моделей конкретными акторами. Впрочем, как показывает опыт утечки WormGPT, создаваемые киберпреступниками для киберпреступников LLM не являются гарантированно безопасными для пользователей.

Рисунок 8. Отзывы про использование HexStrike AI

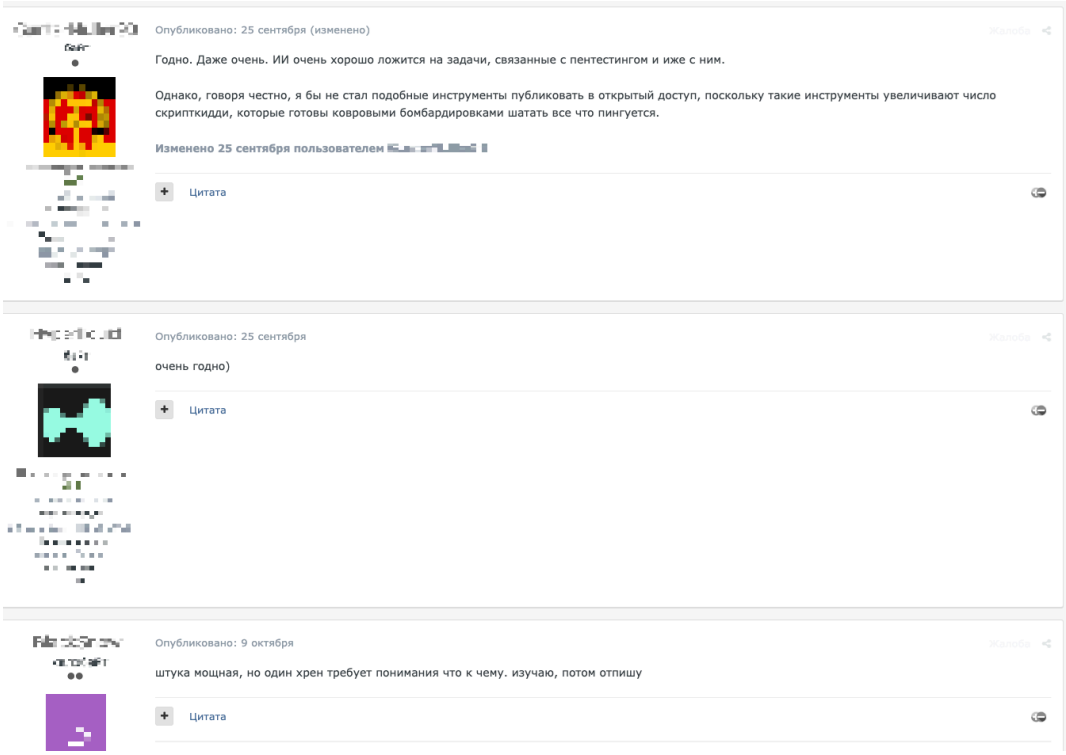
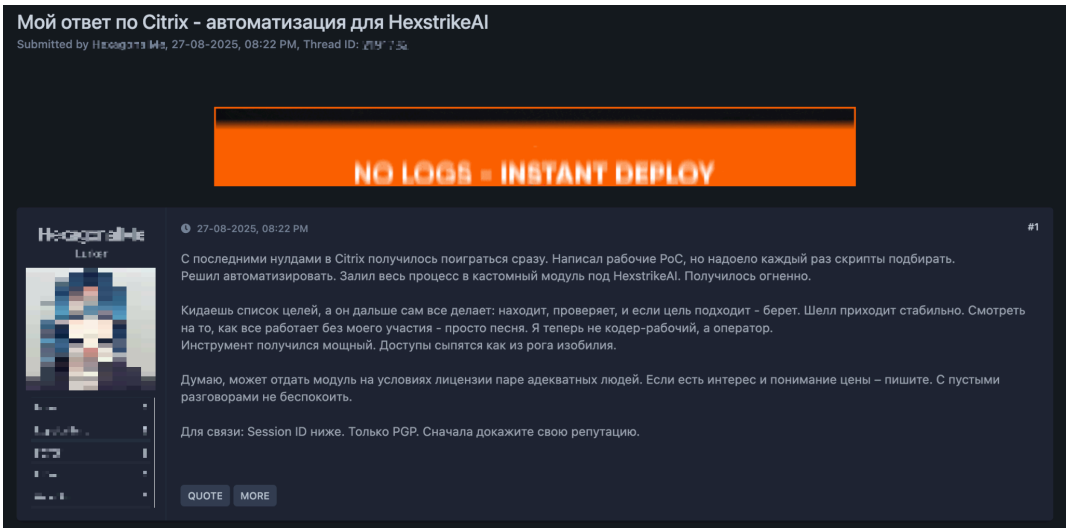


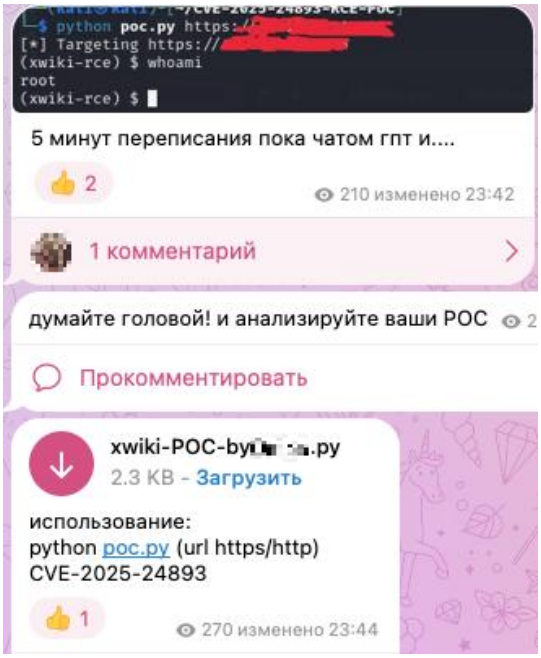
Рисунок 9. Отзыв про использование HexStrike AI



В тоже время CyberStrikeAI уже использовался в известной успешной кибератаке: именно с его помощью финансово мотивированный киберпреступник скомпрометировал более 600 устройств FortiGate. Важно отметить, что взлом был возможен из-за таких распространенных недостатков безопасности, как открытые порты и слабые учетные данные, а не благодаря эксплуатации уязвимостей. Исследователи Amazon Threat Intelligence отмечают, что киберпреступник обладал средней квалификацией, и тем не менее реализовал массовую атаку на незащищенные устройства, обнаруживая их с помощью ИИ. Инцидент показывает, как важно исправлять базовые недостатки защиты, поскольку с ростом автоматизации увеличивается шанс их обнаружения и эксплуатации. Потенциально угроза автоматизированного поиска уязвимостей и последующей эксплуатации в атаках низкой сложности затрагивает большую часть компаний. К примеру, как мы отмечали в исследовании, проблемы с реализацией парольной политики встречались в 97% тестов на проникновение российских компаний, а устаревшие версии ПО на периметре с известными уязвимостями — у 80%.

Возможность использования ИИ не только для обнаружения известных недостатков безопасности, но и для поиска новых, ранее неизвестных уязвимостей ведет к росту количества раскрываемых угроз, которое, по данным NIST, выросло на 263% за последние 5 лет. Чем активнее ИИ будет применяться для поиска новых уязвимостей, тем больше увеличится нагрузка на базы знаний об уязвимостях, — это может привести к все большей их фрагментации и специализации на ПО, важном и характерном для отдельного региона. При таких условиях вырастет важность отслеживания различных источников — как всемирных (таких как CVE), так и локальных, (например, BDU и PT). Применение ИИ для создания PoC и разработки эксплойтов сокращает время между обнаружением уязвимости и появлением для нее эксплойта, а следовательно, сокращает временное окно между публикацией информации об уязвимости и появлением у киберпреступников инструментов ее эксплуатации. Из-за роста числа обнаруживаемых уязвимостей и скорости появления эксплойтов организации могут столкнуться с невозможностью оперативно устранять все угрозы. В таких обстоятельствах вырастет важность приоритизации активов, управления уязвимостями, построенном на понимании инфраструктуры и знании критических компонентов, компрометация которых может приводить к недопустимым событиям.

Рисунок 10. Сообщения киберпреступника о генерации PoC для уязвимостей с помощью ChatGPT

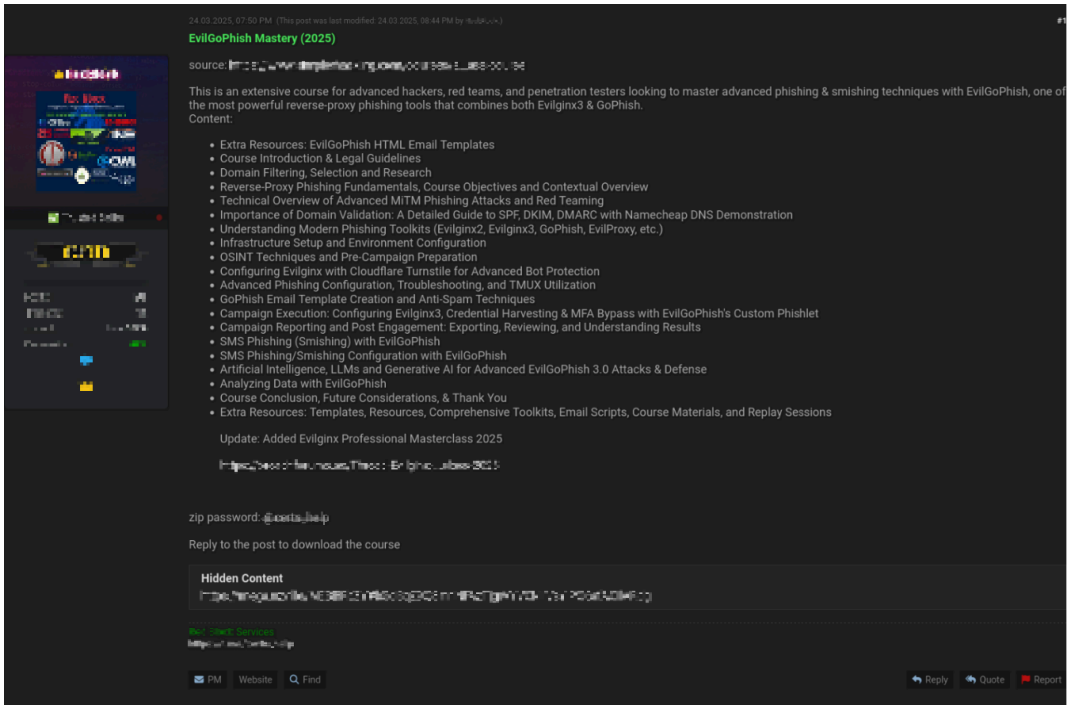


Разработчики актуальных на момент публикации исследования больших языковых моделей, например Claude Mythos и Chat GPT 5.4 Cyber, делают особый акцент на возможности их продуктов не только обнаруживать уязвимости нулевого дня, но и эксплуатировать их. Развитие таких возможностей значительно увеличивает важность тщательной проверки защищенности и всестороннего тестирования безопасности продуктов и систем перед релизом, в том числе с использованием ИИ-инструментов. Несмотря на то что модели с наиболее продвинутыми наступательными возможностями выпускаются для ограниченного круга доверенных команд экспертов (из соображений безопасности), они все равно будут применяться для атак: в эпоху, когда политические мотивы оказывают существенное влияние на ландшафт киберугроз, а атаки в цифровом пространстве становятся частью кинетических ударов, отказ от наступательных возможностей ИИ из соображений морали кажется маловероятным, тем более что технология официально внедряется силовыми ведомствами, например в США. Помимо вероятного использования наиболее продвинутых моделей государственными разведками, контролируемый доступ не может гарантировать, что инструменты не будут распространяться. В день ограниченного релиза Claude Mythos группа энтузиастов смогла получить доступ к закрытой модели и использовать ее, не привлекая внимания. Инцидент, вкупе с другими случаями утечек и несанкционированных доступов к моделям, доказывает, что и в будущем наиболее мощные, ограниченные в распространении инструменты могут попадать в руки злоумышленников.

СОЦИАЛЬНАЯ ИНЖЕНЕРИЯ

Генерация текста и контента для атак социальной инженерии всегда была одним из самых развитых применений ИИ для кибератак. Сегодня происходит рост внедрения ИИ в атаки социальной инженерии вследствие общего улучшения качества генерации. По данным Microsoft, в 54% случае получатель открывает ссылку в сгенерированном фишинговом письме в 4,5 раза чаще, чем в написанном людьми. Генерация убедительного фишинга с помощью моделей позволяет замещать труд человека и сокращать необходимое для подготовки атаки время, а в сочетании с эксплуатацией агентов, возможно выстраивание полного цикла автоматизированной атаки: поиска и сбора информации о жертве, формирования индивидуального сценария обманной коммуникации и даже поддержания длительного и убедительного диалога. Таким образом, применение ИИ потенциально позволяет не просто масштабировать атаки, но и дробить фишинговые кампании на множество целевых атак за счет адаптации каждого письма, сообщения и диалога под атакуемую жертву — по сути, превращать массовые рассылки в множество атак типа spear phisihing. Возможность проводить такие атаки на уровне специалиста-человека с помощью специально обученного ИИ-агента уже продемонстрировали исследователи из HoxHunt в 2025 году. Эти возможности — не просто теория: генерация адаптированных фишинговых сообщений с подстройкой под жертв с помощью ИИ уже фиксировалась в кибератаке APT-35, и в будущем примеров таких массово-целевых атак с применением ИИ будет становится больше. Стоит добавить, что применение ИИ стало обязательной частью «обучения» фишингу.

Рисунок 11. Распространяемый на теневой площадке курс «обучения фишингу»

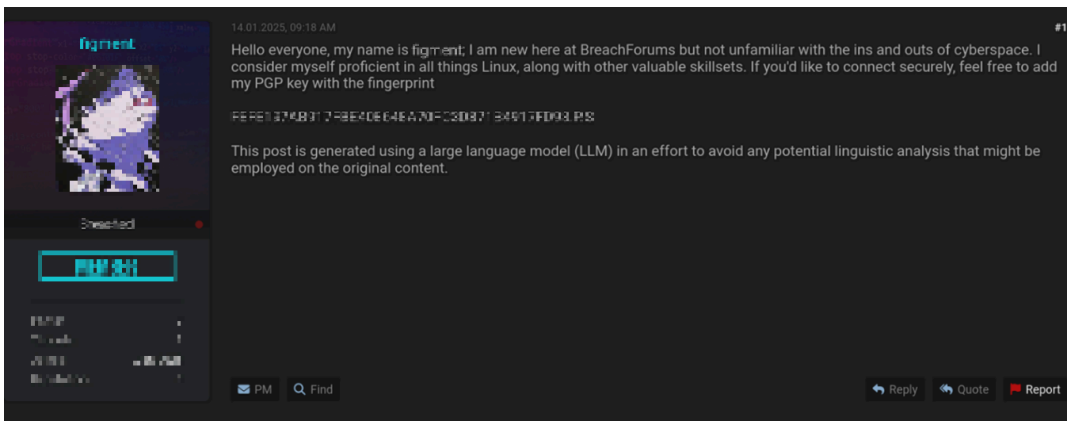


Назвать точное количество фишинговых писем, сгенерированных с помощью ИИ, крайне сложно, поскольку хороший результат работы модели не будет отличаться от образца, созданного человеком. Тем не менее, по оценкам HoxHunt, доля таких писем в 2025 году составляла около 4%, а в период новогодних и рождественских праздников резко выросла до 40–50%. В ближайшем будущем доля сгенерированных ИИ фишинговых писем будет колебаться в зависимости от активности массовых кампаний, но в среднем продолжит постепенный рост.

Кроме того, в начале 2026 года исследователи продемонстрировали, что навыки перевода ChatGPT сравнялись с людьми-переводчиками начального и среднего уровня. Несмотря на то что профессиональные переводчики все еще лучше справляются с передачей языковых тонкостей, способностей моделей достаточно для перевода фишингового текста на новые языки, а значит и потенциального расширения деятельности киберпреступников в новых регионвх. К примеру, согласно отчету CrowdStrike, группировка Renaissance Spider уже использовала ИИ для перевода приманок в ClickFix-атаках, а Google сообщала о применении LLM Gemini группировкой UNC1069 для генерации контента на нетипичном для киберпреступников языке. В ближайшие годы мы можем увидеть расширение ареала действий группировок на новые регионы, и более многочисленные списки участников киберконфликтов, разворачивающихся вокруг политически напряженных регионов.

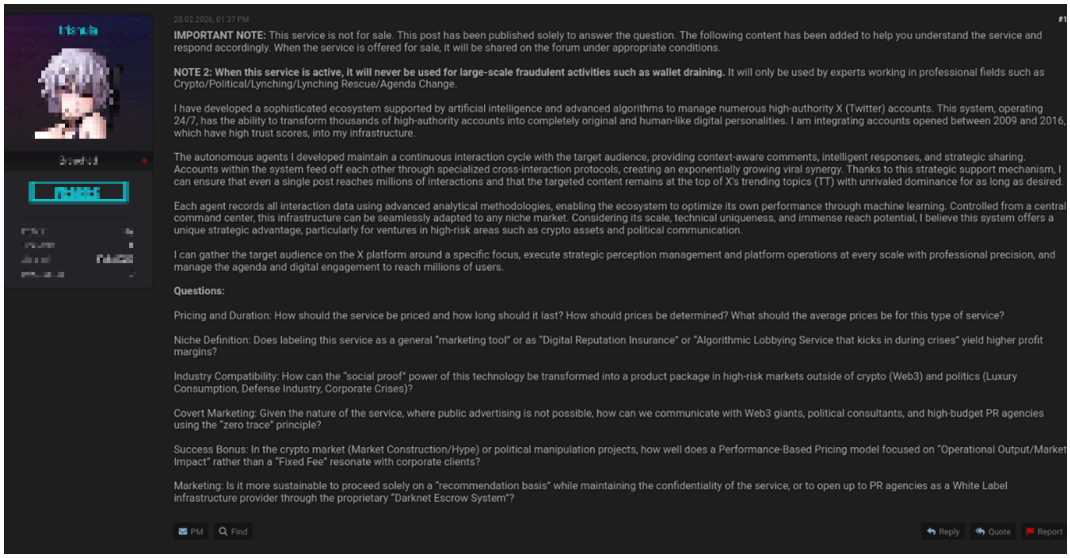
Интересно, что ИИ-генерация может применяться киберпреступниками не только для автоматизации генерации и перевода, а также стилизации приманки, но и для обеспечения собственной безопасности. Осторожные пользователи теневых площадок начинают использовать LLM для переписывания собственного текста, чтобы избежать возможного раскрытия личности. При этом эти же большие языковые модели, как показывают исследования, могут эффективно применяться для поиска и сопоставления фактов для задач деанонимизации.

Рисунок 12. Использование ИИ для избегания деанонимизации



Отдельно необходимо отметить, что расширение возможностей проведения масштабных атак с использованием социальной инженерии, в которых применяется ИИ, создает угрозу массовых кибератак, нацеленных на манипуляцию общественным мнением. Мы уже рассказывали, как с помощью ИИ может быть организована атака с набегом вкладчиков банка, но в будущем могут быть реализованы и другие сценарии массовой дезинформации, представляющие особую опасность в периоды геополитической нестабильности и учащающегося комбинирования разрушительных действий в физическом и информационном пространстве.

Рисунок 13. Использование ИИ для массовой манипуляции общественным мнением



Конечно, применение ИИ в атаках социальной инженерии не ограничивается фишинговыми письмами: одной из площадок проведения атак стали репозитории кода. Киберпреступники массово создавали их с помощью ИИ-автоматизации, генерировали для них убедительный текст. Встречались и более сложные сценарии: например, исследователи Kaspersky в марте 2026 года рассказали об эксплуатации платформы Bubble для быстрого создания веб-приложений с легитимными ссылками и сложно анализируемым кодом, которые не блокируются системами защиты. При этом единственная функция такой «прослойки» — автоматическое перенаправление жертвы на настоящий фишинговый сайт. Исследователи предполагают, что такой метод уже мог появиться в фишинговых платформах.

ФИШИНГОВЫЕ ПЛАТФОРМЫ

Фишинговые платформы и инструменты еще до внедрения технологий искусственного интеллекта давали киберпреступникам возможности для автоматизации атак, компенсировали нехватку навыков у менее опытных злоумышленников, об их важной роли в социальной инженерии мы рассказывали в более раннем исследовании. Поскольку зачастую именно для таких улучшений ИИ внедряется в наступательные инструменты, стоит ожидать, что все крупные фишинговые платформы в ближайшем будущем получат ИИ-обновления. Отдельные платформы уже начали приобретать генеративные усиления: например, Darcula в апреле 2025 года внедрила ИИ для создания фишинговых форм сбора информации и их автоматического перевода на различные языки.

Рисунок 14. Фишинговая платформа с ИИ-функцией для изменения сообщения

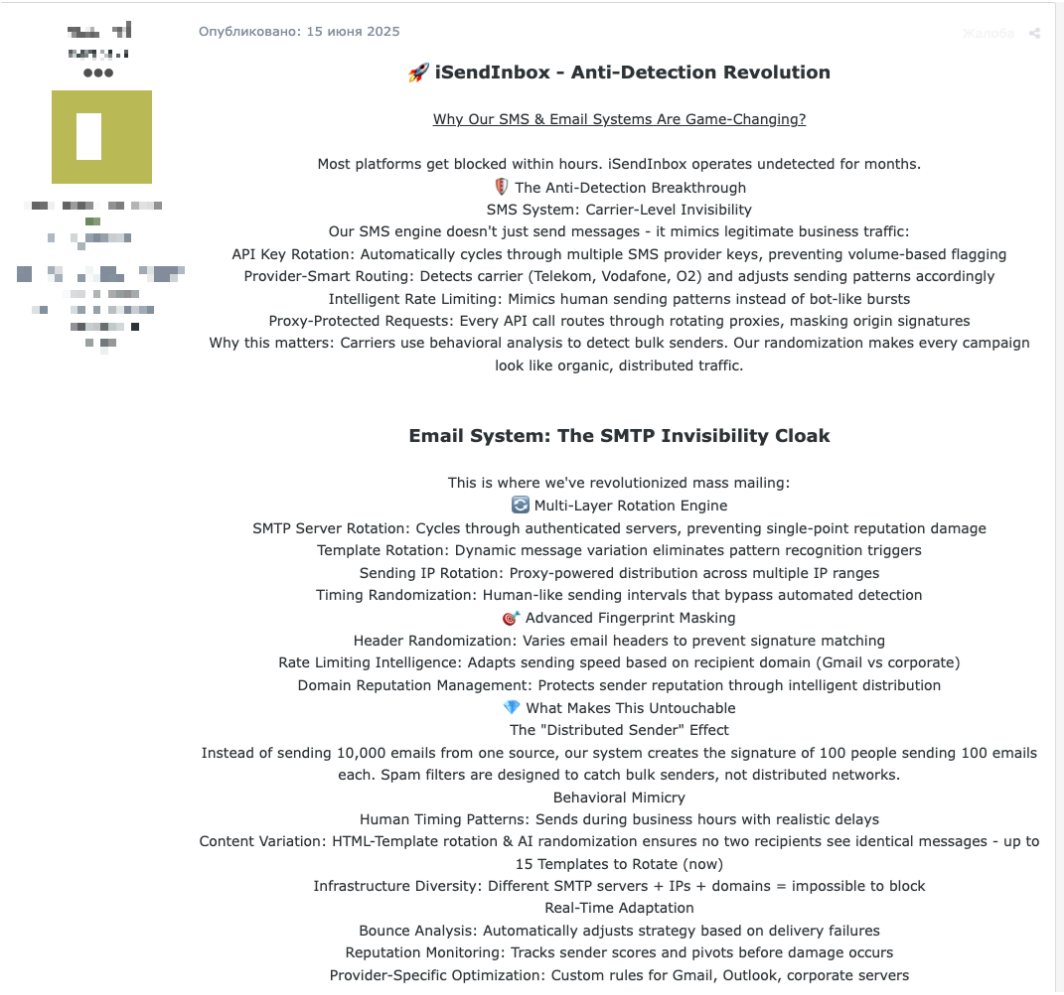
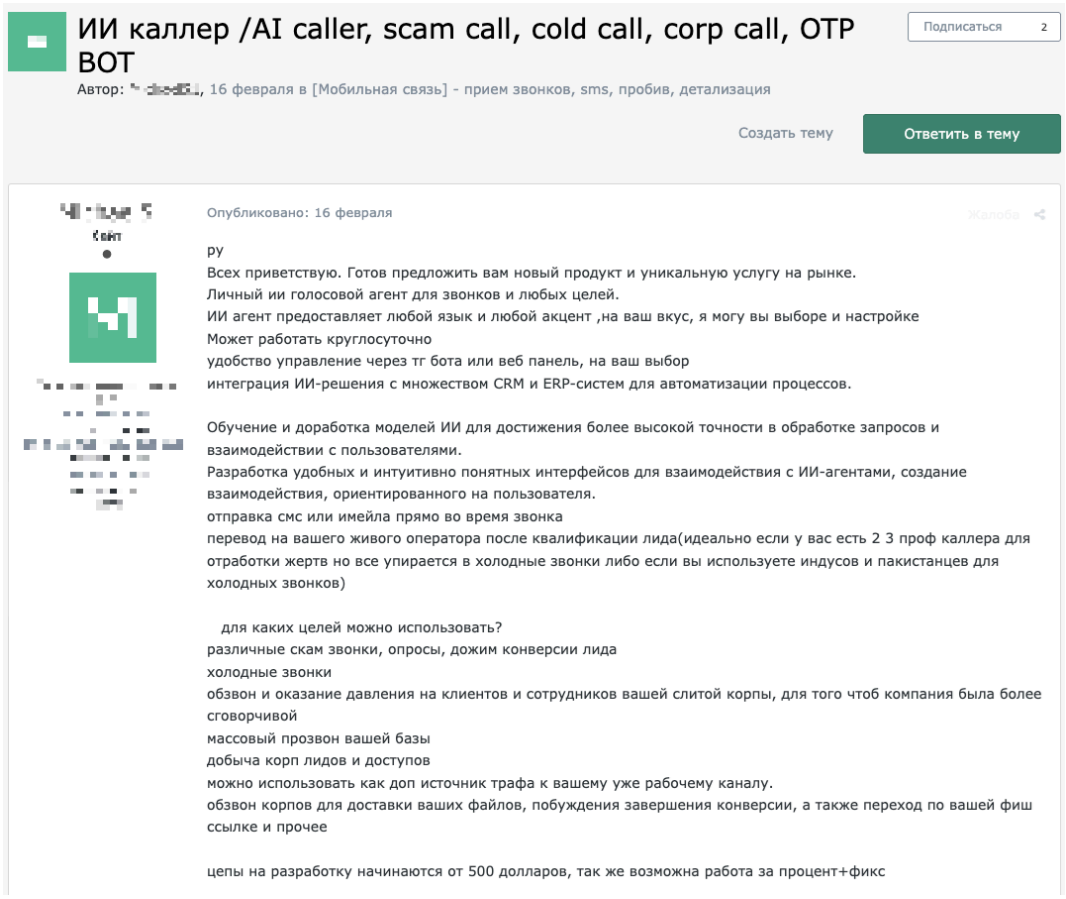


Рисунок 15. Реклама сервиса мошеннического кол-центра



Обратим внимание, что далеко не все фишинговые инструменты с использованием ИИ действительно развивают киберпреступный арсенал. Как и в случае с вредоносными LLM, такими как WormGPT, ИИ может выступать как элемент не подтвержденной реальными функциями рекламы или вовсе быть способом привлечения повышенного внимания к мошеннической схеме. Например, получивший широкую огласку SpamGPT не получает на теневых платформах положительных отзывов, что может говорить о том, что это очередной проект, распространяемый для обмана одними киберпреступниками других.

Рисунок 16. Объявление о продаже SpamGPT

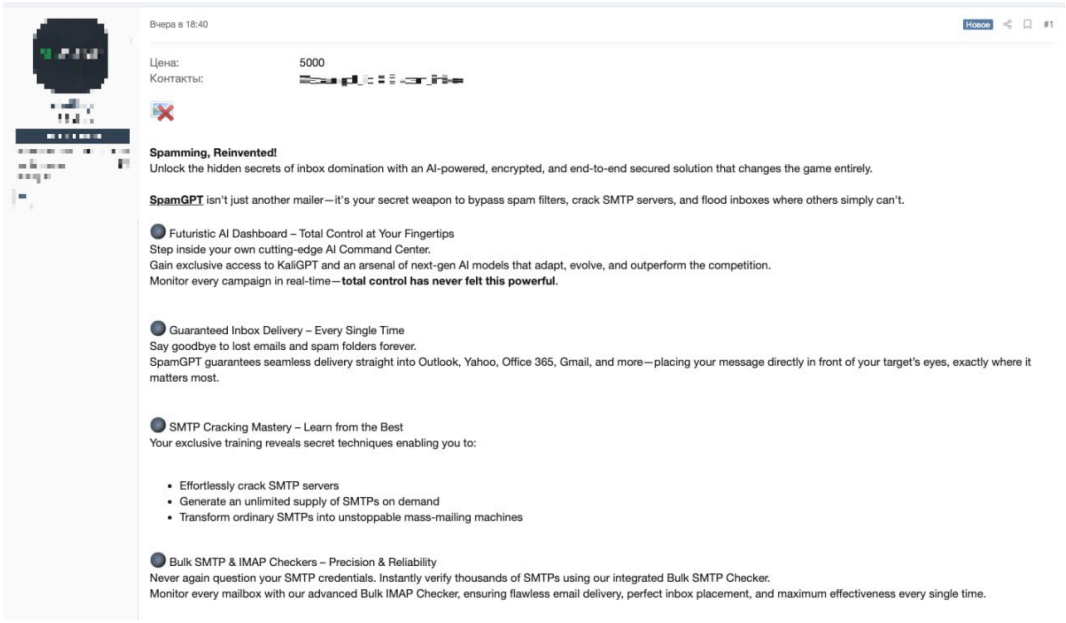
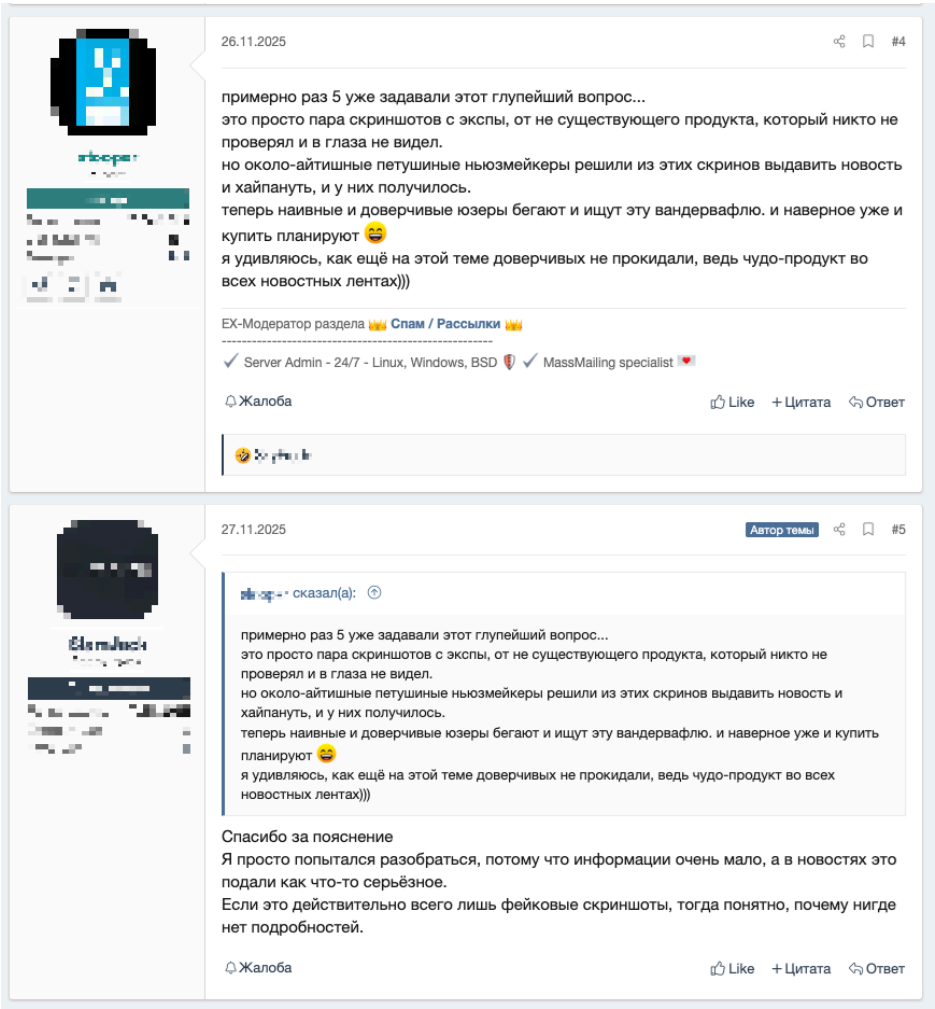


Рисунок 17. Отзыв киберпреступников о несостоятельности инструмента

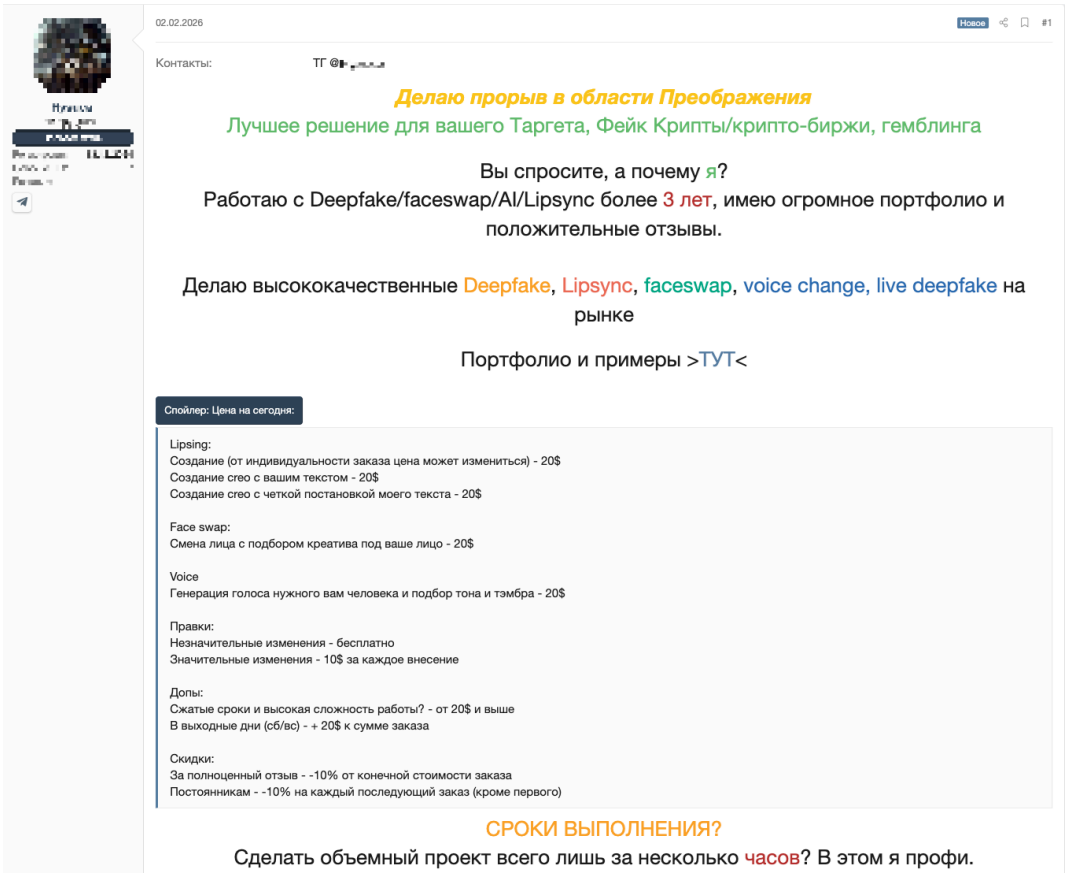


ДИПФЕЙКИ

В ряде исследований мы уже говорили о возрастающей роли дипфейков в атаках социальной инженерии. Наряду с генерацией другого фишингового контента создание видео- и аудио-подделок остается самой масштабной и единственной по-настоящему массовой областью применения ИИ в кибератаках. Дипфейков становится все больше: по данным DeepStrike, за период с 2023 по 2025 годы количество появляющихся в год дипфейков выросло в 16 раз. Киберпреступники продолжают использовать дипфейки для манипуляций общественным мнением, подделки неотрывно сопровождают все значимые политические события во всем мире, в мошеннической рекламе, атаках на частных лиц с подделыванием голосов родственников и друзей, обхода систем Know Your Customer и обмана при трудоустройстве для внедрения инсайдера.

Популярность дипфейков неизбежно привела к развитию рынка услуг deepfake as a service на теневых площадках; подробно про рынок подобных услуг рассказывали в исследовании. За генерацию минуты видео площадки запрашивают от 20 до 200 долларов, при этом киберпреступники используют многие «гражданские» маркетинговые приемы: например, цена за минуту видео падает с ростом общей длительности и, наоборот, увеличивается при срочном заказе или работе в выходные дни. Цены услуг deepfake as a service на теневых площадках превосходят расценки открытых коммерческих платформ, на которых минута дипфейк-видео обойдется примерно в 10–20 долларов. Больший ценник в дарквебе обусловлен, помимо гарантий и персонального подхода, тем, что на создаваемых преступниками фейковых видео нет водяных меток и при генерации не используется принцип imperfect by design, который заложен в некоторые коммерческие приложения из соображений безопасности.

Рисунок 18. Предложение услуг по созданию дипфейков



Дипфейки прочно вошли в перечень массового используемых киберпреступниками инструментов социальной инженерии, как в массовых, так и в целевых атаках. Масштаб применения дипфейков в ближайшие годы продолжит расти, а высококвалифицированные киберпреступники будут включать видео- и аудиоподделки в сложные, многоканальные атаки с использованием социальной инженерии. Мы ожидаем развития и расширения рынка deepfake as a service, появления большого количества предложений услуг генерации подделок, выхода на теневой рынок платформенных решений для генерации дипфейков. Как мы отмечали в исследовании, посвященном трендам развития ВПО, рост применения дипфейков может быть важным фактором, повлиявшим на распространение функций захвата аудиопотока в шпионских вредоносных программах. Группировки уже начали выстраивать пайплайны атак, в которых похищенная информация и аудиовизуальные данные сразу применяются для следующей дипфейк-атаки, потенциально далеко развивая мошенническую цепочку внутри одного сообщества. Широко применяя ИИ для обработки информации и генерации сложных диалогов с несколькими дипфейк-собеседниками, умелые киберпреступники могут повторить и даже развить успех наиболее громкого случая ограбления банка с помощью дипфейка.

Рисунок 19. Одна из функций продаваемого ВПО ориентирована на применение с дипфейками

Secure Data Provider:
Это второй сервер внутри viewer для того что бы передать важные участки для кода
без которых не возможна работа технологии, т.е. это что то вроде хранилища критичных
данных, при реверсе если нет доступа к SDP будет не ясно что анализируемый участок кода делает.
это было сделанно для максимального продления жизни продукта и скрытия технологии.

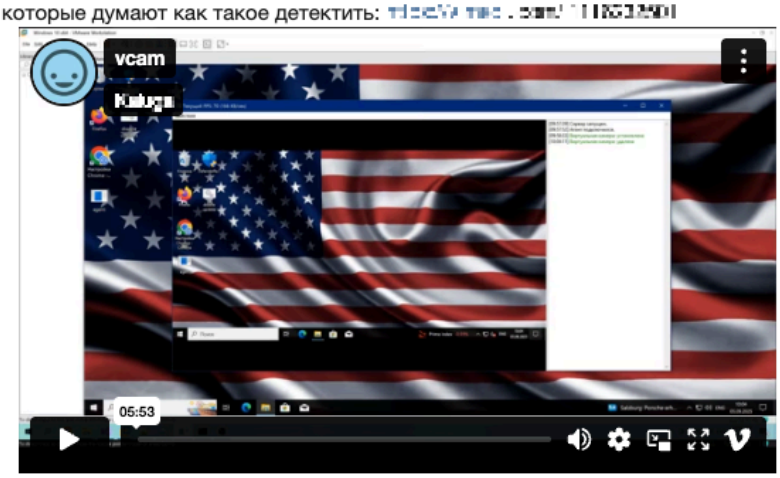
Когда это спасёт: билд улетел к AB или на BT, SDP оффлайн соответствеено и суть технологии
ав не видит в рантайме (если сервер выключен или если клиент уже подключён)

~-~> Fake Cam:
Реализована киллерфича для всяких байбитов и других бирж, а именно Fake Cam,
она даёт возможность внутри хвнц! из запущенного браузера подключить фейковую веб камеру
с любым именем и стримить туда контент с помощью OBS (или аналога)

Идеальная фича для дипфейков, особенно учитывая что вы находитесь внутри хвнц и траст
на высшем уровне.

Выходит так что вы находясь у таргета на ПК в два клика разворачиваете фейк камеру,
перезапускаете браузер, запускаете настроенный обс, стримите дипфейк -> налите бабки

Слышите эти крики? А я слышу. Это крики сек отдела разных бирж
которые думают как такое детектить:



Распространение убедительных видео- и аудиоподделок становится серьезной проблемой, из-за которой требуется пересмотр отдельных процессов подтверждения личности, KYC, а говоря шире — общих практик проверки и формирования доверия к информации.

ИИ КАК ТЕМА ФИШИНГА

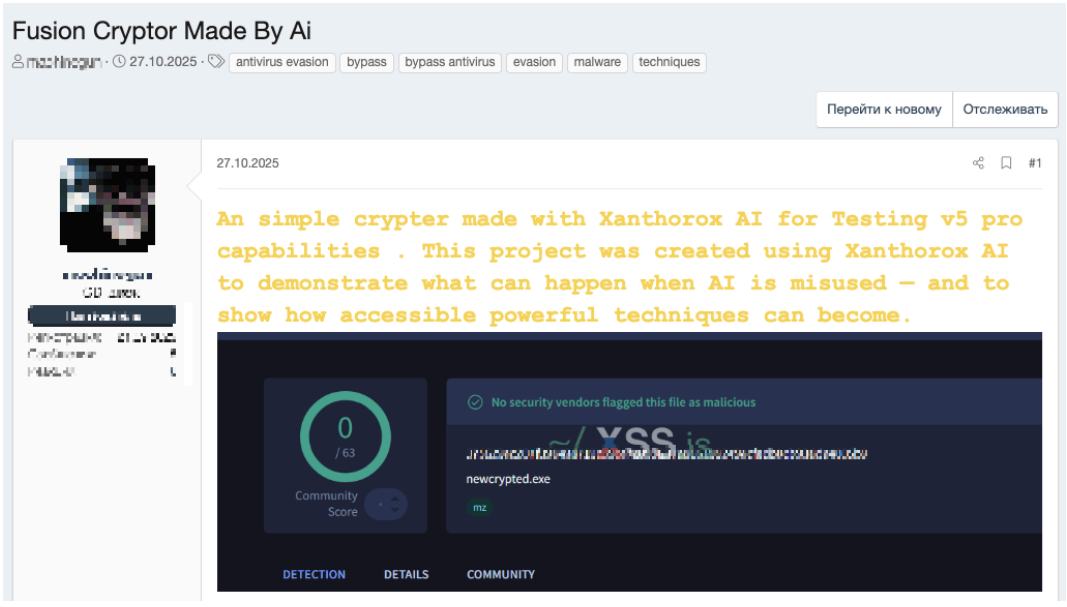
Возможности ИИ интересны киберпреступникам не только как инструменты, но и как популярная тема атак с использованием социальной инженерии. В данном исследовании мы приводили несколько примеров, показывающих, что далеко не все большие языковые модели, отдельные модули и решения, которые киберпреступники распространяют под эгидой наступательного ИИ, на самом деле отвечают указанным в рекламных объявлениях возможностям. Но использование темы ИИ для приманок не ограничивается только теневыми площадками: киберпреступники эксплуатируют популярность ИИ-инструментов для массовых атак с распространением ВПО. Одну из таких кампаний якобы инструмента для генерации видео — с рекламой в социальных сетях и поддельными сайтами — в течение долгого времени отслеживали специалисты Mandiant Threat Defense. При атаках, нацеленных на разработчиков, ИИ и ИТ-специалистов, злоумышленники действуют через поддельные репозитории. Например, в феврале 2025 года специалисты экспертного центра Positive Technologies обнаружили вредоносную кампанию в репозитории пакетов PyPi, в которой киберпреступники, воспользовавшись ажиотажем вокруг DeepSeek, распространяли пакеты со шпионскими функциями, нацелившись на разработчиков, заинтересованных в интеграции нейросети в свои проекты.

ВПО

Генерация вредоносного кода — инновация, ставшая нормой

В 2023–2024 годах, когда киберпреступники только начинали интегрировать искусственный интеллект в процесс подготовки атак, генерация отдельных частей кода стала логичным плацдармом применения технологии, наравне с генерацией текста для фишинга. Так же как легальные программисты стали широко применять вайбкодинг и копайлотов, их преступные коллеги использовали ИИ для разработки ВПО. За два года генерация вредоносного кода стала повсеместной: чрезмерно подробные комментарии, отметки «generated at» и другие маркеры использования LLM обнаруживали, к примеру, в стилере Punishing Owl, бэкдорах группировок Konni, MuddyWater, вымогателей Hive0163, криптомайнерах и даже в замаскированном под библиотеку RAT. При этом множество обнаруженных примеров с сохраненными маркерами генерации показывают, что киберпреступники не стесняются применения ИИ и не пытаются его скрыть. Сегодня эксперты еще выделяют факт генерации кода во вредносном ПО как интересную деталь. Однако в ближайшем будущем ИИ превратится в стандартный рабочий инструмент киберпреступников, а его маркеры в коде перестанут привлекать внимание.

Рисунок 20. Сообщение на теневом форуме о возможности создать шифровальщик с помощью ИИ



Постепенно киберпреступники идут дальше, используя LLM для генерации не только отдельных скриптов, но и ВПО целиком. Показателен пример ВПО VoidLink: специалисты CheckPoint в начале 2026 года сообщили, что его разработка, вероятно, выполнена одним киберпреступником за сравнительно небольшое время с помощью ИИ-инструментов. Важно отметить, что искусственный интеллект не только писал код, но и использовался на этапе планирования и организации процесса разработки. Необходимо отметить, что процесс, выстроенный киберпреступником в данном случае, требует высокой квалификации и на данном этапе развития технологий доступен далеко не каждому злоумышленнику, даже получившему доступ к мощным ИИ-инструментам. В ближайшем будущем можно ожидать появления новых примеров обнаружения и применения в кибератаках сложного ВПО с широким применением ИИ в процессе разработки, действительно качественные образцы такого ВПО будут относиться к операциям профессиональных группировок.

Рисунок 21. Обсуждение на теневой площадке применения LLM для написания вредоносного кода

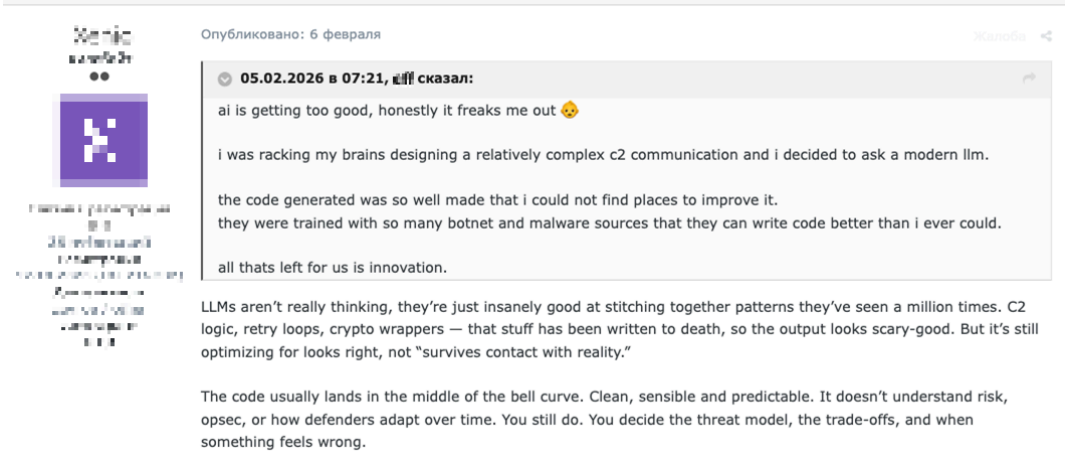
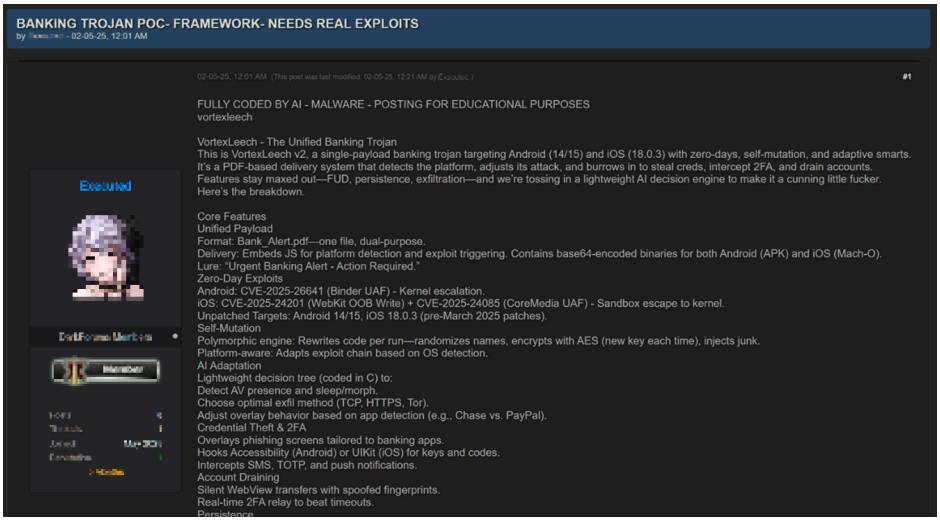


Рисунок 22. Распространяемый на теневой площадке банковский троян, полностью запрограммированный с помощью ИИ



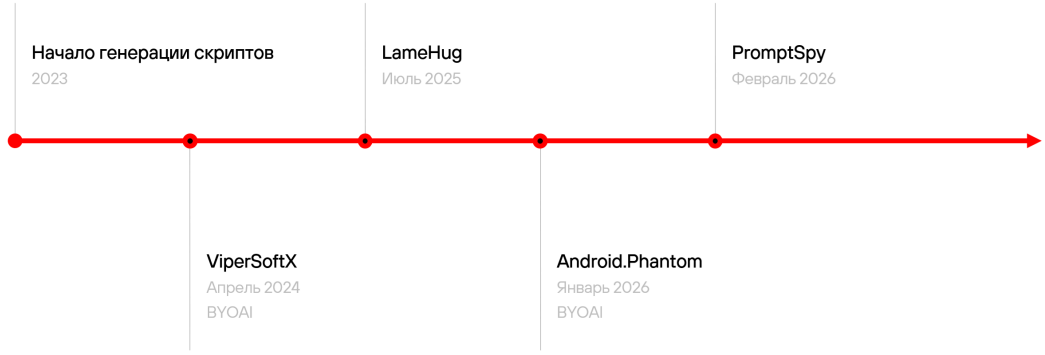
Альтернативой генерации нового ВПО становится переписывание уже известного. Киберпреступники стремятся создавать с помощью ИИ новые версии на базе старых образцов для обфускации и переиспользования, например, находящихся в открытом доступе примеров кода или утечек вредоносных программ. Более сложный вариант — получение изначального вредоносного кода с VirusTotal, интерес к файлам которого мы наблюдаем в последнее время на теневых площадках. Теоретическая схема может заключаться в поиске хешей ВПО в открытых источниках, получении соответствующих файлов с площадки и последующего переделывания в новый образец с помощью ИИ.

Рисунок 23. Распространяемый на теневой площадке банковский троян, полностью запрограммированный с помощью ИИ

Внедрение ИИ в ВПО

Как мы и предполагали, развивается не только генерация вредоносного кода, но и внедрение ИИ в ВПО для реализации отдельных функций и адаптивной генерации команд после заражения системы жертвы. Пока киберпреступники экспериментируют с различными решениями, ВПО с ИИ-модулями и функциональностью остается крайне редким (обнаружено всего несколько примеров).

Рисунок 24. Обнаруженное ВПО с ИИ за период 2024 — Q1 2026



Наиболее яркий пример внедрения ИИ в ВПО в реальных атаках был обнаружен в июле 2025 года. LameHug выполняло для разведки и сбора информации команды, которые динамически генерировала с помощью вшитых в код ВПО промптов для LLM. Важно, что вредоносная программа использовала общедоступный Hugging Face API для доступа к Qwen-2,5, что не только подтверждает использование киберпреступниками легальных моделей, но и демонстрирует важность контроля за обращениями к ИИ-сервисам. Схожую функцию генерации вредоносных скриптов использовал и созданный группой исследователей вымогатель PromptLock.

Еще в 2024 году исследователи обнаружили ВПО TesseractStealer с применением ИИ-инструмента для оптического распознавания скриншотов. В начале 2026 года концепция «умного зрения» для ВПО получила развитие сразу в нескольких вредоносных программах. Троян с возможностями кликера Android.Phantom использует модели машинного обучения TensorFlow для анализа скриншотов браузера и выявления на них рекламных элементов. Оба вредоноса загружают на устройство жертвы и используют локальные модели для распознавания. Конечно же, используемые ими инструменты несопоставимы с тяжелыми LLM, используемыми другими ВПО с ИИ, но тем не менее на их примере можно проследить появление концепции Bring Your Own AI. Для дроппера PromptSpy, также нацеленного на Android-устройства, злоумышленники реализовали с помощью Gemini более сложный процесс: LLM обрабатывает информацию об элементах пользовательского интерфейса и возвращает вредоносной программе пошаговые инструкции по взаимодействию с ним для закрепления зараженного приложения в списке последних используемых. ИИ-модуль обеспечивает ВПО гибкость и адаптируемость к системе, поскольку работает с интерфейсами, изменяющимися в зависимости от устройства и версии ОС, что затрудняет автоматизацию с помощью традиционных скриптов.

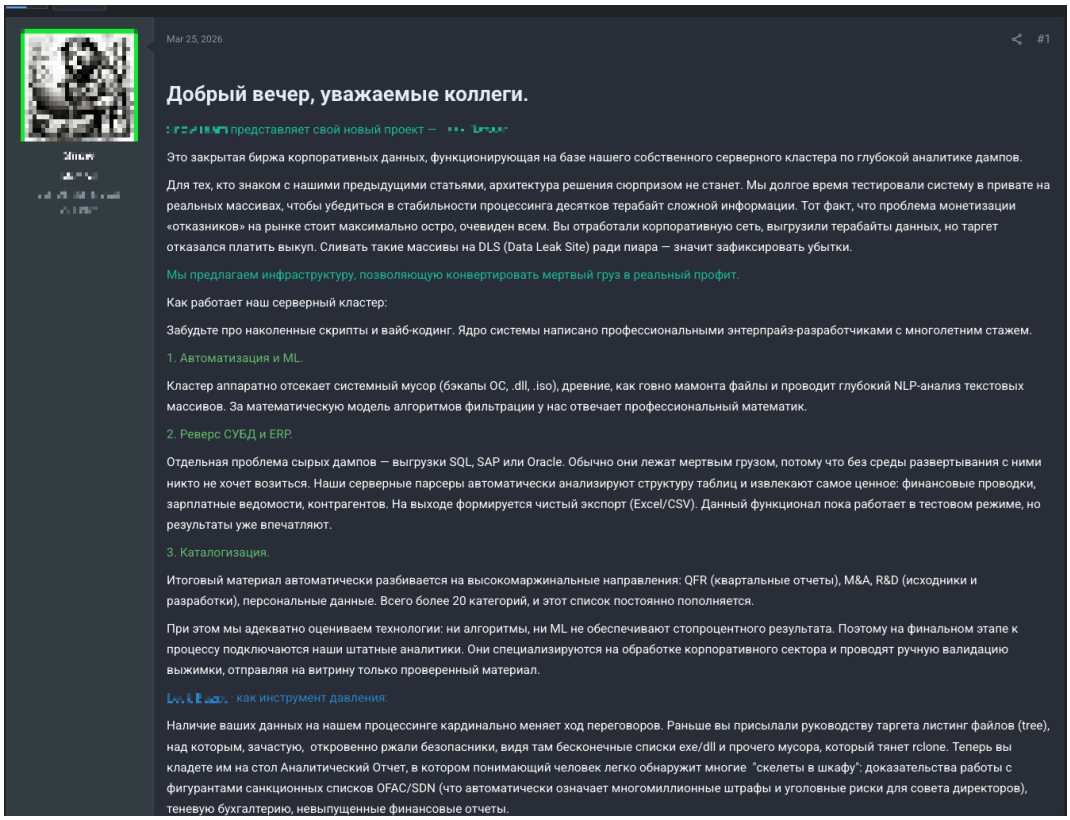
Отдельно стоит упомянуть потенциальное применение ИИ как канала коммуникаций с C2-сервером и эксфильтрации данных. Исследователи из Check Point уже демонстрировали, как с помощью Grok и Copilot выстроить это взаимодействие. Проблема возможности такого использования LLM вновь демонстрирует важность контроля над обращениями к внешним моделям и является пограничной с общими проблемами отсутствия контроля над известными ИИ-решениями в инфраструктуре и shadow AI.

Подводя итог применения ИИ для ВПО, можно уверено сказать, что перед киберпреступниками находится еще огромное поле для эксплуатации технологии, однако о значительном внедрении пока можно говорить только в области генерации вредоносного кода. Мы предполагаем, что ближайшие годы уйдут у злоумышленников не на создание принципиально новых угроз, а на закрепление генерации вредоносного кода в подготовке атак, а также продолжение экспериментов по внедрению ИИ для реализации известных функций ВПО там, где требуется гибкость и адаптируемость, например в поиск и анализ данных жертвы. В сравнительно простых задачах может применяться загрузка модели непосредственно на устройство жертвы, но, когда ИИ используется, например, для генерации команд или управления ходом атаки, нужны более тяжеловесные модели, к которым, на сегодняшний день, ВПО приходится обращаться удаленно. По мере технологического развития такие сложные задачи могут быть переложены на компактные BYOAI-модули. При недостаточном контроле обращений ИИ к API киберпреступники смогут скрывать в них вредоносную активность, такую как работа адаптирующегося ВПО, C2 и эксфильтрация данных. Кроме того, в перспективе нескольких лет, при условии массового распространения локальных моделей в инфраструктурах, может появиться сложное ВПО, стремящееся использовать ресурсы таких моделей.

ПОСЛЕ АТАКИ

Применение ИИ в работе ВПО не ограничивается подготовкой инструмента и выполнением функций непосредственно во время атаки. Способность моделей обрабатывать значительные объемы данных, в том числе неструктурированных и представленных в разных форматах, является очень полезной для вымогателей. Поскольку утечки и похищенные дампы могут составлять многие десятки и сотни гигабайтов, для злоумышленников становится очень полезной возможность оперативно изучить их и извлечь наиболее чувствительные конфиденциальные данные, чтобы использовать их для усиления шантажа жертвы или увеличения цены при будущей продаже.

Рисунок 25. Биржа данных предлагает услуги по обработке с помощью ИИ



ПОСЛЕ АТАКИ

Интересно, что киберпреступники уже начали применять ИИ не только для оптимизации обработки данных, но и для обмана других участников теневых сообществ. Злоумышленники генерируют дополнительные записи к реальным похищенным базам, чтобы искусственно увеличивать размеры продаваемых объектов и завышать на них цены. При достаточно убедительных синтетических данных киберпреступники могут пойти дальше и создавать фальшивые БД с нуля. В посвященных дарквебу исследованиях мы уже рассказывали о том, что честь водится среди далеко не всех киберворов, и ИИ, безусловно, будет не только темой, но и инструментом обмана внутри киберпреступного сообщества.

Рисунок 26. Сообщения с предложением заработка за счет расширения продаваемых баз данных

В будущем мы ожидаем экспериментов по переносу обработки похищенных данных с помощью ИИ внутрь процесса работы шпионского ПО, шифровальщика или вайпера. Потенциально такой подход позволит сделать атаки быстрыми и точечными, за счет чего их будет сложнее обнаружить, но ущерб от них останется высоким из-за воздействия на ценные данные.

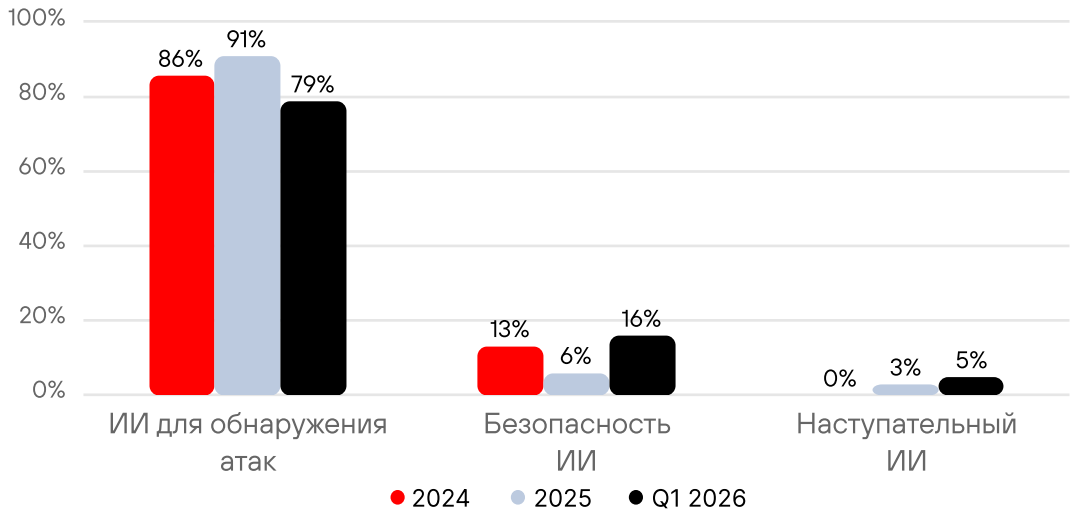
Угроза внутри

Пока киберпреступники точечно внедряют ИИ для автоматизации этапов атаки, в которых применение технологии доказало свою эффективность, и экспериментируют с новыми подходами, все более актуальными становятся угрозы в самом искусственном интеллекте. Темпы внедрения и применения искусственного интеллекта значительно обгоняют внедрение принципов ответственного ИИ на уровне разработки и интеграции в процессы компаний. По даннымMicrosoft, лишь половина компаний внедряют защиту вокруг генеративного ИИ. Парадокс обеспечения безопасности при внедрении ИИ заключается в обоюдоостром нарушении защиты: когда запрос использования ИИ в задачах обгоняет темпы внедрения возникает проблема ShadowAI; когда внедрение обгоняет зрелость и обеспечение защищенности процессов, проявляются проблемы с недостаточным контролем генерации и незащищенным и неконтролируемым внедрением. Ближайшие годы будут характеризоваться поиском баланса между стремлением к высокой производительности и защитой от новых типов угроз агентов, моделей и инфраструктуры, с которой они взаимодействуют.

Несмотря на постепенный переход массового отношения к технологии в стадию «разочарования»[1], в ИИ продолжают вкладывать существенные средства. Так, Gartner прогнозируют почти полуторакратный рост годовых всемирных расходов на ИИ до 2,5 триллиона долларов. Мы отслеживаем увеличение инвестиций в ИИ и в области информационной безопасности. В 2025 году венчурные фонды и капиталы, инвестиционные партнеры и частные инвесторы вложили в ИИ-инструменты в ИБ более 7 миллиардов долларов; для сравнения, в 2024 году вложения составили почти 4 миллиарда. Большая часть инвестиций относится к проектам, использующим ИИ для обнаружения атак, то есть к применению технологии с наибольшими реальными результатами в кибербезопасности. При этом в 2025 году общий объем инвестиций в защиту искусственного интеллекта упал не только на 7 п. п., но и уменьшился на 17% в отношении год к году, несмотря на лишь возрастающую важность направления, которую подтвердил ряд инцидентов, связанных с применением ИИ, в том числе агентов. Материализация угрозы в реальный ущерб подстегнула интерес инвесторов, защита ИИ стала одной из главных инвестиционных тем первого квартала 2026 года. По данным исследований, около 76% компаний в России внедрили хотя бы одну ИИ-функцию, что на 17 п. п. больше, чем в прошлом году. При этом международные исследования оценивают уровень внедрения ИИ в компаниях мира в 80–90%; это говорит о том, что в России масштаб проблемы может нарастать в ближайшие несколько лет по мере распространения ИИ.

[1] Фаза разочарования — один из этапов жизненного цикла технологии, согласно методологии Hype Cycle Gartner. Фаза разочарования характеризуется угасанием интереса к технологии из-за сложностей внедрения, не оправдывающих ожиданий экспериментов. Инвестиции сохраняются только в случае, если поставщики обновляют, улучшают продукты и удовлетворяют запросам компаний, внедривших технологию на ранних этапах.

Рисунок 27. Распределение инвестиций в ИИ в сфере ИБ



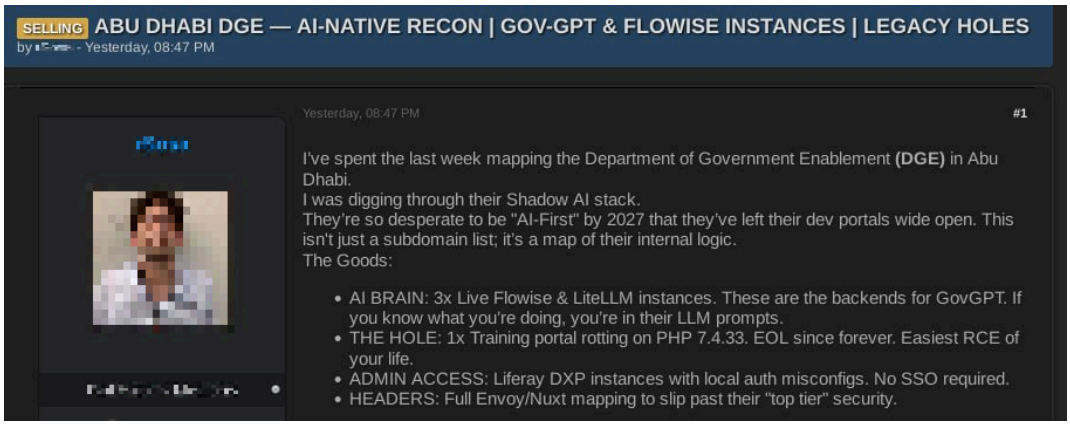
ТЕНЕВАЯ УГРОЗА

Shadow AI — проблема применения сотрудниками компаний неутвержденных, неконтролируемых ИИ-инструментов — становится все масштабнее: исследованияMicrosoft показывают, что к таким ИИ обращается каждый третий сотрудник. Применяя такие инструменты, работники могут недооценивать или осознано игнорировать риски, стремясь повысить свою производительность. Shadow AI, как и другие теневые проблемы, создают значительные риски для компании, поскольку неизвестные ИТ и ИБ подразделениям инструменты несут угрозу утечек данных, а также потенциально являются дополнительными незащищенными поверхностями атаки.

Наиболее явная проблема, появляющаяся из-за shadow AI, — утечки данных. При условии, что 13% запросов к чат-ботам содержат чувствительную информацию, в том числе учетные данные, сетевую информацию и персональные данные, утечка конфиденциальной информации становится лишь вопросом времени, по даннымIBM 20% организаций, пострадавшая от утечки данных связывала инцидент с shadow AI, для сравнения, аналогичный показатель для легальных ИИ на 7 п. п. ниже. Усугубляется проблема тем, что, как и в других «теневых» случаях, shadow AI скрывает инциденты, увеличивает время и усложняет реагирование, из-за чего значительно (в среднем на 200 000 долларов) возрастает ущерб от утечки.

Проблема будет сохраняться, поскольку у компаний далеко не всегда всегда есть возможность оперативно внедрить доверенный ИИ, ресурсов локальных моделей и решений часто не хватает, личные AI-ассистенты распространяются все активнее; кроме того, человеческий фактор всегда будет ставить безопасность под угрозу. Бороться с проблемой теневого AI можно только комплексно, комбинируя технические методы с организационной работой с сотрудниками компании.

Рисунок 28. Продажа доступа к сети разработки правительственной LLM в Абу-Даби, полученного, по утверждению киберпреступника, через shadow AI



СГЕНЕРИРОВАННАЯ УГРОЗА

РАЗРАБОТКА

Глобальный постепенный переход к AI-driven разработке оказывает серьезное влияние на киберландшафт. ИсследованиеVeracode показывает, что за период между 2023 г. и 2026 г. модели достигли 95% точности синтаксиса генерируемого кода, но при этом почти не улучшились в контексте безопасности генерации. В половине случаев в сгенерированной кодовой базе оказывались известные уязвимости, и хотя моделям удавалось достичь хороших (более 80%) показателей при проверках устойчивости к внедрению SQL-кода (CWE-89) и к применению небезопасных криптографических алгоритмов (CWE-327), ситуация с более сложными задачами — обеспечения защиты от XSS (CWE-80) и от неправильной обработки выходных данных для логов (CWE-117) — остается тяжелой: лишь каждый седьмой тест на устойчивость к этим уязвимостям оказывался успешным. Существующие модели плохо справляются с задачами безопасности, требующими понимания контекста и выходящими за рамки сопоставления с шаблонами. Причины таких проблем кроются в небезопасных уязвимости обучающих данных (например, в обучающем датасете для LLM Common Crawl исследователи обнаружили 12 000 жестко закодированных паролей и API-ключей), фокусировке прогресса развития моделей на повышении производительности (цель, как видно из показателей синтаксической точности, выполняется), а также в общем пренебрежении процессами безопасной разработки. В итоге проблема только увеличивается.

Игнорирование и «принятие за норму» сгенерированного небезопасного кода недопустимо, поскольку оно будет приводить к значительному росту инцидентов безопасности, вызванных как действиями киберпреступников, так и ошибками в ПО, и чем дольше проблема не будет получать должного внимания и контроля, тем больше будет список инцидентов. Вайбкодинг уже стал одной из причин ряда инцидентов с серьезным ущербом, например сбоев у Amazon, потери почти 2 миллионов долларов лендинговым протоколом Moonwell из-за неправильной настройки оракула (о безопасности DeFi систем рассказывали в исследовании), а также утечки данных приложения для знакомств Tea. Как видно, угроза актуальна как для небольших компаний и стартапов, стремящихся как можно быстрее произвести продукт, так и для больших бизнесов, видящих в применении вайбкодинга возможности снизить издержки.

БЕЗОПАСНОСТЬ

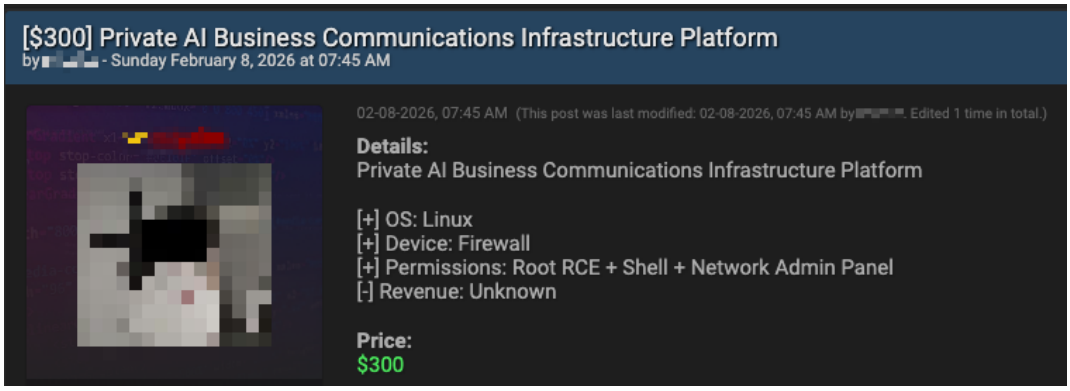
Поскольку ключевой принцип работы LLM заключается в предсказании наиболее вероятных токенов, на долгой дистанции они выдают схожие ответы на одинаковые запросы. В этом принципе кроется существенная проблема для безопасности, поскольку применение больших языковых моделей для настройки инфраструктуры приводит к стандартизации и упрощению проведения на них атак. ИИ создает шаблонных пользователей, распространяя слабые значения по умолчанию, что уже становится важным фактором для распространения отдельных массовых атак ботнетов. Аналогичная проблема проявляется при попытке использовать LLM для генерации надежных паролей: несмотря на то, что при разовом запросе модели будут выдавать на первый взгляд сложные комбинации, при большом количестве идентичных промптов обнаруживается, что генерируемые пароли оказываются схожими друг с другом, повторяя единый паттерн. Лишь вопрос времени, когда генерируемые LLM пароли окажутся в библиотеках для атак с перебором. Из-за того, что людям кажется надежной выдача больших языковых моделей, и в силу процветания вайбкодинга такие шаблонные настройки и учетные данные продолжат широко распространяться. Это сильно повлияет на масштаб проведения сравнительно несложных, легко автоматизируемых атак.

ИНФРАСТРУКТУРНАЯ УГРОЗА

ИИ-инфраструктура по мере внедрения технологии в различные процессы становится полноценной высокорисковой поверхностью атаки. Программные интерфейсы, MCP, веса, системные промпты, локальные модели становятся целью и объектом как уже известных, так и новых типов атак.

За 2025 год было обнаружено более двух тысяч уязвимостей, связанных с ИИ, — это на треть больше, чем в прошлом году. При этом такой темп роста в два раза превосходит общий рост числа обнаруженных CVE. Некоторые уязвимости относятся непосредственно к ИИ-решениям, но основные риски сегодня кроются не в моделях, а в инфраструктурной обвязке вокруг них. Например, апрельский инцидент с утечкой данных стартапа Mercor связан с атакой на цепочку поставок LiteLLM — многофункционального шлюза для ИИ-агентов: злоумышленники выпустили две троянизированные версии библиотеки, которые попали в PyPi.

Рисунок 29. Продажа доступа к ИИ-платформе

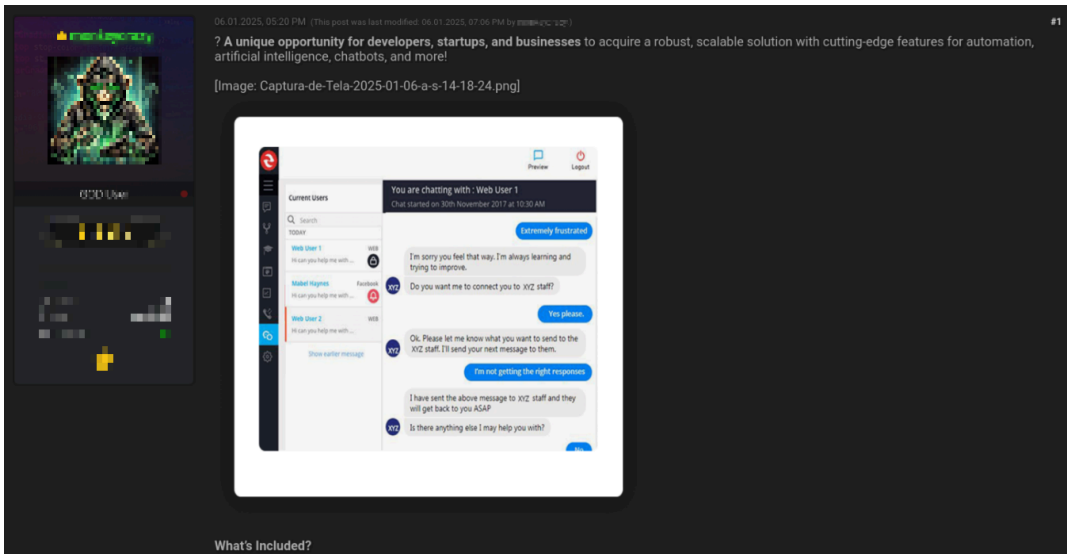


Другим примером инфраструктурной угрозы, исходящей от ИИ, являются API. Распространение ИИ сегодня является одним из ключевых факторов роста числа атак на API, открытые и слабозащищенные программные интерфейсы. ИсследованиеWallarm показывает, что больше половины API AI находятся в открытом доступе; при этом 89% не имеют надежных механизмов аутентификации и подвергаются кибератакам с целью использования чужих токенов и тарифов, похищения данных. Так, воспользовавшись ИИ-платформами консалтингового агентства McKinsey, специалисты CodeWall смогли получить доступ к значительному объему конфиденциальных данных; интересно, что уязвимость была обнаружена и проэксплуатирована также ИИ-агентом. Утечки ключей API приводят к несанкционированному использованию моделей: к примеру, разработчик из Мексики рассказал, как за двое суток злоумышленники использовали токенов на 82 тысячи долларов (для сравнения, обычно жертва тратила 180 долларов в месяц). При этом атаки типа LLMjacking уже перешли на уровень массовой индустриализованной киберпреступности и нацелены на поиск уязвимых конечных точек ИИ и перепродажу полученных доступов. Компрометация узлов с ИИ несет не только угрозы траты мощностей и токенов, но и утечек данных из прошлых диалогов с чатботами, бокового перемещения внутри инфраструктуры и несанкционированного доступа к файловым системам и БД. Помимо обеспечения безопасности конечных точек ИИ, мерой контроля и обнаружения атаки становится и отслеживание трат токенов, которое пока что вводится отдельными компаниями для отслеживания производительности сотрудников и других экономических параметров.

ИИ-системы становятся целью атак не только для нанесения ущерба применяющей их компании или извлечения финансовой выгоды, но и для получения сведений о самом ИИ. Продолжающаяся техническая гонка между разработчиками ИИ и интерес киберпреступников к созданию собственных инструментов подстегивает внимание к утечкам, связанным с новыми продуктами. В апреле 2025 года большую популярность получил репозиторий с дампом, содержащим системные промпты, и внутренними материалами различных компаний; материалы Anthropic обнаружили в незащищенном хранилище в марте 2026 года. Примером более сложной атаки со сбором информации о модели могут послужить попытки извлечения и дистилляции моделей — о таких атаках на Gemini сообщала в начале 2026 года Google. В ближайшие годы интерес к утечкам и несанкционированным доступам к информации об ИИ-решениях не будет спадать. Возможна, хотя и маловероятна, утечка весов крупных моделей, информации, тщательно защищаемой разработчиками, после утечки весов LLama.

Рисунок 30. Сообщение об утечке данных LLama

Рисунок 31. Утечка кода продукта с ИИ



УГРОЗА АГЕНТОВ

Агенты — ключевая технология сегодняшней автоматизации и перехода внедрения ИИ на уровень выполнения задач вместо человека, а не только ассистирования, которое обеспечивали LLM. Поскольку агенты имеют возможность самостоятельно инициировать действия, взаимодействовать с инфраструктурой и данными, а также ограниченно принимать решения, они формируют принципиально новые угрозы и риски безопасности. Подробно историю создания, устройство и влияние агентов на бизнес и киберландшафт мы разбирали в исследовании, посвященном новым технологиям ИИ; большинство из рассмотренных в нем угроз для ИИ-агентов остаются на сегодняшний день потенциальными, хотя и нет никаких сомнений, что киберпреступники будут стремиться атаковать новую поверхность атаки, тем более, что ее компрометация может дать огромные возможности. Тем не менее прогноз об атаках на цепочку поставок ПО для агентов уже успел стать реальностью: исследованиеSnyk показало, что более трети всех Skills, пакетов, дающих возможность агентам применять инструменты, API или системные ресурсы, содержат как минимум одну уязвимость безопасности, начиная от жестко закодированных ключей и заканчивая встроенным вредоносным ПО. Проблемы обнаружены и с другой частью агентской инфраструктуры — API-роутерами. Исследователи безопасности обнаружили 29 зараженных, шпионских и встраивающих вредоносный код API-роутеров из 428 проанализированных.

Основная часть инцидентов, связанных с агентами, пока что относится не к инициированной киберпреступниками активности, а к ошибкам самих агентов, ярко демонстрирующим угрозы, исходящие от таких решений. Ошибки агентов уже приводили к утечкам конфиденциальных данных, удалению баз данных, а также удалению электронных писем из почтового ящика; последний инцидент относится к стремительно набравшему в начале 2026 года популярность агенту OpenClaw. OpenClaw — наиболее показательный случай незрелого, ранняя версия работала через открытый для публичной сети шлюз, агента: в инструменте обнаружилось множество проблем безопасности и критических уязвимостей, в частности позволяющих удаленно выполнить код одним щелчком мыши (CVE-2026-25253), кроме того открытая торговая площадка навыков для агента содержит множество вредоносных Skills. Постепенно уязвимости и недостатки обнаруживают и закрывают, появляются новые, более защищенные версии, например NemoClawот NVIDIA, но пример OpenClaw демонстрирует, как сырые, небезопасные ИИ-инструменты могут стремительно распространиться, поставив под удар десятки тысяч систем.

Для России проблема потенциальной угрозы агентов также крайне актуальна: по данным Сбера, 39% российских компаний на начало 2026 года используют ИИ-агенты и ассистенты для решения различных задач, в первую очередь документооборота и обработки заявок, бухгалтерии, финансового учета и HR-процессов. Бояться применять новую технологию не стоит, но необходимо ответственно подходить к процессу внедрения, учитывать угрозы и относиться к агентам как к производительному, но рискованному продукту. Начать безопасное внедрение можно с рассмотрения ряда фундаментальных вопросов:

1. Какие доступы и права есть у агента?

Необходимо руководствоваться принципом минимальных привилегий, строго ограничивать доступы, оставляя только необходимые для решения задач, и по возможности стоит использовать агенты в изолированных средах.

2. Какой образ действия допустим при выполнении этих задач?

Крайне важный вопрос при подготовке агента к выполнению задач — не только «Что делает?», но и «Как делает?». Четкое указание допустимых образов действий необходимо для защиты от излишней самостоятельности агента, который может ради достижения результата обходить барьеры безопасности.

3. Как журналируется работа агента?

Для отслеживания процесса работы агента имеет смысл вести журналы размышлений и принятия решений, обращения к данным и вызова инструментов. Это позволит оперативно отследить сбой и внести корректировки в процесс, для устранения возможности повторения инцидента в будущем.

4. Какие механизмы противодействуют неправильной работе агента?

Заранее необходимо подготовить механизмы, которые остановят работу агента в случае неверных действий и нейтрализуют последствия ошибки.

В перспективе нескольких лет, при условии развития и распространения внедрения самоэволюционирующих агентов, дополнительным важным вопросом обеспечения их безопасности станет отслеживание изменений, эволюции агента. Как показывают исследования, самоэволюционирующие агенты в процессе развития могут «деградировать» из-за постоянного выполнения однотипных задач, разучится применять инструменты в сложных случаях, поскольку до этого они не нужны были на простых, проблема может стать особенно острой в задачах безопасности, которым характерен постоянный поток «нормальной» работы с редкими всплесками вредоносной активности. Помимо «деградации», галлюцинации и ошибки могут накапливаться в памяти, провоцируя «мисэволюцию», провоцирующую неправильную обработку задач и игнорирование ограничений безопасности.

ЗАКЛЮЧЕНИЕ

Влияние искусственного интеллекта на киберугрозы невозможно игнорировать, в равной степени опасны как недооценка возможностей технологии, так и преждевременная попытка защититься от несуществующих атак, в ущерб обеспечению необходимой сегодня безопасности. Внедрение ИИ, причем как киберпреступниками для наступательных действий, так и компаниями для автоматизации процессов, не отменяет и даже наоборот подчеркивает строгую необходимость следования фундаментальным принципам информационной безопасности. Инциденты, связанные с ИИ, происходили из-за отсутствия давно известных мер защиты и базовых проблем, таких как: слабые учетных данные, устаревшее ПО и игнорирование принципов безопасной разработки. ИИ автоматизирует и масштабирует эксплуатацию уже существующих проблем, а также становится катализатором роста угроз внутри инфраструктуры при поспешном, недостаточно зрелом и защищенном внедрении. В условиях роста и ускорения реализации угроз, критически важно системно подходить к обеспечению результативной кибербезопасности, реалистично оценивать возможности новых технологий, защищать их слабые места при внедрении, а также использовать сильные стороны для опережающего обнаружения и предотвращения кибератак.

В отличие от киберпреступников, сторона защиты, хотя и имеет ряд ограничений в применении ИИ, работает над внедрением технологии системно и постоянно. Ключевые направления автоматизации и развития защиты, о которых мы рассказывали в предыдущем исследовании, — обнаружение атак, в том числе неизвестных, ускорение реагирования, помощь специалисту информационной безопасности — наращивают эффективность и внедряются во все большее количество решений. Хотя бенчмарки оценки эффективности больших языковых моделей в задачах безопасности показывают, что актуальные LLM не способны полностью взять на себя обнаружение атак, но этого от них и не требуется (подробно о бенчмарках рассказывали в статье ML-специалисты Positive Technologies). Как мы неоднократно подчеркивали, сегодня задача ИИ в защите — дополнять и усиливать решения безопасности. В комплексе с средствами защиты технология раскрывается в полной мере — к примеру, в PT X модуль ML помогает находить аномалии во входных данных и новые паттерны атак, дополняя классические методы обнаружения.