kaspersky

Чек-лист

Как безопасно внедрять ИИ в процессы разработки

Как безопасно внедрять ИИ в процессы разработки: чек-лист

Разработка ПО — сфера, в которой ИИ является сейчас главным драйвером изменений. Компании в погоне за эффективностью тратят огромные бюджеты на ИИ-вычисления, а лидеры индустрии пророчат закат профессии программиста. Если в штате вашей компании есть разработчики, они, скорее всего, уже пользуются агентными IDE (интегрированными средами разработки) вроде Cursor или отдельными агентными инструментами, такими как Claude Code или Codex — или в работе, или в личных экспериментах.

Учитывая перспективы, которые дает агентный ИИ (Agentic AI) в разработке — от роста производительности труда и скорости реализации нового функционала до покрытия тестами и автоматизированного создания документации — компании начинают активно внедрять ИИ в свою разработческую практику, не дожидаясь, пока в компании расцветет неподконтрольный «теневой ИИ».

Kaspersky AI Technology Research совместно с другими подразделениями начал внедрять LLM (большие языковые модели) во внутренние сервисы, включая инструменты разработчиков, еще в 2023 году, и сегодня ими еженедельно пользуются почти 2 тысячи сотрудников. LLM помогают им в написании кода, создании тестов, ориентировании в задачах и кодовой базе, написании документации и код-ревью. На основании этого опыта мы составили чек-лист важнейших практик, связанных с ответственным и безопасным внедрением ИИ-технологий в разработку.

Ответы на приведенные в чек-листе вопросы помогут как составить план реализации ИИ-трансформации разработки с учетом потенциальных рисков, так и обнаружить пробелы тем компаниям, которые уже активно применяют ИИ в разработке.

Продукты

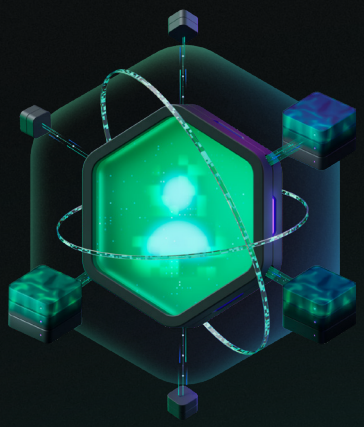
- Endpoint Security
- Kaspersky Endpoint Detection and Response
- Kaspersky Unified Monitoring and Analysis Platform
- Kaspersky Symphony XDR

1 Защищено ли рабочее место сотрудника?

Защита конечных точек в агентную эпоху не перестает быть актуальной. Напротив, когда система на базе LLM может автономно выполнять такие действия, как скачивание и запуск скриптов из интернета или поиск файлов и отправка электронной почты, причем со скоростью, значительно превышающей человеческую, именно защита конечных устройств становится основой безопасности. Учитывая, что действуя под воздействием LLM-специфичных атак (непрямых промпт-инъекций), ИИ-агенты могут выполнять опасные действия, на лету генерируя вредоносные команды, вторым важным фактором становится наличие продуктов класса EDR и SIEM для обнаружения опасных действий, расследования и купирования потенциальных инцидентов.

Кейсы

s1ngularity — инцидент, связанный с атакой на билд-систему nx. Будучи примером атаки на цепочку поставок, эта атака примечательна тем, что для поиска и эксфильтрации секретов с компьютера жертв использовались (при наличии) установленные на них ИИ-ассистенты разработчика, такие как Codex и Amazon Q. Наличие мониторинга действий на компьютере разработчика могло помочь обнаружить аномальную активность и предотвратить ущерб.



Продукты

- Kaspersky Open Source Software Threats Data Feed
- Kaspersky Scan Engine
- Kaspersky Container Security

2 Защищены ли цепочки поставок?

Атаки на цепочку поставок — один из наиболее актуальных факторов риска при разработке ПО. К сожалению, в сфере ИИ эта проблема особенно выражена: огромная скорость развития технологий зачастую приводит к тому, что пользователи стремятся использовать самые последние версии, а разработчики — выпускать обновления как можно быстрее, что может приводить к проблемам. При этом автономный ИИ-агент для разработки может установить не только скомпрометированный пакет, но и изначально вредоносный, например, в результате галлюцинации. Наличие централизованных корпоративных репозиториев для пакетов с их проверкой на безопасность и соответствие политикам может значительно снизить риск подобных атак.

Кейсы

Mini Shai-Hulud — атака, в результате которой в мае 2026 года оказалось скомпрометировано множество пакетов. Среди них — Mistral AI, один из пакетов, необходимый для доступа к LLM (в данном случае, конкретного провайдера). В марте 2026 из-за компрометации сканера уязвимостей Trivy вредоносный код оказался внедрен в популярный AI-гейтвей LiteLLM. Атака в случае успеха приводила к краже учетных данных.

Продукт

Kaspersky Automated Security Awareness Platform

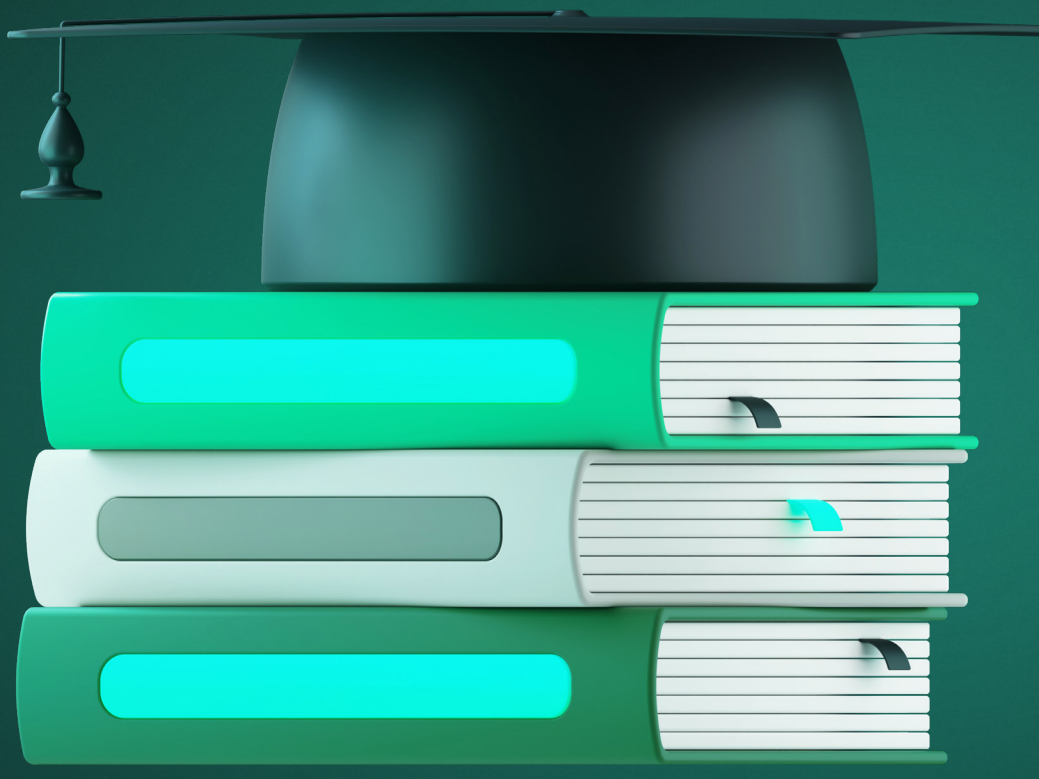
3 Проведено ли обучение сотрудников?

Пользователи, даже технически грамотные разработчики, не всегда полностью осведомлены об особенностях LLM, которые отличают их от традиционных программных продуктов. В частности, LLM подвержены галлюцинациям, что означает, что любые данные, полученные от них, требуют проверки в первоисточнике, а потенциально деструктивные действия основанных на LLM агентов — подтверждения человеком. Агенты должны запускаться в изолированной среде с минимальными привилегиями и отключенной телеметрией, а сами исполняемые файлы агентов должны браться из доверенного источника.

Кроме того, в случае с ассистентами, формат общения в виде чата может вызывать необоснованные ожидания, связанные с конфиденциальностью диалога, в то время как на деле переписки с чат-ботами могут изучаться владельцами сервиса и даже использоваться для обучения модели.

Кейсы

В одном из наиболее ранних LLM-инцидентов компания обнаружила, что ее сотрудники бесконтрольно отправляют конфиденциальный код в ChatGPT. Ущерб, к сожалению, может не ограничиваться утечками — из-за неверных действий агентов компании в рамках нескольких отдельных случаев сталкивались с потерей данных. Обучение пользователей правильному использованию LLM поможет им принять правильное решение в сложной ситуации.



4 Правильно ли выбрана модель доступа к LLM?

Продукты

- Endpoint Security
- Kaspersky Container Security
- Kaspersky Cloud Access Security Broker Data Feed

ИИ-инструменты разработчика требуют огромного объема вычислений. В зависимости от уровня критичности и конфиденциальности кода, над которым работает сотрудник, эти вычисления могут выполняться в облаке или в контуре компании. Минимальным требованием для менее критичного кода являются централизованные подписки от авторитетного провайдера, предоставляющего такой функционал, как централизованный контроль аккаунтов, SSO и zero data retention-гарантии. Это поможет защитить от ситуаций, когда из-за компрометации личного аккаунта разработчика корпоративные данные выставляются на продажу в даркнете.

Если же к коду или инфраструктуре применяются более жесткие требования (например, регуляторные), доступ может осуществляться по on-premise-модели: выделенная команда или интегратор разворачивают GPU-инфраструктуру в контуре компании, поверх которой разворачиваются модели. В таком случае критичной становится защита этих серверов, контейнерной инфраструктуры, а также вопрос безопасности самих моделей. Существуют также и гибридные варианты — от запуска моделей прямо на устройствах разработчиков (допустимо для небольших autocomplete-моделей) до аренды GPU-мощностей в private cloud; каждый из вариантов имеет свои риски, которые нужно принимать во внимание.

Кейсы

Команда Digital Footprint Intelligence «Лаборатории Касперского» обнаружила в даркнете объявления о продаже доступа к аккаунтам OpenAI, включая корпоративные. Доступ к аккаунту подразумевал доступ и к истории переписок.

5 Правильно ли реализована модель доступа к корпоративному контексту?

Для того, чтобы ИИ-ассистенты были максимально эффективны, они должны иметь доступ к необходимому рабочему контексту: кодовым репозиториям, трекеру задач, корпоративной вики, документации и другим базам знаний. При этом все эти источники данных могут иметь свои ролевые модели доступа, которым агент, запускаемый от имени того или иного разработчика, должен подчиняться, наследуя права пользователя. К сожалению, этим правилом часто пренебрегают, как результат, инструмент для доступа к контексту работает с правами создателя этого инструмента или и вовсе администратора с полным доступом. Это означает, что имея доступ к скомпрометированной УЗ разработчика, злоумышленник имеет потенциал получить доступ к привилегированным данным.

Кейсы

В 2025 году сообщалось о случае, когда чат-бот-копайлот бухгалтерской фирмы имел полный доступ к базе данных, не реализуя ролевую модель доступа. В результате этого пользователи могли получить доступ к чужой отчетности.

Продукт

- Kaspersky Symphony XDR
- Kaspersky Vulnerability Management

6 Реализованы ли LLM-специфичные меры безопасности?

LLM-системы, кроме традиционных методов защиты, требуют дополнительных средств, которые обеспечивают безопасное взаимодействие с ними разработчиков и работу базирующихся на них агентов. В первую очередь речь идет об observability: постоянном трекинге, записи и анализе действий агентов и запросов к LLM. Эти действия могут записываться как в SIEM, так и в специализированные средства LLM observability, такие как Langfuse или Arize Phoenix. Кроме того, для противодействия утечкам данных, например, персональных (при использовании облачной модели доступа), а также защиты от промпт-инъекций, могут применяться решения класса AI Firewall, которые реализуют защиты (guardrails), снижающие вероятность успешных атак. Наконец, in-house инструменты должны следовать лучшим практикам защиты от таких атак на уровне дизайна. Общий уровень защищенности при этом поможет оценить ИИ-специфичное тестирование на проникновение (AI Redteaming).

Кейсы

В конце 2025 года исследователи продемонстрировали, что промпт-инъекция в файле в кодовом репозитории на платформе GitHub может привести к утечке приватных данных (например, токены) на серверы злоумышленника, причем агент в ходе атаки принимает активное участие в поиске и эксфильтрации секретов.

Продукты

- Kaspersky Unified Monitoring and Analysis Platform
- Kaspersky Security Services
- Endpoint Security
- Kaspersky Vulnerability Management

7 Есть ли единый репозиторий доверенных моделей, навыков, MCP-серверов?

Важной чертой современных агентов является их простая расширяемость: MCP-серверы (Model Context Protocol) дают им нативный доступ к API сторонних и внутренних сервисов, а навыки (Agent Skills) позволяют им легко адаптироваться к новым задачам и инструментам. К сожалению, обратной стороной их доступности и простоты использования является высокий риск, что очередной полезный агентный навык окажется вредоносным — та же проблема, которая знакома всем по плагинам для браузеров и IDE. Еще одним вектором атаки являются файлы моделей: загрузка недоверенных файлов моделей может привести к исполнению произвольного кода, внедренного в эту модель злоумышленником. Это особенно важный риск, если команды или индивидуальные разработчики в принятой модели имеют право разворачивать модели под собственные нужды. Чтобы предотвратить негативные последствия, необходимо поддерживать единый централизованный репозиторий одобренных моделей, навыков и других артефактов, проходящих сканирование на наличие угроз.

Продукты

- Kaspersky Endpoint Detection and Response
- Kaspersky Scan Engine

Кейсы

Исследователи обнаружили, что некоторые модели на самом популярном репозитории открытых моделей Huggingface имеют скрытые закладки кода, который исполняется при загрузке моделей, приводя, например, к получению злоумышленником доступа к устройству. Кроме того, в популярном репозитории агентных навыков обнаруживались файлы навыков с вредоносным кодом.

8 Организованы ли накопление экспертизы по безопасности ИИ, обмен опытом и лучшими практиками?

Этот наименее технический из всех вопросов подразумевает создание организационной культуры, в которой сочетаются прозрачность, заинтересованность ключевых лиц, принимающих решения, и наличие ресурсов на обеспечение безопасности ИИ у задействованных в трансформации департаментов. В связи с высоким интересом к технологиям ИИ необходимо отслеживать изменения в законодательстве и текущие требования регуляторов. Параллельно следует изучать текущие лучшие практики в оценке рисков и обеспечении безопасности агентных систем, оценивать применимость фреймворков, таких как OWASP Top-10 for Agentic Security. Для эффективности этого процесса можно создать единый центр компетенций в области безопасности ИИ, задача которого — транслировать экспертизу в рабочие команды и процессы: методология, аудит, помощь при инцидентах. Для повышения уровня вовлеченности команд может пригодиться роль AI Security-чемпионов.

Продукт

- Регуляторный хаб знаний в области информационной безопасности

Кейсы

В США с 2022 года идет судебное разбирательство против GitHub, связанное с законностью использования открытого кода для обучения моделей для разработки. Исход этого иска может повлиять на всю текущую инфраструктуру разработки с применением ИИ.

Заключение

Внедрение ИИ в процессы разработки — это сложный процесс, который может принести ощутимые преимущества, такие как повышение продуктивности и уменьшение time-to-market, но который также привносит и риски, которые местами отличаются от традиционных. Чтобы их минимизировать, необходимо, во-первых, не пренебрегать традиционными мерами защиты, во-вторых, адаптировать процессы безопасности под новые вызовы, в-третьих — поставить во главу угла пользователей, их обучение и вовлеченность, а также процессы накопления и распространения знаний о безопасном применении ИИ. Приведенные выше вопросы не являются исчерпывающим гайдом по обеспечению безопасности ИИ-разработки, но мы надеемся, что ответы на них позволят обнаружить некоторые белые пятна в безопасности, запустить необходимые процессы и сделать внедрение ИИ успешным и безопасным.

www.kaspersky.ru

#kaspersky
#активируйбудущее