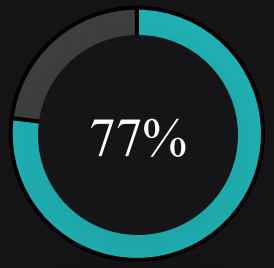
INFERA AI.Firewall

защита ИИ-моделей и их безопасное использование

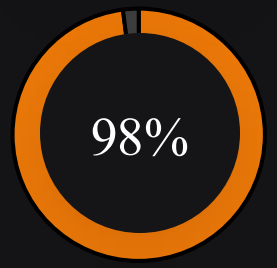
решение для защиты и безопасного использования LLM-моделей, включая контроль доступа и защиту конфиденциальных данных, фильтрацию вредоносных запросов и аудит взаимодействия с ИИ

Безопасность ИИ



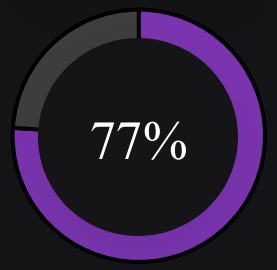
компаний сообщили о взломах своих ИИ-систем и утечку персональных данных за последний год

AI Threat Landscape Report 2024



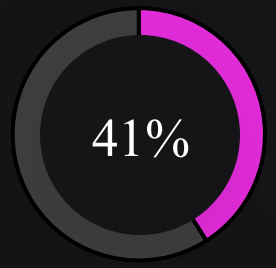
компаний считают свои ИИ-модели критически важными для бизнеса

AI Threat Landscape Report 2024



столкнулись с инцидентами безопасности, связанными с ИИ, в течение последнего года

AI Threat Landscape Report 2024



уже сталкивались с инцидентами безопасности или утечками конфиденциальных данных при работе с AI/ML

Gartner

среди наиболее частых проблем:

- подмена или отравление обучающих данных (data poisoning attacks) — до 30%
- некорректное использование приватных данных клиентов в датасетах
- утечки информации через модель или её API

кто игнорирует безопасность ИИ, рискует столкнуться не только с киберугрозами, но и с репутационными и юридическими последствиями

Gartner

INFERA AI.Firewall

система защиты для мониторинга, фильтрации и контроля запросов и ответов, обрабатываемых ИИ-моделями (LLM, ML API, генеративный ИИ)

интеграция через API

— устанавливается между приложением и моделью

анализ запросов в реальном времени

— анализирует все запросы, находит в них конфиденциальные и персональные данные, к которым пользователь не должен иметь доступ

ИИ-подход

— LLM внутри INFERA AI.Firewall распознает сложные паттерны и отыскивает сложные взаимосвязи

результат

— безопасное и контролируемое использование искусственного интеллекта в продуктах и сервисах

— если в результате выдачи модели будет присутствовать конфиденциальная информация, то она будет маскирована или доступ к запросу будет запрещен

INFERA AI.Firewall. Схема работы

1. запрос от пользователя проходит через INFERA AI.Firewall
2. система анализирует и фильтрует его, маскируя конфиденциальную информацию, блокируя опасные или запрещённые данные
3. запрос передается LLM только после прохождения всех проверок
4. ответ LLM также проходит проверку перед отправкой пользователю
5. все этапы фиксируются в аудит-логе

данные компании

обучение на всех данных компании, включая конфиденциальные

Внутренняя
Внешняя
LLM модель

запрет доступа

«Вы не имеете доступ к данной информации, обратитесь к администратору» или

ответ с маскированными данными

«Клиент [Клиент_1] (паспорт [Документ_1]) запрашивает кредитную историю по счету [Счет_1]»

данные о правах доступа

системы учёта и управления доступом (AD, keylock,…)

пользователь

сформированный prompt

«выдай финансовые данные Иванова И.И.»

Утечка данных в LLM и пример маскирования

какой ответ может получить пользователь LLM	ответ при маскировании данных
Клиент Иванов И.И., паспорт 1234 567896, выдан УВД г. Москвы, запрашивает кредитную историю по счету № 40702810123450000123	**Клиент [Клиент_15], {паспорт [Документ_12_15]} запрашивает кредитную историю по счету [Счет_1]**
Пациент Петров А.В. , полис ОМС №1234567890123456, с диагнозом J45.0 нуждается в консультации	Пациент [Пациент_103] , {полис ОМС [Полис 103_02]}, с диагнозом [Диагноз_1] нуждается в консультации
Логин: admin@company.ru Пароль: P@ssw01rd123 Сервер: 192.168.1.1 Доступ к репозитарию: https://github.com/company/project	Логин: [Email_178] Пароль: [Пароль_178] Сервер: [IP_45] Доступ к репозитарию: [URL_7]

какие данные подлежат маскированию

личные данные (ПДн)
- ФИО
- дата рождения
- профессия
- телефон
- почта
- адрес

конфиденциальная информация
- данные о клиентах и поставщиках
- информация о зарплатах и бонусах
- учетные записи

документы
- паспорт
- ИНН
- номер ОМС
- номер ПФР
- номер счета страхователя
- паспорт ТС
- данные по недвижимости

финансовая информация
- номер банковской карты
- срок действия и CVV/CVC
- счет в банке
- SWIFT/BIC-коды
- информация о транзакциях
- доходы, задолженности, налоговые сведения
- номер договора
- суммы на счете

медицинские данные
- диагноз
- результаты исследований
- номер ОМС и ДМС
- заключения врачей
- назначения, назначения лечения и схемы терапии
- данные о психическом здоровье
- факт наличия инвалидности

INFERA AI.Firewall. Возможности решения

- маскирование данных
- управление доступом к данным, выдаваемым от LLM модели для разграничения доступа пользователей к информации на основе правил доступа к их источнику/типу данных
- интеграция с LLM по принципу proxy
- настройка через конфигурационные файлы: пользователи, права, интеграция с AD/ALD, действия при выявлении недоступного файла
- автоматическое выявление критической информации в массиве данных
- гибкая настройка параметров маскирования
- автоматический анализ структуры ответов
- замена персональных данных и чувствительной информации внутри сообщения
- сохранение форматов и структуры данных, а также обратной замены

уникальные свойства

- интеграция с облачными и частными LLM моделями и агентами
- автоматическое определение типов информации
- возможность обучения на документах организации
- различные режимы реагирования – маскирование или запрет доступа
- возможность получения информации о правах доступа с распространёнными системами хранения информации (Confluence, Wiki и другие)
- возможность защиты собственных On-Premise LLM моделей от угроз утечки информации

Законодательные требования по использованию ИИ

— в условиях активного внедрения ИИ в критически важные сферы (банки, финансы, здравоохранение, госсектор) растет внимание регуляторов

— базовые требования к безопасности, этике и устойчивости моделей ИИ становятся обязательными

Центральный банк РФ готовит проект рекомендаций для финансового сектора по безопасному использованию ИИ

защита на 4х этапах и обеспечение безопасности:

— при подготовке данных

— при разработке модели ИИ

— при обучении и тестировании модели ИИ

— при функционировании модели ИИ

законодательные инициативы и рекомендации

— Проект от ЦБ РФ: рекомендации по управлению ИИ-моделями в банках

— ГОСТ Р ИСО/МЭК 23894-2023: принципы ИИ-этики и управления рисками

— Будущий ГОСТ «Об ИИ» в РФ — в разработке

— Международные аналоги: AI Act (ЕС), NIST AI RMF (США), ISO/IEC 42001

новый стандарт по защите ИИ

напишите нам, и мы пришлем актуальную версию документа

INFERA SECURITY

безопасность встроенная в ДНК

спасибо за внимание

www.InfEraSecurity.ru

тел.: +7 915 234 9900

почта: info@InfEraSecurity.ru