СБЕР
КИБЕР
БЕЗОПАСНОСТЬ

Модель угроз для кибербезопасности AI

для этапов сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с приложениями

v.1.0 (март 2025)

Общие сведения о модели угроз

Рост использования искусственного интеллекта (AI) в критически важных областях требует превентивного подхода к безопасности.

Модель угроз кибербезопасности для искусственного интеллекта, охватывает полный жизненный цикл AI-решений — от подготовки данных и разработки модели до интеграции её в приложения.

Документ систематизирует 70 угроз, классифицированных по трем этапам:

- сбор и подготовка данных,
- разработка модели и обучение,
- эксплуатация модели и интеграции с приложениями.

Для каждой угрозы представлено:

- виды моделей, для которых угроза релевантна (PredAI[1], GenAI[2]),
- описание угрозы,
- возможные последствия,
- нарушаемое свойство информации:
    - конфиденциальность,
    - целостность,
    - доступность,
    - достоверность[3].
- объект, который может быть подвергнут воздействию нарушителя

Модель угроз предназначена для специалистов, вовлеченных в создание, внедрение и управление AI-решениями: разработчиков, ответственных за разработку и обучение моделей; специалистов кибербезопасности, обеспечивающих защиту данных и ИТ-инфраструктуры; архитекторов, интегрирующих AI-решения в бизнес-процессы; а также руководителей, оценивающих риски внедрения AI в критически важные операции.

Модель угроз служит инструментом для команд, которым требуется системный подход к моделированию угроз — от этапа подготовки обучающих данных до эксплуатации моделей, и позволяет организациям не только выявлять слабые места в процессах, но и формировать превентивные меры защиты, опираясь на синтез собственного экспертного опыта команды Сбера и лучших практик по кибербезопасности AI, в том числе документов OWASP, MITRE, NIST.

---

[1] PredAI (predicative AI) — разновидность моделей искусственного интеллекта, специализирующаяся на решении конкретных прикладных задач с использованием строго структурированных данных. Обучение модели осуществляется на репрезентативных выборках с четко определёнными признаками, что позволяет ей распознавать шаблоны в новых данных, аналогичных по структуре обучающей выборке. Основная функция — имитация человеческих функций в предсказании и прогнозировании для задач с фиксированной структурой данных (например, анализ рисков, классификация изображений, прогноз спроса).

[2] GenAI (generative AI) — разновидность моделей искусственного интеллекта, которая специализируется на решении разнородных задач и генерации нового контента (текст, изображения, аудио, видео) на основе анализа как структурированных, так и неструктурированных данных. Обучение осуществляется на разнородных неструктурированных и структурированных данных. Основная функция — имитация широкого набора функций человека, генерация новых данных и прогнозирование для данных новой структуры.

[3] Достоверность — соответствие информации фактам, отсутствие субъективизма, предвзятости или смысловых искажений.

СБЕР
КИБЕР
БЕЗОПАСНОСТЬ

Оглавление

Объекты, входящие в модель угроз

На рисунке приведена общая схема объектов воздействия нарушителем на этапах сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с приложениями, на которой отмечены потенциальные угрозы.

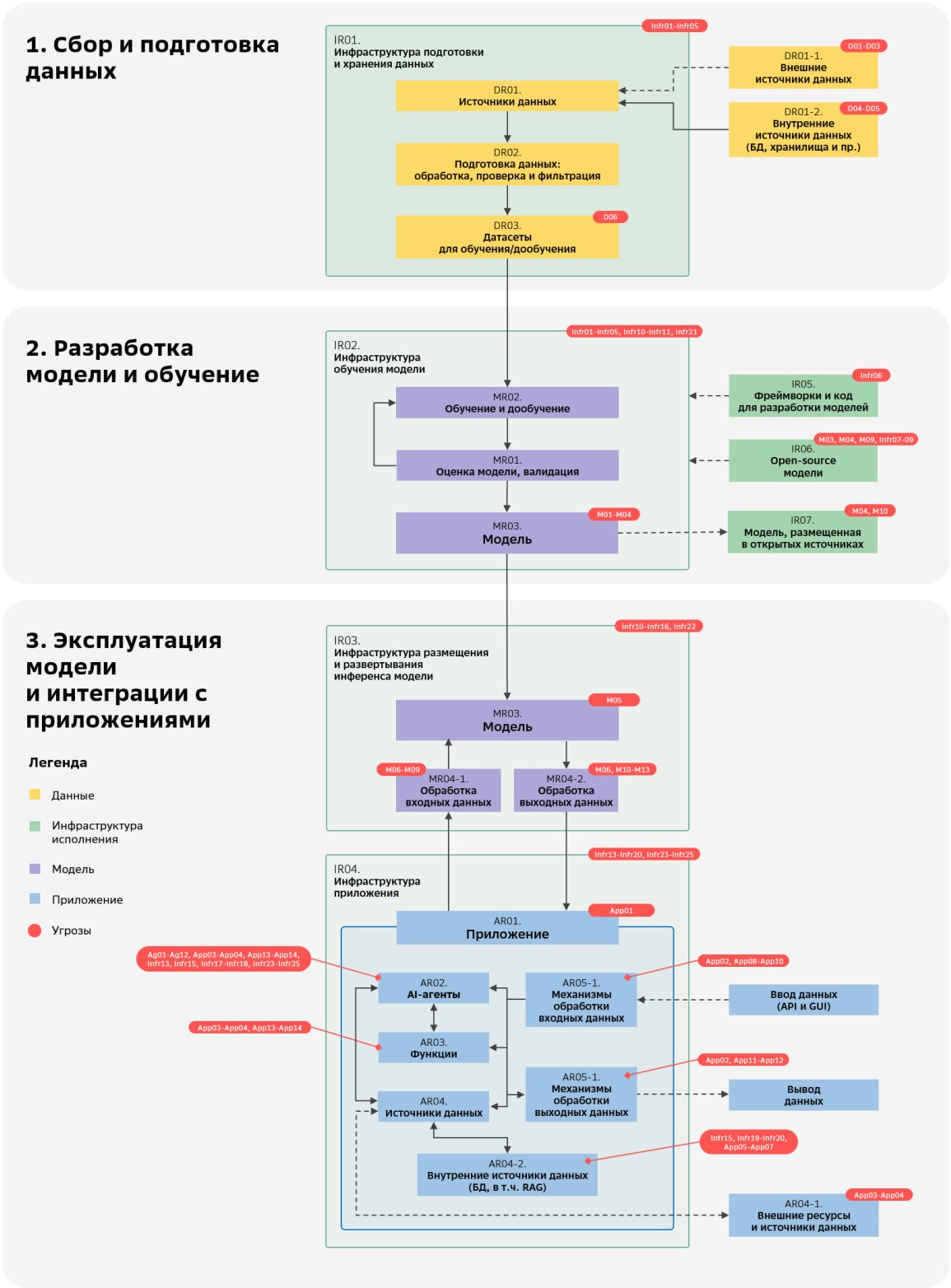

Рисунок 1 — Обобщенная схема объекта защиты и актуальных угроз на этапах сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с приложениями

Объекты на схеме:

1. Этап «Сбор и подготовка данных»

- IR01 Инфраструктура подготовки и хранения данных — совокупность программных и аппаратных средств, необходимых для сбора, обработки, хранения и управления данными на этапе подготовки к обучению моделей.
- DR01 Источники данных — данные из различных источников, используемые для обучения и дообучения моделей
    - DR01-1 Внешние источники — публичные датасеты, API, социальные сети, открытые базы данных.
    - DR01-2 Внутренние источники — корпоративные базы данных, логи пользовательского взаимодействия, исторические данные организации.
- DR02 Подготовка данных: обработка, проверка, фильтрация — действия по очистке, нормализации, преобразованию форматов и фильтрации данных для создания обучающих выборок.
- DR03 Датасеты для обучения и дообучения — наборы данных, прошедшие подготовку и используемые для обучения, тестирования и дообучения моделей.

2. Этап «Разработка модели и обучение»

- IR02 Инфраструктура обучения модели — совокупность программных и аппаратных средств, необходимых для обучения и оптимизации моделей.
- MR01 Оценка модели, валидация — действия по тестированию и валидации качества модели, включая ее безопасность.
- MR02 Обучение и дообучение — действия по применению алгоритмов машинного обучения для создания и адаптации моделей.
- MR03 Модель — итоговая модель, полученная в результате обучения.
- IR05 Фреймворки и код для разработки моделей — программные библиотеки и исходный код для создания моделей.
- IR06 Open-source модели — предобученные модели из открытых репозиториев.
- IR07 Модель, размещенная в открытых источниках — проприетарная модель, размещаемая в открытых репозиториях для использования сообществом.

3. Этап «Эксплуатация модели и интеграции с приложениями»

- IR03 Инфраструктура размещения и развертывания инференса модели — совокупность программных и аппаратных средств, необходимых для развертывания обученных моделей в production.
- MR04 Обработка входных и выходных данных — механизмы преобразования данных перед передачей в модель и после получения результатов.
    - MR04-1 Механизмы обработки входных данных — предобработка данных, поступающих на вход модели.
    - MR04-2 Механизмы обработки выходных данных — постобработка результатов модели, может включать форматирование, фильтрацию, в том числе нежелательного контента.
- IR04 Инфраструктура приложения — совокупность программных и аппаратных средств, необходимых для выполнения AI-приложений.
- AR01 Приложение — конечная система, использующая AI-модель для решения задач
    - AR02 AI-агенты — автономные системы, принимающие решения на основе выводов моделей.
    - AR03 Функции — внутренние функции, реализующие бизнес-логику приложения, или внешние по отношению к приложению функции, связанные с работой AI-модели.
    - AR04 Источники данных — данные, используемые приложением в реальном времени.

        - AR04-1 Внешние источники данных — веб-сайты, публичные базы данных, потоковые данные, поступающие из внешних источников.
        - AR04-2 Внутренние источники данных (БД, в т.ч. RAG) — локальные базы данных, векторные хранилища для RAG (Retrieval-Augmented Generation)
    - AR05 Обработка пользовательского ввода/вывода
        - AR05-1 Механизмы обработки входных данных — валидация, санитизация и предобработка пользовательских запросов.
        - AR05-2 Механизмы обработки выходных данных — форматирование ответов модели, фильтрация нежелательного контента.

Перечень угроз

Угрозы, связанные с данными

D01. Использование для обучения/дообучения модели отравленных данных или датасетов, загруженных из внешних источников

Описание
Загрузка, обучение или дообучение модели на вредоносных данных или наборах данных, в которые были внедрены вредоносные примеры, из внешних источников (например, репозиториев или страниц, с которых выполняется сбор данных)

Последствия
Модификация (искажение) модели, смещение результатов работы модели, снижение точности или создание бэкдоров в модели

Объект воздействия
DR01-1 Внешние источники данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец данных

D02. Использование для обучения/дообучения модели модифицированных данных или датасетов, загруженных из внешних источников

Описание
Использование скомпрометированного источника данных, обучение или дообучение модели на загруженных вредоносных данных или наборах данных, в которые были внедрены вредоносные примеры, из внешних источников, которые были модифицированы после добавления в перечень источников

Последствия
Модификация (искажение) модели, смещение результатов работы модели, снижение точности или создание бэкдоров в модели

Объект воздействия
DR01-1 Внешние источники данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец данных

D03. Воспроизведение в ответах модели персональных данных (ПДн), полученных из внешних источников

Описание
Загрузка, обучение или дообучение модели на данных или наборах данных, которые содержат ПДн из внешних источников

Последствия
Запоминание данных и непреднамеренное воспроизводство персональной информации, что создает риски нарушения ФЗ-152 и приватности пользователей

Объект воздействия
DR01-1 Внешние источники данных

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец данных

D04. Использование для обучения/дообучения модели отравленных данных или датасетов, загруженных из внутренних источников

Описание
Загрузка, обучение или дообучение модели на данных или наборах данных, в которые были внедрены вредоносные примеры из внутренних источников

Последствия
Модификация (искажение) модели, смещение результатов работы модели, снижение точности или создание бэкдоров в модели

Объект воздействия
DR01-2 Внутренние источники данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец данных

D05. Неконтролируемая загрузка данных, содержащих конфиденциальную информацию, в датасеты для обучения моделей

Описание
Загрузка и обучение модели на данных или наборах данных, которые содержат конфиденциальную информацию

Последствия
Запоминание данных и непреднамеренное воспроизводство конфиденциальной информации, что создает риски утечек и хищения конфиденциальной информации

Объект воздействия
DR01-2 Внутренние источники данных

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец данных

D06. Неконтролируемое использование, модификация, удаление данных для обучения или дообучения модели

Описание

Отсутствие надлежащего контроля и управления данными для обучения или дообучения модели, включая их классификацию, наличие метаданных, процессы управления жизненным циклом (включая хранение, удаление, архивирование и пр.)

Последствия

Увеличение поверхности атаки, утечки конфиденциальной информации, риски нарушения ФЗ-152

Объект воздействия

DR03 Датасеты для обучения/дообучения

Нарушаемое свойство

Конфиденциальность, целостность

Виды моделей, подверженных угрозе

PredAI и GenAI

Лица, ответственные за митигацию угрозы

Владелец данных

Угрозы, связанные с инфраструктурой

Infr01 Несанкционированная модификация реестра источников данных, датасетов

Описание
Использование скомпрометированного источника данных или датасетов вследствие несанкционированной модификации реестра источников

Последствия
Модификация (искажение) модели, смещение результатов работы модели, снижение точности или создание бэкдоров в модели

Объект воздействия
IR01 Инфраструктура хранения данных, IR02 Инфраструктура обучения модели

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры хранения данных; владелец ИТ-инфраструктуры обучения модели

Infr02 Несанкционированная модификация обучающих данных

Описание
Использование скомпрометированных данных или датасетов, используемых для обучения/дообучения модели, вследствие несанкционированной модификации

Последствия
Модификация (искажение) модели, смещение результатов работы модели, снижение точности или создание бэкдоров в модели

Объект воздействия
IR01 Инфраструктура хранения данных, IR02 Инфраструктура обучения модели

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры хранения данных; владелец ИТ-инфраструктуры обучения модели

Infr03 Небезопасная передача данных/датасетов между этапами подготовки

Описание
Перехват или модификация при передаче между этапами обработки, хранения данных или обучения модели

Последствия
Утечки данных, компрометация модели, внедрение вредоносных данных в датасеты или нарушение целостности данных

Объект воздействия
IR01 Инфраструктура хранения данных, IR02 Инфраструктура обучения модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры хранения данных; владелец ИТ-инфраструктуры обучения модели

Infr04 Кража обучающих данных

Описание
Хищение обучающих данных вследствие слабостей разграничения доступа

Последствия
Восстановление модели, создание теневой модели

Объект воздействия
IR01 Инфраструктура хранения данных, IR02 Инфраструктура обучения модели

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры хранения данных; владелец ИТ-инфраструктуры обучения модели

Infr05 Утечка конфиденциальной информации из наборов обучающих данных

Описание
Хищение конфиденциальной информации из наборов обучающих данных вследствие слабостей разграничения доступа

Последствия
Хищение конфиденциальной информации

Объект воздействия
IR01 Инфраструктура хранения данных, IR02 Инфраструктура обучения модели

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры хранения данных; владелец ИТ-инфраструктуры обучения модели

Infr06 Использование уязвимых версий сторонних библиотек, использование программного кода с закладками

Описание
Использование для разработки модели и обучения сторонних библиотек или программного кода с уязвимостями или программными закладками. Угроза обусловлена слабостями механизмов анализа библиотек и кода на наличие уязвимостей или отсутствием соответствующих проверок.

Последствия
Внедрение вредоносного кода

Объект воздействия
IR05 Фреймворки и код для разработки моделей

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Эксперт по безопасной разработке

Infr07 Использование open-source моделей, содержащих программные закладки в файлах

Описание
Возможность осуществления нарушителем деструктивного воздействия на систему путем внедрения программных закладок в файлах open-source моделей

Последствия
Внедрение вредоносного кода

Объект воздействия
IR06 Open-source модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Эксперт по безопасной разработке

Infr08 Использование open-source моделей, содержащих логические закладки, заложенные при обучении

Описание
Возможность осуществления нарушителем деструктивного воздействия на систему путём эксплуатации уязвимостей open-source моделей, содержащих логические закладки и бэкдоры

Последствия
Смещение результатов работы модели, снижение точности или эксплуатация бэкдоров в модели

Объект воздействия
IR06 Open-source модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Эксперт по безопасной разработке

Infr09 Использование open-source моделей, содержащих закладки в весах

Описание
Возможность осуществления нарушителем деструктивного воздействия на систему путём эксплуатации уязвимостей open-source моделей, содержащих закладки в весах модели

Последствия
Смещение результатов работы модели, снижение точности или эксплуатация бэкдоров в модели

Объект воздействия
IR06 Open-source модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Эксперт по безопасной разработке

Infr10 Подмена или модификация модели

Описание
Несанкционированная модификация или подмена модели вследствие слабостей разграничения доступа

Последствия
Смещение результатов работы модели, снижение точности или эксплуатация бэкдоров в модели

Объект воздействия
IR02 Инфраструктура обучения модели, IR03 Инфраструктура размещения и развертывания инференса модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры обучения модели; владелец ИТ-инфраструктуры размещения и развертывания инференса модели

Infr11 Кража модели

Описание
Хищение модели вследствие слабостей разграничения доступа

Последствия
Создание теневой модели или восстановление обучающих данных из-за наличия white-box доступа к украденной модели

Объект воздействия
IR02 Инфраструктура обучения модели, IR03 Инфраструктура размещения и развертывания инференса модели

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры обучения модели; владелец ИТ-инфраструктуры размещения и развертывания инференса модели

Infr12 Нарушение доступности модели

Описание
Атаки типа DDoS (Distributed Denial of Service), в том числе атаки на API, или эксплуатация уязвимостей инфраструктуры размещения и развертывания инференса модели или зависимых компонентов

Последствия
Прекращение или снижение скорости оказания услуг AI-сервисом всем потребителям (или группе потребителей) из-за нарушения доступности для них инфраструктуры размещения и развертывания инференса модели

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения и развертывания инференса модели

Infr13 Утечки конфиденциальной информации из систем логирования, в том числе логирования запросов и вызовов функций

Описание
Утечки информации из логов (журналов) запросов, содержащих ПДн и иную конфиденциальную информацию, в том числе об особенностях реализации AI-агентов и мультиагентных систем

Последствия
Хищение конфиденциальной информации

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели, IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения и развертывания инференса модели; владелец приложения

Infr14 Невозможность или несвоевременное выявление, реагирование и расследование событий безопасности и инцидентов из-за отсутствия логирования взаимодействий

Описание
Отсутствие или неполнота данных логирования взаимодействий (включая запросы и ответы модели, данные телеметрии, вызовы функций и пр.) между моделью и интеграциями, AI-агентами или пользователями в ИТ-инфраструктуре размещения и развертывания инференса модели

Последствия
Увеличение времени или невозможность выявления, реагирования и расследования событий безопасности и инцидентов

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели, IR04 Инфраструктура приложения

Нарушаемое свойство
Целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения и развертывания инференса модели; владелец приложения

Infr15 Перехват или подмена запросов или ответов модели или данных, передаваемых при взаимодействии с БД RAG

Описание
Перехват или подмена запросов или ответов модели путем реализации атаки man-in-the-middle (MiTM), вмешательство в процесс обмена данными между интеграциями и компонентами AI-агентами или пользователями и моделью, что позволяет выполнить перехват, изменение или подмену входных данных, отправляемых модели, или изменение выходных данных, а также ответов модели

Последствия
Перехват запросов и ответов, содержащих ПДн или конфиденциальную информацию, модификация запросов для получения некорректных предсказаний, или подмена ответов модели для введения в заблуждение. Это может привести к утечке конфиденциальной информации, компрометации системы или принятию ошибочных решений на основе искаженных данных

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели, IR04 Инфраструктура приложения, AR02 AI-агенты, AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения и развертывания инференса модели

Infr16 Несанкционированное отключение или модификация механизмов фильтрации или контроля входных и выходных данных

Описание
Отключение или изменение систем, предназначенных для проверки и очистки данных, поступающих в модель или возвращаемых потребителю

Последствия
Обработка вредоносных или некорректных входных данных, сбои в работе модели, возвращение потребителю нежелательного, опасного или конфиденциального контента

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели, IR04 Инфраструктура приложения

Нарушаемое свойство
Целостность, конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения и развертывания инференса модели

Infr17 Хищение системного промпта

Описание
Получение доступа к системным промптам, которые управляют поведением модели, приложения или AI-агента, с целью их кражи для последующего анализа, применения для копирования особенностей функционирования модели, приложения или AI-агента

Последствия
Раскрытие конфиденциальной информации, облегчение проведения промпт-атак, утрата конкурентных преимуществ

Объект воздействия
IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения

Infr18 Несанкционированная модификация системного промпта

Описание
Получение доступа к системным промптам, которые управляют поведением модели, приложения или AI-агента, с целью их модификации для манипуляции выводом модели, функционированием приложения или AI-агента

Последствия
Нарушение функциональности модели, приложения или AI-агента, в том числе некорректные, вредоносные или нежелательные генерации, или раскрытие конфиденциальной информации

Объект воздействия
IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения

Infr19 Несанкционированная модификация данных во внутренних источниках данных (в т.ч. в БД RAG)

Описание
Доступ к внутренним хранилищам и изменение данных, используемых при работе модели, для внедрения вредоносной информации, включая непрямые промпт-инъекции или некорректную/неэтичную информацию

Последствия
Искажение результатов работы модели, нарушение целостности ответов, некорректные или вредоносные ответы, генерации некорректных инструкций, нарушение принятия решений

Объект воздействия
IR04 Инфраструктура приложения, AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения; владелец ИТ-инфраструктуры хранения данных

Infr20. Утечки информации из внутренних источников данных

Описание
Несанкционированное копирование, передача или раскрытие конфиденциальной информации или ПДн, хранящихся в базах данных, файлах или других внутренних источниках данных

Последствия
Утечка конфиденциальной информации, финансовые потери, репутационный ущерб и юридические последствия

Объект воздействия
IR04 Инфраструктура приложения, AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения; владелец ИТ-инфраструктуры хранения данных

Infr21. Несанкционированная модификация тестовых и валидационных датасетов

Описание
Несанкционированное изменение тестовых и валидационных датасетов, добавление или удаление данных для искажения данных о качестве работы модели

Последствия
Внедрение ненадежной модели в эксплуатацию, утечки данных, некорректные предсказания и генерации

Объект воздействия
IR02 Инфраструктура обучения модели

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры обучения модели

Infr22. Несанкционированные подключения к модели

Описание
Подключение и обращение к модели с использованием скомпрометированных учетных данных или через системы-посредники, проксирующие системы или шлюзы с использованием их учетных данных

Последствия
Невозможность или несвоевременное выявление, реагирование и расследование событий безопасности, DoW[4] системы-посредника

Объект воздействия
IR03 Инфраструктура размещения и развертывания инференса модели

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

Infr23. Утечки данных AI-агента или информации об особенностях его реализации

Описание
Несанкционированное копирование, передача или раскрытие информации, связанной с AI-агентом, в том числе о его цели, описании функций, инструкциях механизма планирования, содержимом памяти.

Последствия
Нарушение конфиденциальности цели, описания функций, содержимого памяти или инструкций механизма планирования AI-агента

Объект воздействия
IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения; владелец приложения

Infr24. Несанкционированная модификация AI-агента

Описание
Несанкционированная модификация AI-агента через добавление промпт-атак в цель, описание функций, память или механизм планирования AI-агента, добавление вредоносных функций в список доступных функций, добавление в список доступных ресурсов AI-агента записей, содержащих ресурсы с промпт-атаками, добавление промпт-атак в память AI-агента

Последствия
Нарушение функциональности AI-агента или приложения его реализующего

Объект воздействия
IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность, целостность, доступность

---

[4] DoW (Denial of wallet) — отказ в обслуживании вследствие исчерпания лимитов на затраты по потреблению ресурсов модели.

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец ИТ-инфраструктуры размещения приложения; владелец приложения

Infr25. Утечка информации об архитектуре мультиагентной системы через интерфейсы инструментов разработки или взаимодействия пользователя с AI-агентом или мультиагентной системой

Описание
Извлечение информации о мультиагентной системе (включая ее архитектуру, состав, правила взаимодействия) из интерфейсов инструментов разработки или взаимодействия пользователя с AI-агентом или мультиагентной системой

Последствия
Нарушение конфиденциальности состава мультиагентной системы и устройства взаимодействия между AI-агентами в мультиагентной системой, облегчение проведения кибератак

Объект воздействия
IR04 Инфраструктура приложения, AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Угрозы, связанные с моделью

M01. Невозможность реагирования и расследования событий безопасности и инцидентов из-за отсутствия информации о данных, на которых выполнено обучение модели

Описание
Отсутствие документации и прозрачности в отношении данных, использованных для обучения модели. Угроза создает риски безопасности и соответствия нормам законодательства РФ и наличия потенциальных смещений (bias) или уязвимостей к специфичным атакам (состязательным и промпт-атакам)

Последствия
Невозможность или несвоевременное выявление, реагирование и расследование событий безопасности

Объект воздействия
MR03 Модель

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Разработчик модели

M02. Использование модели с высокой уязвимостью к состязательным атакам (в том числе промпт-атакам)

Описание
Недостаточная подготовка или валидация модели, приводящая к возможности осуществления нарушителем деструктивного воздействия на систему путём эксплуатации уязвимостей модели. Обусловлена слабостями самой модели и слабостями механизмов обработки входных данных. Реализуется при использовании специально подобранных входных данных (например, изображений, текстов или запросов) для эксплуатации уязвимостей модели

Последствия
Некорректное или недекларированное поведение модели, утечки конфиденциальной информации

Объект воздействия
MR03 Модель

IR06 Open-source модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Разработчик модели

M03. Нежелательное поведение, вредоносные генерации, галлюцинации

Описание
Недостаточная подготовка или валидация модели, приводящая к некорректному или недекларированному поведению модели, вредоносным генерациям или галлюцинациям. Выявленные злоумышленником галлюцинации (наименования программных пакетов, URL-адреса, имена организаций или электронные адреса) могут быть использованы для создания вредоносных целей, таких как поддельные программные пакеты, фишинговые сайты или фальшивые организации, которые затем публикуются и распространяются

Последствия
Некорректное или недекларированное поведение модели

Объект воздействия
MR03 Модель

Нарушаемое свойство
Целостность, достоверность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Разработчик модели

M04. Подбор атак с использованием знаний уровня white-box об open-source модели

Описание
Подбор атак на целевую модель с использованием знаний уровня white-box (полный доступ к архитектуре, параметрам и обучающим данным) об уязвимости open-source моделей (моделей сторонних разработчиков и проприетарных моделей при их размещении в открытом доступе)

Последствия
Уязвимость модели к состязательным атакам и промпт-атакам

Объект воздействия
MR03 Модель, IR06 Open-source модели, IR07 Модель, размещенная в открытых источниках

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Эксперт по безопасной разработке

M05. Отсутствие информации об инференсах модели

Описание
Отсутствие или неактуальность информации о состоянии инференсов моделей

Последствия
Невозможность или несвоевременное выявление, реагирование и расследование событий безопасности, уязвимость интеграций к промпт-атакам

Объект воздействия
MR03 Модель

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M06. Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели

Описание
Обнаружение способов обхода или нарушения работы механизмов обработки входных или выходных данных, включая механизмы санитизации, валидации и фильтрации, что позволяет внедрять вредоносные данные, манипулировать результатами или получать несанкционированный доступ к информации. Угроза связана с отсутствием или недостаточностью механизмов обработки входных/выходных данных, реализованных на уровне модели

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности ответов или предсказаний

Объект воздействия
MR04-1 Механизмы обработки входных данных, MR04-2 Механизмы обработки выходных данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M07. Нарушение доступности модели (DoS) из-за отсутствия единого контроля запросов на уровне модели

Описание
Использование специально подобранных входных данных (например, вычислительно сложных) или отправка большого количества запросов, специально созданных для максимальной нагрузки на среду развертывания инференса модели

Последствия
Прекращение или снижение скорости оказания услуг AI-сервисом всем потребителям (или группе потребителей) или увеличение затрат на эксплуатацию из-за перегрузки AI-сервиса или модели запросами

Объект воздействия
MR04-1 Механизмы обработки входных данных

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M08. Исчерпание лимитов интеграции (DoW) из-за отсутствия единого контроля запросов на уровне модели

Описание
Направление чрезмерного количества запросов от интеграции к модели, что приводит к повышенному использованию токенов и лимитов

Последствия
Прекращение работы интеграции или увеличение затрат на эксплуатацию

Объект воздействия
MR04-1 Механизмы обработки входных данных

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M09. Обход встроенных защитных механизмов модели в том числе с использованием методов состязательных атак и промпт-атак

Описание
Обход встроенных защитных механизмов модели или механизмов, основанных на технологиях искусственного интеллекта

Последствия
Некорректное или недекларированное поведение модели

Объект воздействия
MR04-1 Механизмы обработки входных данных

IR06 Open-source модели

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Разработчик модели

Владелец инфраструктуры размещения, подключения и предоставления контента

M10. Утечка информации о модели

Описание
Использование различных методов (например, повторяющихся запросов или анализа документации и файлов модели, размещенных в открытых источниках) для получения информации о модели, её онтологии, семействе модели, границах принимаемых решений, а также о связанных артефактах, таких как данные, программное обеспечение и инфраструктура

Последствия
Утечка конфиденциальной информации, облегчение проведения состязательных и промпт-атак

Объект воздействия
MR04-2 Механизмы обработки выходных данных, IR07 Модель, размещенная в открытых источниках

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M11. Утечки конфиденциальной информации из дообученной модели или LoRA

Описание
Использование специально подобранных входных данных (например, с техниками джейлбрейков) для извлечения конфиденциальной информации из дообученной модели или LoRA

Последствия
Утечка конфиденциальной информации

Объект воздействия
MR04-2 Механизмы обработки выходных данных

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Разработчик модели; владелец инфраструктуры размещения, подключения и предоставления контента

M12. Эксфильтрация, инверсия или реверс-инжиниринг модели

Описание
Отправка многократных запросов к модели через API доступа к модели, анализ и изучение ответов для создания функциональной копии модели, воссоздающей поведение и функциональность целевой модели без прямого доступа к её архитектуре или обучающим данным, реализация атак инверсии (Invert ML Model) или извлечения (Extract ML Model) модели

Последствия
Кража модели, создание теневой модели

Объект воздействия
MR04-2 Механизмы обработки выходных данных

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
PredAI

Лица, ответственные за митигацию угрозы
Владелец инфраструктуры размещения, подключения и предоставления контента

M13. Эксфильтрация данных

Описание
Отправка многократных запросов к модели через API доступа к модели, анализ ответов для извлечения конфиденциальной информации или обучающих данных, реализация атак на определение принадлежности данных (Infer Training Data Membership)

Последствия

Восстановление фрагментов обучающих данных, которые могут включать ПДн или конфиденциальную информацию

Объект воздействия

MR04-2 Механизмы обработки выходных данных

Нарушаемое свойство

Конфиденциальность

Виды моделей, подверженных угрозе

PredAI и GenAI

Лица, ответственные за митигацию угрозы

Владелец инфраструктуры размещения, подключения и предоставления контента

Угрозы, связанные с приложениями

App01. Ошибки в проектировании, использование небезопасных интеграций компонентов

Описание
Использование небезопасных интеграций функциональных компонентов (например, AI-агентов, функций, плагинов) приложений, в том числе отсутствие проверки всех входных/выходных данных, использование небезопасных API, отсутствие шифрования данных в процессе передачи, некорректная настройка прав доступа

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности ответов или предсказаний, некорректное или недекларированное поведение модели

Объект воздействия
AR01 Приложение

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App02. Обход механизмов обработки входных/выходных данных, реализуемых на уровне приложения

Описание
Обнаружение способов обхода или нарушения работы механизмов обработки входных или выходных данных, включая механизмы санитизации, валидации и фильтрации. Это позволяет внедрять вредоносные данные, манипулировать результатами или получать несанкционированный доступ к информации. Угроза связана с отсутствием или недостаточностью механизмов обработки входных/выходных данных, реализованных в приложении

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности ответов или предсказаний, некорректное или недекларированное поведение модели

Объект воздействия
AR05-1 Механизмы обработки входных данных, AR05-2 Механизмы обработки выходных данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App03. Загрузка вредоносного программного обеспечения (ВПО) из внешних источников (Интернет)

Описание
Использование специально созданных источников данных (в том числе баз данных, репозиториев кода или веб-сайтов), на которых размещается вредоносный код. Возможна реализация в условиях наличия функционала поиска по внешним источникам и выполнения кода

Последствия
Выполнение вредоносной нагрузки, которая может повлечь дальнейшие нарушения безопасности

Объект воздействия
AR04-1 Внешние источники данных, AR02 AI-агенты, AR03 Функции

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App04. Загрузка отравленных данных из внешних источников (Интернет)

Описание
Использование специально созданных источников данных (в том числе баз данных или веб-сайтов), на которых размещены данные (файлы, текст, мультимедиа), содержащие непрямые промпт-инъекции. Возможна реализация в условиях наличия функционала поиска по внешним источникам

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности ответов или предсказаний, некорректное или недекларированное поведение модели

Объект воздействия
AR04-1 Внешние источники данных, AR02 AI-агенты, AR03 Функции

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App05. Внедрение непрямых промпт-инъекций во внутренние источники (в т.ч. БД RAG)

Описание
Использование специально подобранных входных данных для внедрения непрямых промпт-инъекций во внутренние источники. Возможна реализация в условиях сохранения результатов обработки данных моделью во внутренние источники

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности ответов или предсказаний, некорректное или недекларированное поведение модели

Объект воздействия
AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App06. Утечки информации из внутренних источников (в т.ч. БД RAG)

Описание
Использование специально подобранных входных данных (например, с техниками джейлбрейков) для извлечения конфиденциальной информации из внутренних источников (в т.ч. БД RAG)

Последствия
Утечка конфиденциальной информации

Объект воздействия
AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App07. Выполнение вредоносных инструкций, созданных моделью

Описание
Использование специально подобранных входных данных для формирования вредоносных инструкций, направляемых во внутренние источники (в т.ч. в БД SQL). Возможна реализация в условиях формирования моделью запросов ко внутренним источникам

Последствия
Несанкционированная модификация или уничтожение информации во внутренних источниках, утечка конфиденциальной информации

Объект воздействия
AR04-2 Внутренние источники данных (БД, в т.ч. RAG)

Нарушаемое свойство
Конфиденциальность, целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App08. Реализация прямых промпт-инъекций из-за отсутствия контроля входных данных

Описание
Использование специально подобранных входных данных для проведения промпт-атак

Последствия
Некорректное или недекларированное поведение модели, утечка конфиденциальной информации

Объект воздействия
AR05-1 Механизмы обработки входных данных

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App09. Нарушение доступности (DoS/DoW) интеграции

Описание
Использование специально подобранных входных данных или чрезмерного количества запросов от интеграции к модели, что приводит к DoS интеграции или повышенному использованию токенов и лимитов DoW

Последствия
Прекращение или снижение скорости оказания услуг интеграции с AI-сервисом или увеличение затрат на эксплуатацию интеграции

Объект воздействия
AR05-1 Механизмы обработки входных данных

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App10. Нарушение логики выполнения задачи из-за отсутствия контроля входных данных

Описание
Использование специально подобранных входных данных (состязательных атак или промпт-атак), чтобы обойти инструкции, заложенные в системном промпте, ограничения или иные системные настройки

Последствия
Изменение логики функционирования, выполнение недекларированных действий, которые были явно запрещены или не предполагались при разработке

Объект воздействия
AR05-1 Механизмы обработки входных данных

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
PredAI и GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App11. Утечка информации о системном промпте из-за некорректной обработки выходных данных

Описание
Использование специально подобранных входных данных для получения информации об используемом системном промпте

Последствия
Утечка конфиденциальной информации, облегчение проведения промпт атак

Объект воздействия
AR05-2 Механизмы обработки выходных данных

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App12. Токсичная или вредоносная генерация из-за некорректной обработки выходных данных

Описание
Использование специально подобранных входных данных для получения токсичной генерации (противоречащей законодательным и морально-этическим нормам) или вредоносной генерации (опасные инструкции, инструкции по подготовке киберпреступлений, генерации вредоносного кода, уязвимого кода)

Последствия
Некорректное или недекларированное поведение приложение вследствие генераций модели

Объект воздействия
AR05-2 Механизмы обработки выходных данных

Нарушаемое свойство
Целостность, достоверность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App13. Вывод информации о среде

Описание
Использование специально подобранных входных данных для получения информации о конфигурации системы, внутренних процессах, API-ключах, версиях программного обеспечения или других деталях, которые могут быть использованы для дальнейших атак

Последствия
Утечка конфиденциальной информации, облегчение проведения кибератак

Объект воздействия
AR02 AI-агенты, AR03 Функции

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

App14. Автоматическое распространение вредоносной инструкции на другие приложения

Описание
Использование специально подобранных входных данных для реализации самовоспроизводящихся промпт-атак. Угроза может быть реализована в случае сохранения результатов обработки данных во внутренние источники (в том числе на базе RAG) или при передаче результатов обработки данных в другие приложения

Последствия
Реализации кибератак на другие приложения, выполнение вредоносной нагрузки (вредоносные запросы распространяются на другие модели или приложения)

Объект воздействия
AR02 AI-агенты, AR03 Функции

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Угрозы, связанные с AI-агентами

Ag01. Ошибки в проектировании AI-агентов и мультиагентных систем

Описание
Использование небезопасных интеграций AI-агентов, функций и плагинов, используемых AI-агентами, в том числе отсутствие проверки всех входных/выходных/промежуточных шагов выполнения, некорректная настройка прав доступа

Последствия
Утечки конфиденциальной информации, нарушение доступности, нарушение целостности выполнения задач, некорректное или недекларированное поведение AI-агентов или мультиагентных систем

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag02. Вредоносные генерации в ответе AI-агента на запрос пользователя

Описание
Использование специально подобранных входных данных для генерации токсичного или вредоносного ответа

Последствия
Некорректное или недекларированное поведение AI-агента вследствие генераций модели, риск репутационного ущерба

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Достоверность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag03. Отправка информации из среды исполнения функций AI-агента (действий) на внешние ресурсы

Описание
Использование специально подобранных входных данных для вызова функций для размещения на внешних ресурсах конфиденциальной информации (в том числе значений переменных среды)

Последствия
Утечка конфиденциальных данных, содержащихся в среде исполнения, облегчение проведения кибератак

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag04. Удаление или модификация файлов в среде исполнения функций AI-агента (действий)

Описание
Использование специально подобранных входных данных для вызова функций, удаляющих или модифицирующих файлы, доступные в среде исполнения

Последствия
Нарушение целостности данных, содержащихся в среде исполнения

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag05. Размещение в среде исполнения функций AI-агента (действий) файлов с ВПО, полученных с внешних ресурсов

Описание
Использование специально подобранных входных данных для вызова функций (загружающих из внешних источников файлы ВПО и исполняющих его), подмена легитимных ресурсов или внедрение ВПО во внешние ресурсы, используемые AI-агентами

Последствия
Выполнение вредоносной нагрузки, которая может повлечь дальнейшие нарушения безопасности

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag06. Нарушение доступности (DoS/DoW) среды исполнения AI-агента (в т.ч. функций)

Описание
Использование специально подобранных входных данных для исполнения функций, потребляющих избыточное количество ресурсов среды исполнения

Последствия
Прекращение или снижение скорости оказания услуг AI-агентом/приложением или увеличение затрат на эксплуатацию

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag07. Утечка информации об архитектуре мультиагентной системы через интерфейсы пользовательского ввода

Описание
Использование специально подобранных входных данных для извлечения информации о мультиагентной системе (включая ее архитектуру, состав, правила взаимодействия) из ответов AI-агента

Последствия
Нарушение конфиденциальности состава мультиагентной системы и устройства взаимодействия между AI-агентами в мультиагентной системе, облегчение проведения кибератак

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag08. Передача другому AI-агенту ложной информации в мультиагентной системе

Описание
Использование специально подобранных входных данных для модификации и искажения семантической составляющий информации, передаваемой AI-агентом в рамках мультиагентной системы

Последствия
Нарушение взаимодействия и кооперации AI-агентов, нарушение рабочего процесса приложения или мультиагентной системы

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag09. Нарушение цели другого AI-агента при кооперативном взаимодействии в мультиагентной системе

Описание
Использование специально подобранных входных данных для модификации цели другого AI-агента в мультиагентной системе путем передачи запроса с промпт-атакой

Последствия
Нарушение взаимодействия и кооперации AI-агентов, нарушение рабочего процесса приложения или мультиагентной системы

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Целостность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag10. Распространение промпт-атаки по AI-агентам в мультиагентной системе для усиления ее эффекта

Описание
Использование специально подобранных входных данных для распространения промпт-атак в мультиагентной системе

Последствия
Распространения в мультиагентной системе промпт-атаки, имеющей целью возникновение различных негативных последствий

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность, целостность, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag11. Нарушение рабочего процесса приложения, реализующей AI-агента

Описание
Использование специально подобранных входных данных для модификации и искажения семантической составляющей информации, структуры или формата её представления AI-агентом

Последствия
Искажение результатов работы, нарушение доступности приложения, реализующей AI-агента

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Целостности, доступность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Ag12. Утечка информации о цели, функциях, содержимом памяти или инструкциях механизма планирования AI-агента

Описание
Использование специально подобранных входных данных для получения информации, связанной с AI-агентом, в том числе о его цели, описании функций, инструкциях механизма планирования, содержимого памяти

Последствия
Нарушение конфиденциальности цели, описания функций, содержимого памяти или инструкций механизма планирования AI-агента

Объект воздействия
AR02 AI-агенты

Нарушаемое свойство
Конфиденциальность

Виды моделей, подверженных угрозе
GenAI

Лица, ответственные за митигацию угрозы
Владелец приложения

Материалы, использованные при подготовке

1. OWASP Top-10 LLM 2025
2. OWASP Top-10 Machine Learning Security
3. OWASP AI Security Solutions Landscape
4. OWASP Agentic Threats Taxonomy (draft)
5. OWASP AI Exchange 4.5
6. MITRE ATT&CK
7. MITRE ATLAS
8. Google SAIF (Secure AI Framework)
9. NIST Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations
10. AWS Generative AI Security Scoping Matrix