OFFENSIVE SECURITY

КОМПЛЕКСНЫЙ АНАЛИЗ ЗАЩИЩЕННОСТИ ПРИЛОЖЕНИЙ, ИСПОЛЬЗУЮЩИХ AI

КЛЮЧЕВЫЕ ОБЛАСТИ АНАЛИЗА

- Анализ архитектуры и площади атаки: от пользовательского интерфейса и API до конкретных агентов в комплексных мультиагентных системах, включая LLM, интеграции и защитные механизмы
- Оценка реализуемости специфических угроз - Prompt Injection, манипуляция агентами и другие (покрывает OWASP Top 10 for LLM и не только)
- Анализ интеграций и прав доступа: например, проверка привилегий, которыми обладают AI-агенты при взаимодействии с внутренними и сторонними сервисами
- Поиск возможностей злоупотребления реализованной GenAI-функциональностью с целью нарушения бизнес-логики приложения и обхода заложенных ограничений

РЕАЛЬНЫЕ УГРОЗЫ ДЛЯ AI-ПРИЛОЖЕНИЙ

- **Prompt Injection**
  Злоумышленники могут перехватить управление AI-агентом через специально сконструированные запросы, получив доступ к данным и действиям от имени системы
- **Утечка конфиденциальных данных**
  LLM может раскрыть в ответах чувствительную информацию – данные пользователей, внутренние документы, API-ключи или детали инфраструктуры, включая содержимое базы данных для RAG
- **Чрезмерное потребление ресурсов**
  Чрезмерное потребление токенов в результате атакующих воздействий
- **Избыточные привилегии агентов**
  AI-агенты с чрезмерными правами доступа к инструментам и сервисам создают возможности для масштабных атак, включая использование интегрированных инструментов для получения доступа во внутреннюю инфраструктуру, проведения спам-атак и фишинга
- **Обход бизнес-ограничений**
  Контекст агента может не учитывать различные особенности заложенной бизнес-логики, чем может воспользоваться злоумышленник для обхода ограничений даже без использования специализированных полезных нагрузок

ЧТО МЫ АНАЛИЗИРУЕМ

Архитектура
Анализ архитектуры и поверхности атаки

Интеграции
Анализ подключенных интеграций и прав доступа

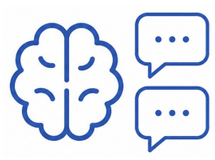
Специфические угрозы
Анализ специфических уязвимостей и сценариев атак

Бизнес-логика
Злоупотребление GenAI-функциональностью

Защитные меры
Анализ реализованных защитных механизмов

Продуктовое окружение
Инфраструктура размещения и доставки приложения

ПОЧЕМУ AI-СИСТЕМАМ НУЖЕН СПЕЦИАЛИЗИРОВАННЫЙ АНАЛИЗ

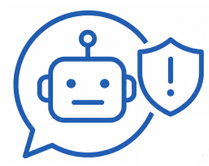


Вероятностное поведение

LLM генерируют разные ответы на одинаковые запросы – выход недетерминирован и может меняться для одинаковых входных данных

Специфические классы угроз

Prompt Injection, Jailbreak, раскрытие данных, ошибки интеграции и избыточные полномочия агентов – угрозы специфичны для систем с AI и требуют специализированных методик



Расширенная поверхность атаки

AI-агент становится самостоятельной исполнительной единицей со своими привилегиями. Дополнительно формируются системы из нескольких агентов. Значительно расширяется поверхность атаки

КОМУ НЕОБХОДИМ АНАЛИЗ AI-СИСТЕМ

- E-commerce
- Финтех и банки
- Телеком и промышленность
- SaaS и продуктовые компании
- Здравоохранение
- Государственные структуры

ЧТО ВЫ ПОЛУЧАЕТЕ

- Технический отчет
- Резюме для руководства
- Рекомендации по устранению
- Повторный тест после внесения исправлений, чтобы подтвердить их эффективность
- Консультации по устранению найденных уязвимостей и защите от AI-специфичных угроз

ЭТАПЫ ПРОВЕДЕНИЯ АНАЛИЗА

Определение границ

Согласование объема работ, идентификация AI-компонентов, анализ документации и архитектуры системы

Анализ архитектуры

Определение поверхности атаки: LLM-модели, агенты, интеграции, API, пользовательские интерфейсы и потоки данных

Отчет и рекомендации

Подготовка подробного отчета

Анализ реализации

Активный поиск уязвимостей в соответствии с рекомендациями и методиками OWASP для AI-систем, MITRE ATLAS и Google SAIF, тестирование интеграций, проверка эффективности защиты

Демонстрация уязвимостей и анализ рисков

Демонстрация реальных последствий найденных уязвимостей, оценка критичности и влияния на бизнес

125040, г. Москва, ул. Вавилова, д. 47А +7 (499) 705-76-57 info@solidlab.ru

Группа компаний SolidLab работает в области информационной безопасности более 15 лет, защищая организации от киберугроз, предупреждая атаки, минимизируя риски финансовых и репутационных потерь. SolidLab предлагает полный комплекс услуг: от пентестов до защиты приложений, реагирования на инциденты и организации процессов SDLC.