СБЕР
КИБЕР
БЕЗОПАСНОСТЬ

Модель угроз кибербезопасности AI

для этапов сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с AI-решениями

v.2.0 (июль 2026)

Оглавление

Общие сведения о модели угроз

Документ подготовлен Управлением кибербезопасности искусственного интеллекта Службы кибербезопасности Сбера на основе анализа актуального ландшафта угроз, мировых практик и собственного опыта расследования инцидентов.

С момента выхода предыдущей версии ландшафт угроз кибербезопасности (КБ) претерпел значительные изменения. Массовое внедрение GenAI, усложнение архитектур AI-решений (мультиагентные системы, RAG, LLM-адаптеры) и накопленная практика эксплуатации выявили новые векторы атак. Актуализированная модель угроз сохраняет преемственность и расширяет охват, отражая современные реалии.

Ключевые особенности модели угроз кибербезопасности AI v.2.0:

- Разделение на угрозы КБ AI и способы их реализации. Угрозы описывают негативные последствия, способы реализации – конкретные действия нарушителя. Это позволяет выстраивать многоуровневую защиту: блокировать техники атак и минимизировать последствия.
- Привязка к цепочке атаки MITRE ATLAS. Способы реализации классифицированы по тактикам (от Reconnaissance до Exfiltration), угрозы – по типу последствий (Impact). Это обеспечивает единый язык с экспертами кибербезопасности.
- Детализация нарушителей. Введены категории внешних и внутренних нарушителей с разным уровнем технических возможностей и осведомлённости нарушителя о системе – от white-box (полное знание архитектуры, данных и параметров) до black-box (только публичные интерфейсы), с промежуточным grey-box (частичное знание), что позволяет точнее оценивать реализуемость угроз для каждого типа нарушителя.
- Расширенный перечень объектов защиты (таких как системные промпты, LLM-адаптеры, память и кэш AI-агентов, механизмы обмена сообщениями в мультиагентных системах, среда исполнения AI-агентов, а также ресурсные системы AI-агентов (внешние сервисы и API). В модель включены компоненты, которые в предыдущей версии не рассматривались, но сегодня являются критическими точками атак. Их выделение позволяет покрыть угрозы, связанные с новой архитектурой AI-решений, и дать командам разработки и кибербезопасности конкретные ориентиры для выработки мер защиты.

В документе представлены 37 угроз КБ AI и 51 способ их реализации. Для каждого элемента указаны: описание, нарушаемые свойства информации и последствия, связанные объекты, этапы жизненного цикла, виды моделей (PredAI/GenAI) и конкретные роли сотрудников (например, владелец данных, разработчик (владелец) модели, разработчик (владелец) AI-решения, владелец ИТ-инфраструктуры, эксперт по кибербезопасности), отвечающие за меры защиты (Приложение 4). Виды нарушителей, способные реализовать каждую угрозу и каждый способ реализации, приведены в разделе «Потенциальные нарушители кибербезопасности AI» и Приложении 3.

Модель предназначена для разработчиков AI-решений, MLOps/DevOps-инженеров, архитекторов, экспертов по кибербезопасности и руководителей, отвечающих за управление рисками при внедрении AI-решений.

Термины и определения

AI-агент – система, использующая LLM, способная планировать и совершать автономные действия во внешней среде, реагировать на изменения и взаимодействовать с человеком и другими агентами для достижения поставленных целей.

AI-модель – программа для электронных вычислительных машин (ее составная часть), предназначенная для выполнения интеллектуальных задач на уровне, сопоставимом с результатами интеллектуального труда человека или превосходящем их, использующая алгоритмы и наборы данных для выведения закономерностей, принятия решений или прогнозирования результатов. AI-модели разделяются на два класса: модели генеративного (GenAI) и предиктивного искусственного интеллекта (PredAI).

AI-приложение – приложение, которое использует AI-модель (PredAI или GenAI) как компонент для решения задачи и не являющееся AI-агентом.

AI-решение – цифровой программный продукт или его часть с функцией искусственного интеллекта, реализованный как AI-приложение или AI-агент.

GenAI – модель, создающая новый контент (текст, изображения, аудио и пр.) на основе анализа и обработки существующих данных.

LLM (Большая языковая модель) – система искусственного интеллекта, основанная на нейронных сетях, которая обучена на больших объемах текстовых данных и способна обрабатывать, понимать и генерировать текст, аналогичный человеческому. Такие модели используют методы глубокого обучения и могут содержать миллиарды параметров, что позволяет им улавливать нюансы языка, грамматики и контекста.

LLM-адаптер – компонент, обученный на данных специализированной задачи, подключаемый к LLM без изменения её исходных весов (например, LoRA).

LoRA (Low-Rank adaptation) – метод адаптации LLM, который позволяет эффективно настраивать их под конкретные задачи, изменяя лишь некоторые параметры или слои вместо того, чтобы обучать всю модель заново.

PredAI – модель, обученная предсказывать результаты на основе репрезентативных данных.

RAG (Retrieval Augmented Generation) – подход к работе с LLM, при котором входной запрос к модели дополняется релевантными данными, извлечёнными из внешних источников, для генерации более фактологически точных и контекстуально релевантных ответов.

БД RAG, векторная база данных – хранилище эмбеддингов, позволяющее быстро искать похожие векторы.

Внешний нарушитель – нарушитель кибербезопасности, не имеющий прав доступа в контролируемую (охраняемую) зону (территорию) и (или) полномочий по доступу к объекту защиты или его компонентам, требующим авторизации.

Внутренний нарушитель – нарушитель кибербезопасности, имеющий права доступа в контролируемую (охраняемую) зону (территорию) и (или) полномочий по доступу к объекту защиты и его компонентам.

Выравнивание (alignment) – процесс согласования целей, поведения и ответов AI-модели с намерениями пользователя, этическими нормами и установленными ограничениями.

Дообучение (Fine-tuning) – процесс адаптации предварительно обученной модели для решения конкретных задач или использования в определенных сценариях.

Инференс – эксплуатация модели для обработки входных данных и генерации предсказаний или ответов в реальном времени без внесения изменений в веса модели.

Искусственный интеллект (ИИ или AI) – комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые с результатами интеллектуальной деятельности человека или превосходящие их.

Мультиагентная система – система, состоящая из множества AI-агентов, которые взаимодействуют друг с другом для достижения общих целей.

Набор данных, датасет – набор упорядоченных данных, используемых для анализа и обучения моделей.

Нарушитель КБ AI (нарушитель) – субъект (физическое лицо, группа лиц, AI-агент), имеющий цели и потенциал для реализации угроз КБ AI в отношении объекта защиты в процессе сбора и подготовки данных, разработки и обучения модели, эксплуатации модели и интеграций с AI-решениями.

Негативные последствия – совокупность последствий, которые являются результатом реализации угроз КБ AI на объект защиты и которые приводят к какому-либо ущербу.

Объект воздействия – информационные ресурсы и компоненты систем и сетей, несанкционированный доступ к которым или воздействие на которые в ходе реализации (возникновения) угроз КБ AI может привести к негативным последствиям.

Объект защиты – отдельный элемент или функционально связанная группа элементов, подлежащих защите с использованием организационных и технических мер. К объектам защиты относятся: информационные активы и процессы; ИТ-инфраструктура: совокупность технологий, оборудования и прикладного / платформенного / системного ПО (в том числе ПО и программно-аппаратные комплексы центров обработки данных), обеспечивающая функционирование автоматизированных систем и сервисов.

Потенциал нарушителя КБ – совокупность уровня технических возможностей (компетенции, инструментарий и доступные ресурсы для реализации атакующих воздействий) и уровня осведомлённости и доступа (объём внутренней информации об архитектуре, данных, коде и конфигурациях AI-решения, а также характер легитимного взаимодействия с его компонентами) для реализации нарушителем угроз КБ AI.

Предобучение (Pretrain) – процесс начального обучения модели на большом объеме данных с целью создания базовых представлений и знаний, которые затем могут быть адаптированы для решения конкретных задач. Этот этап позволяет модели развить общее понимание структуры и закономерностей данных, в том числе естественного языка.

Промпт – запрос или инструкция, предоставляемая LLM-модели пользователем или AI-решением и используемая для генерации ответа.

Промпт-атака – класс кибератак, эксплуатирующих уязвимости GenAI, выполняемых путем отправки GenAI-модели вредоносных входных данных.

Системный промпт – набор инструкций, используемый для управления поведением LLM. Описывает контекст задачи, роль, правила и ограничения, которым LLM должна следовать при генерации ответа. Может включать статическую часть с описанием цели, роли, допустимых ограничений и динамическую с динамическими параметрами, например, имя пользователя, текущие дату и время, другую дополнительную информацию.

Способы реализации угроз КБ AI – техники и процедуры с использованием которых нарушитель КБ осуществляет атаки и, впоследствии, реализацию угрозы КБ AI.

Токен (token) – основная единица обработки данных для LLM. Может являться элементом, кодирующим фрагмент текста, аудио или изображения.

Угроза КБ – совокупность условий и факторов, создающих потенциально или реально существующую опасность нарушения КБ в отношении объекта защиты.

Угроза КБ AI – совокупность условий и факторов, создающих потенциально или реально существующую опасность нарушения КБ в отношении объекта защиты на этапах сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с AI-решениями.

Функции / инструменты – в контексте AI-приложений - внешние по отношению к AI-решению инструменты, которые приложение может использовать в процессе работы для достижения поставленных целей. В контексте AI-агентов - программный компонент (в т.ч. другой AI-агент), реализующий функциональность и предоставляющий стандартизированный интерфейс доступа (в виде API), вызов которого инициируется на основании ответа LLM.

Эмбеддинги – числовые представления данных, которые позволяют моделям машинного обучения анализировать и интерпретировать текст, изображения и другие типы информации. Они представляют собой векторы, которые сохраняют семантические связи между элементами данных.

Группы угроз кибербезопасности AI и способов их реализации

В основе данной версии модели угроз лежит разделение на угрозы КБ AI и способы их реализации. Такое разделение позволяет выстраивать многоуровневую защиту — блокировать действия нарушителя на ранних этапах и минимизировать последствия в случае реализации угрозы.

Для соотнесения с отраслевыми подходами используется цепочка атаки (killchain) по MITRE ATLAS: способы реализации соответствуют тактикам подготовки и проведения атаки (от Reconnaissance до Exfiltration), а угрозы – финальной тактике Impact, то есть наступившим негативным последствиям.

Под **угрозой КБ AI** понимается совокупность условий и факторов, создающих потенциально или реально существующую опасность нарушения КБ в отношении объекта защиты на этапах сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с AI-решениями. Характеристики:

- Описывают, что может произойти, как описание последствий
- Привязаны к объектам защиты, в которых реализуется негативное последствие
- Сгруппированы по типам последствий – 8 групп

№	Группы угроз КБ AI	Количество
1.	Несанкционированный сбор информации (разведка и утечка данных)	7
2.	Утечка конфиденциальной информации	7
3.	Несанкционированное воздействие на информационные ресурсы и среду исполнения	7
4.		
5.	Отказ в обслуживании	7
6.	Нарушение мультиагентного взаимодействия	2
7.	Распространение противоправной информации	2
8.	Нарушение функционирования (работоспособности)	4
9.	Нарушение процессов мониторинга и расследования	1
Итого		37

Способы реализации – техники, процедуры с использованием которых нарушитель КБ осуществляет атаки и, впоследствии, реализацию угрозы КБ AI. Характеристики:

- Описывают, как реализуется угроза, механику атаки
- Привязаны к объектам воздействия
- Классифицированы по этапам killchain в соответствии с тактиками MITRE ATLAS[1]

№	Группы способов реализации	Количество
1.	Reconnaissance	1
2.	Resource Development	2
3.	Initial Access	7
4.	AI Model Access	1
5.	Execution	10

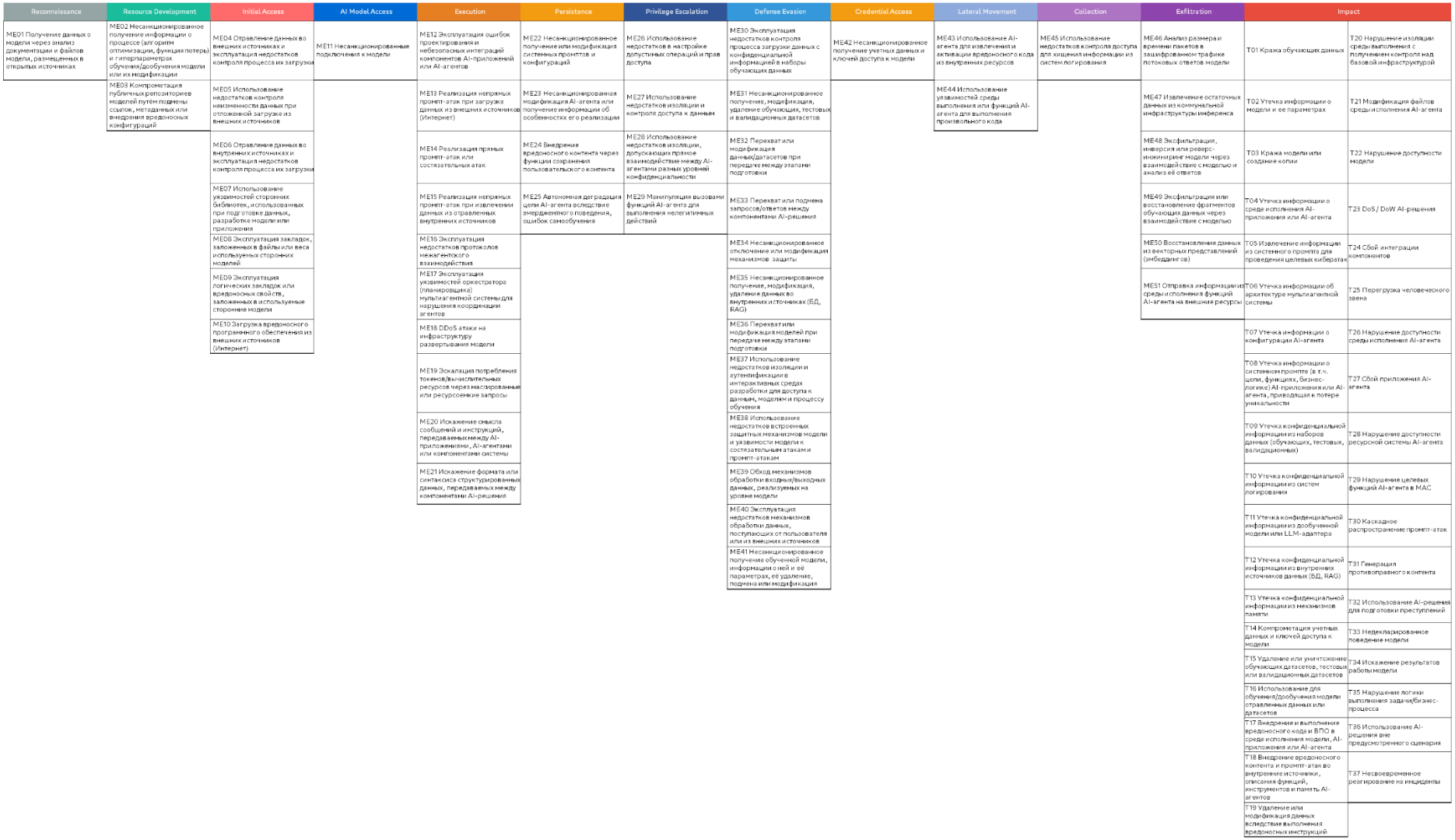
[1] Discovery и AI Attack Staging не выделены в отдельные группы, поскольку покрываются составом смежных способов реализации (Discovery – через Reconnaissance и Collection; AI Attack Staging – через Resource Development и Initial Access) либо реализуются вне контура организации

6.	Persistence	4
7.	Privilege Escalation	4
8.	Defense Evasion	12
9.	Credential Access	1
10.	Lateral Movement	2
11.	Collection	1
12.	Exfiltration	6
Итого		51

В отличие от предыдущей версии модели, где идентификатор был привязан к первичному объекту, в текущей версии используется сквозная нумерация, не зависящая от объекта: угрозы КБ AI – T01–T37, способы реализации – ME01–ME51.

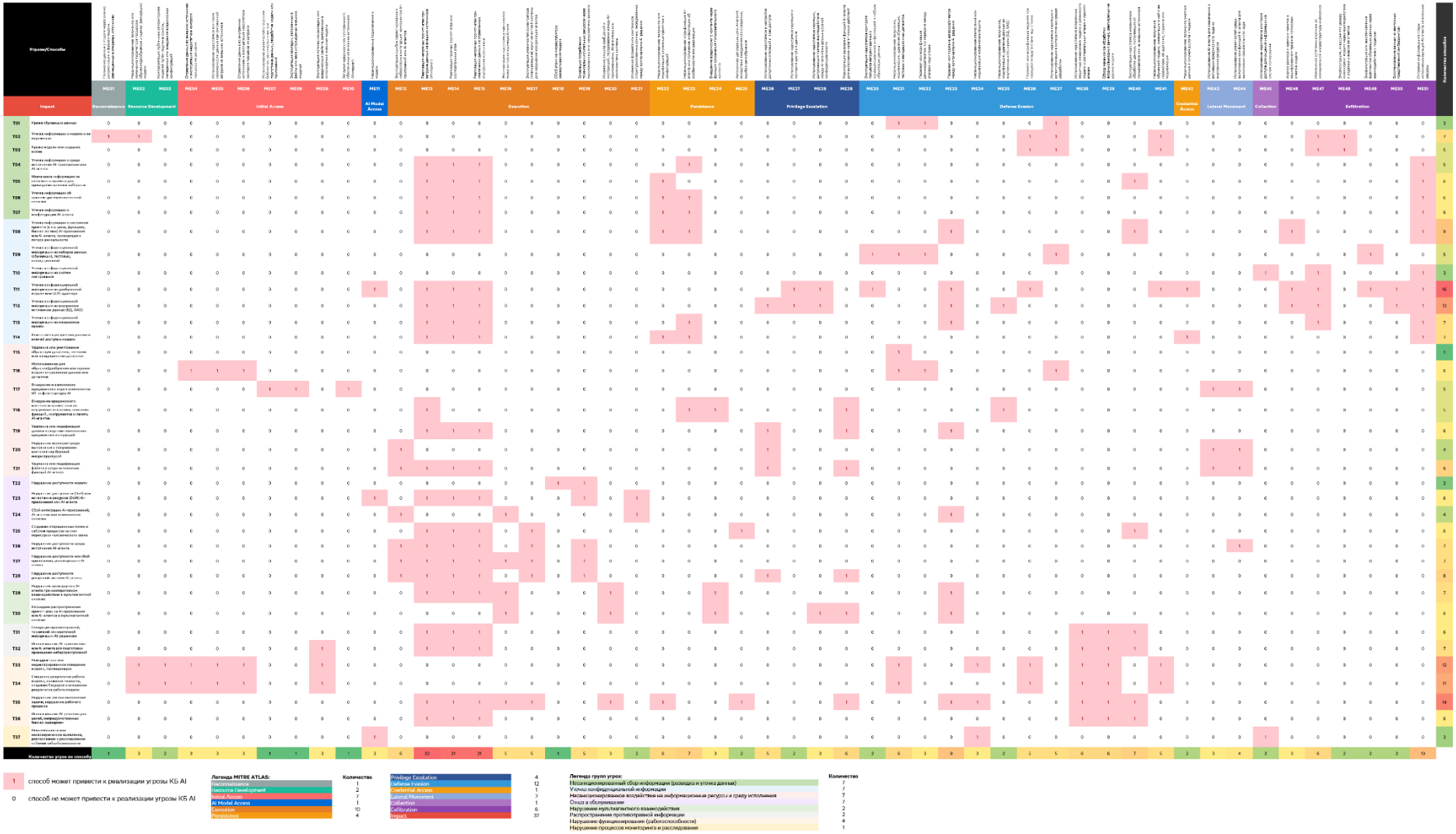
Одна угроза может реализовываться несколькими способами реализации, один способ реализации может приводить к нескольким угрозам (раздел «Матрица угроз кибербезопасности AI и способов их реализации»).

Угрозы кибербезопасности AI и способы их реализации по цепочке атаки (killchain)



Reconnaissance	Resource Development	Initial Access	AI Model Access	Execution	Persistence	Privilege Escalation	Defense Evasion	Credential Access	Lateral Movement	Collection	Exfiltration	Impact	
ME01 Получение данных о модели через анализ документации и файлов модели, размещенных в открытых источниках	ME02 Несанкционированное получение информации о процессе (алгоритм оптимизации, функция потерь) и гиперпараметрах обучения/дообучения модели или их модификации	ME04 Отравление данных во внешних источниках и эксплуатация недостатков контроля процесса их загрузки	ME11 Несанкционированные подключения к модели	ME12 Эксплуатация ошибок проектирования и небезопасных интеграций компонентов AI-приложений или AI-агентов	ME22 Несанкционированное получение или модификация системных промптов и конфигураций	ME26 Использование недостатков в настройке допустимых операций и прав доступа	ME30 Эксплуатация недостатков контроля процесса загрузки данных с конфиденциальной информацией в наборы обучающих данных	ME42 Несанкционированное получение учетных данных и ключей доступа к модели	ME43 Использование AI-агента для извлечения и активации вредоносного кода из внутренних ресурсов	ME45 Использование недостатков контроля доступа для хищения информации из систем логирования	ME46 Анализ размера и времени пакетов в зашифрованном трафике потоковых ответов модели	T01 Кража обучающих данных	T20 Нарушение изоляции среды выполнения с получением контроля над базовой инфраструктурой
	ME03 Компрометация публичных репозиториев моделей путём подмены ссылок, метаданных или внедрения вредоносных конфигураций	ME05 Использование недостатков контроля неизменности данных при отложенной загрузке из внешних источников		ME13 Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет)	ME23 Несанкционированная модификация AI-агента или получение информации об особенностях его реализации	ME27 Использование недостатков изоляции и контроля доступа к данным	ME31 Несанкционированное получение, модификация, удаление обучающих, тестовых и валидационных датасетов		ME44 Использование уязвимостей среды выполнения или функций AI-агента для выполнения произвольного кода		ME47 Извлечение остаточных данных из коммунальной инфраструктуры инференса	T02 Утечка информации о модели и ее параметрах	T21 Модификация файлов среды исполнения AI-агента
		ME06 Отравление данных во внутренних источниках и эксплуатация недостатков контроля процесса их загрузки		ME14 Реализация прямых промпт-атак или состязательных атак	ME24 Внедрение вредоносного контента через функции сохранения пользовательского контента	ME28 Использование недостатков изоляции, допускающих прямое взаимодействие между AI-агентами разных уровней конфиденциальности	ME32 Перехват или модификация данных/датасетов при передаче между этапами подготовки				ME48 Эксфильтрация, инверсия или реверс-инжиниринг модели через взаимодействие с моделью и анализ её ответов	T03 Кража модели или создание копии	T22 Нарушение доступности модели
		ME07 Использование уязвимостей сторонних библиотек, использованных при подготовке данных, разработке модели или приложения		ME15 Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников	ME25 Автономная деградация цели AI-агента вследствие эмерджентного поведения, ошибок самообучения	ME29 Манипуляция вызовами функций AI-агента для выполнения нелегитимных действий	ME33 Перехват или подмена запросов/ответов между компонентами AI-решения				ME49 Эксфильтрация или восстановление фрагментов обучающих данных через взаимодействие с моделью	T04 Утечка информации о среде исполнения AI-приложения или AI-агента	T23 DoS / DoW AI-решения
		ME08 Эксплуатация закладок, заложенных в файлы или веса используемых сторонних моделей		ME16 Эксплуатация недостатков протоколов межагентского взаимодействия			ME34 Несанкционированное отключение или модификация механизмов защиты				ME50 Восстановление данных из векторных представлений (эмбеддингов)	T05 Извлечение информации из системного промпта для проведения целевых кибератак	T24 Сбой интеграции компонентов
		ME09 Эксплуатация логических закладок или вредоносных свойств, заложенных в используемые сторонние модели		ME17 Эксплуатация уязвимостей оркестратора (планировщика) мультиагентной системы для нарушения координации агентов			ME35 Несанкционированное получение, модификация, удаление данных во внутренних источниках (БД, RAG)				ME51 Отправка информации из среды исполнения функций AI-агента на внешние ресурсы	T06 Утечка информации об архитектуре мультиагентной системы	T25 Перегрузка человеческого звена
		ME10 Загрузка вредоносного программного обеспечения из внешних источников (Интернет)		ME18 DDoS атаки на инфраструктуру развертывания модели			ME36 Перехват или модификация моделей при передаче между этапами подготовки					T07 Утечка информации о конфигурации AI-агента	T26 Нарушение доступности среды исполнения AI-агента
				ME19 Эскалация потребления токенов/вычислительных ресурсов через массированные или ресурсоемкие запросы			ME37 Использование недостатков изоляции и аутентификации в интерактивных средах разработки для доступа к данным, моделям и процессу обучения					T08 Утечка информации о системном промпте (в т.ч. цели, функциях, бизнес-логике) AI-приложения или AI-агента, приводящая к потере уникальности	T27 Сбой приложения AI-агента
				ME20 Искажение смысла сообщений и инструкций, передаваемых между AI-приложениями, AI-агентами или компонентами системы			ME38 Использование недостатков встроенных защитных механизмов модели и уязвимости модели к состязательным атакам и промпт-атакам					T09 Утечка конфиденциальной информации из наборов данных (обучающих, тестовых, валидационных)	T28 Нарушение доступности ресурсной системы AI-агента
				ME21 Искажение формата или синтаксиса структурированных данных, передаваемых между компонентами AI-решения			ME39 Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели					T10 Утечка конфиденциальной информации из систем логирования	T29 Нарушение целевых функций AI-агента в MAC
							ME40 Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников					T11 Утечка конфиденциальной информации из дообученной модели или LLM-адаптера	T30 Каскадное распространение промпт-атак
							ME41 Несанкционированное получение обученной модели, информации о ней и её параметрах, её удаление, подмена или модификация					T12 Утечка конфиденциальной информации из внутренних источников данных (БД, RAG)	T31 Генерация противоправного контента
												T13 Утечка конфиденциальной информации из механизмов памяти	T32 Использование AI-решения для подготовки преступлений
												T14 Компрометация учетных данных и ключей доступа к модели	T33 Недекларированное поведение модели
												T15 Удаление или уничтожение обучающих датасетов, тестовых или валидационных датасетов	T34 Искажение результатов работы модели
												T16 Использование для обучения/дообучения модели отравленных данных или датасетов	T35 Нарушение логики выполнения задачи/бизнес-процесса
												T17 Внедрение и выполнение вредоносного кода и BTO в среде исполнения модели, AI-приложения или AI-агента	T36 Использование AI-решения вне предусмотренного сценария
												T18 Внедрение вредоносного контента и промпт-атак во внутренние источники, описания функций, инструментов и память AI-агентов	T37 Несвоевременное реагирование на инциденты
												T19 Удаление или модификация данных вследствие выполнения вредоносных инструкций	

Матрица угроз кибербезопасности AI и способов их реализации



Практическое руководство по работе с документом

Общие принципы

Модель угроз кибербезопасности AI предназначена для системного анализа угроз безопасности информации, возникающих на этапах сбора и подготовки данных, разработки и обучения модели, эксплуатации модели и интеграций с AI-решениями.

Модель угроз является основой для последующей разработки организационных и технических мер защиты. Сами меры в рамках данной модели не определяются — они разрабатываются на основе полученных результатов с учётом требований регуляторов, архитектурных особенностей и допустимого уровня риска.

Работа с моделью угроз строится в соответствии с общими принципами оценки угроз кибербезопасности:

1. Принцип полноты. Угрозы рассматриваются на всех этапах жизненного цикла AI-решения: сбор и подготовка данных, разработка и обучение модели, эксплуатация модели и интеграции с AI-решениями.
2. Принцип системности. Угрозы, способы их реализации и объекты рассматриваются в совокупности и взаимосвязи. Одна угроза может реализовываться несколькими способами реализации, один способ реализации может приводить к нескольким угрозам. Связи между угрозами и способами реализации устанавливаются с помощью «Матрицы угроз кибербезопасности AI и способов их реализации» и карточками угроз в разделе «Перечень угроз кибербезопасности AI».
3. Принцип объектности. Состав угроз и способов реализации определяется по составу объектов AI-решения через установление объектов защиты [2] (для угроз) и объектов воздействия [3] (для способов реализации). Состав объектов представлен в разделе «Объекты модели угроз кибербезопасности AI», а взаимосвязь объектов и угроз/способов реализации приведена в Приложении 1.
4. Принцип учёта нарушителя. Для каждой угрозы определяются виды нарушителей, обладающих возможностью для её реализации, с учётом их технических возможностей, осведомлённости о системе и характера доступа к её компонентам. Виды нарушителей приведены в разделе «Потенциальные нарушители кибербезопасности AI» и в Приложениях 2–3.
5. Принцип обоснованности мер. Информация о связях угроз и способов реализации, полученная в результате применения модели, обеспечивает обоснованный выбор организационных и технических мер защиты на последующих этапах проектирования. Блокирующие меры разрабатываются в отношении способов реализации, компенсирующие — в отношении негативных последствий угроз; компоненты, на которых реализуются меры, определяются при проектировании системы защиты с учётом архитектуры конкретного AI-решения.

Алгоритм работы

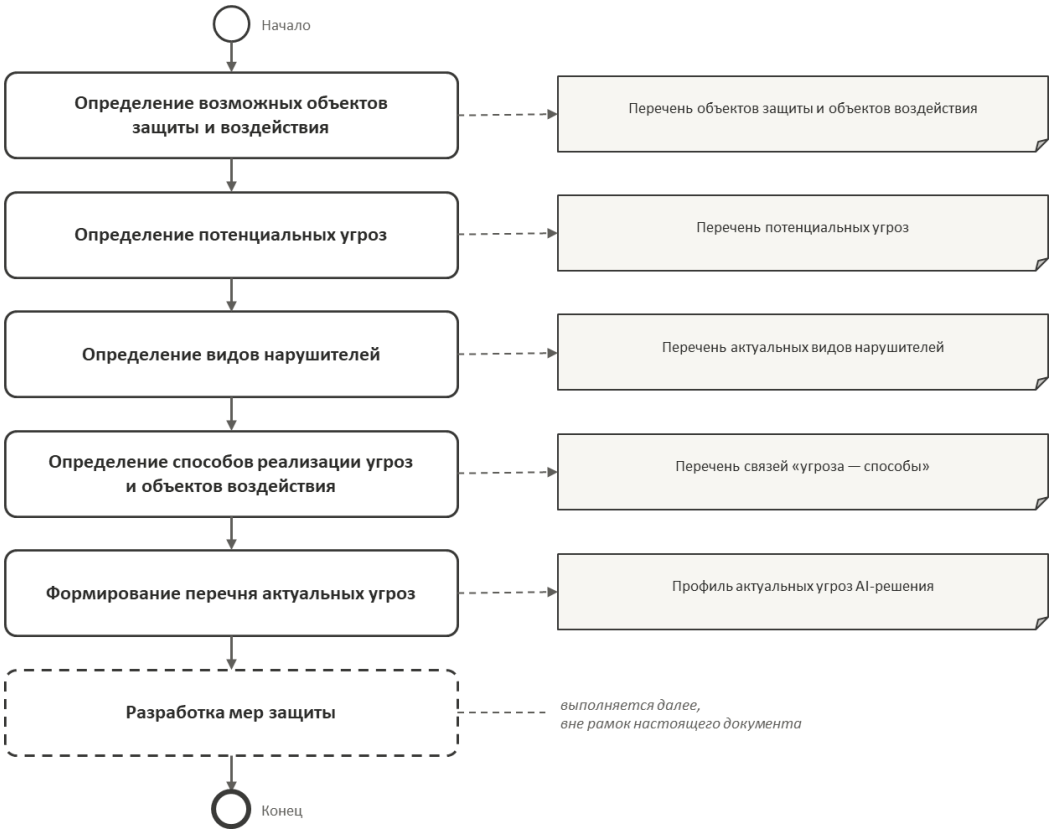
Работа с моделью угроз представляет собой последовательный процесс, который начинается с инвентаризации объектов, присутствующих в архитектуре, затем определяются актуальные угрозы и способы их реализации, и завершается формированием профиля угроз. Ниже приведены

---

[2] объект относительного которого могут наступить негативные последствия (нарушение конфиденциальности, целостности или доступности)

[3] объект, через который или на который нарушитель непосредственно воздействует при применении способа реализации угрозы

шаги, которые рекомендуется выполнять при первом обращении к документу, а затем повторять при изменениях в составе объектов или ландшафте угроз.



Алгоритм работы с моделью угроз и результаты шагов

1. Определение возможных объектов защиты и воздействия

В ходе оценки угроз должен быть определен состав объектов, подлежащих защите, и объектов воздействия. Совокупность объектов защиты и объектов воздействия определяет границы дальнейшего процесса оценки угроз. Эти границы задаются этапами жизненного цикла, которые охватывает AI-решение, и его архитектурными особенностями (наличие мультиагентной системы, RAG, LLM-адаптеров, внешних моделей, внешних источников данных)[4].

Определение границ необходимо, чтобы из множества потенциальных угроз выделить актуальные для конкретного решения, и исключить те, реализация которых невозможна из-за отсутствия соответствующих объектов. Для этого на основании раздела «Объекты модели угроз кибербезопасности AI» из полного перечня отбираются только те объекты, которые соответствуют определённым этапам жизненного цикла и архитектурным элементам. Объекты, не соответствующие этим условиям, исключаются вместе со связанными с ними угрозами и способами реализации.

Результатом является перечень объектов, актуальных для заданной области рассмотрения, с указанием их ролей (объект защиты, объект воздействия). Этот перечень служит основой для последующего определения угроз и способов их реализации.

[4] При использовании внешних сервисов (например, модели как сервис) в границы включаются только объекты, находящиеся под контролем организации; объекты на стороне поставщика исключаются из рассмотрения.

2. Определение потенциальных угроз

На основе перечня объектов, полученного в результате инвентаризации, определяются потенциальные угрозы, актуальные для данного AI-решения.

Для каждого выделенного объекта устанавливаются угрозы, негативные последствия которых могут наступить в отношении этого объекта. Угрозы, не соответствующие типам используемых моделей (PredAI, GenAI), исключаются.

Соответствие угроз объектам зафиксировано в карточках угроз и обобщено в таблице в Приложении 1.

3. Определение видов нарушителей

После того как определён перечень актуальных угроз, для каждой из них определяется перечень нарушителей. Виды нарушителей и их потенциал приведены в разделе «Потенциальные нарушители кибербезопасности AI» и Приложении 2. Из общего перечня исключаются виды нарушителей, неактуальные для условий функционирования конкретного AI-решения.

Для каждого вида нарушителя оцениваются его категория (внешний или внутренний), уровень технических возможностей, уровень осведомлённости о AI-решении и модели (white-box, grey-box, black-box) и характер доступа к компонентам AI-решения; в совокупности это определяет потенциал нарушителя (низкий, средний, высокий).

Результатом шага является перечень актуальных видов нарушителей с указанием их потенциала. Он используется на следующем шаге при отборе способов реализации угроз: способ признаётся актуальным, если хотя бы один из актуальных нарушителей обладает возможностями для его реализации.

4. Определение способов реализации угроз и объектов воздействия

После того как определён перечень актуальных угроз, для каждой из них определяется перечень способов реализации, которые могут к ней привести. Для этого используется матрица соответствия (раздел «Матрица угроз кибербезопасности AI и способов их реализации») или карточки угроз, где перечислены все способы реализации.

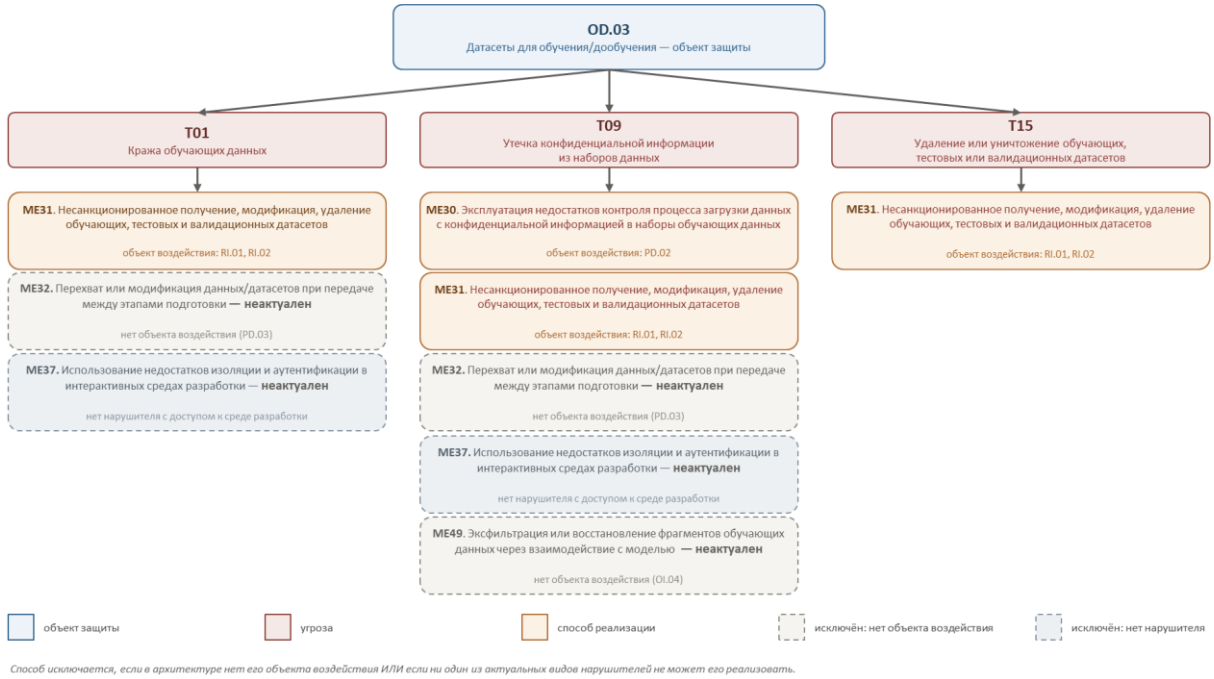
Из этого перечня отбираются только те способы реализации, для которых одновременно выполняются два условия:

- объект воздействия, указанный в карточке способа реализации, присутствует в составе объектов AI-решения, определённом на этапе инвентаризации;
- способ реализации доступен хотя бы одному из нарушителей, рассматриваемых для данного AI-решения.

Способы реализации, не удовлетворяющие этим условиям, исключаются из рассмотрения.

Сопоставление объектов и способов реализации выполняется с помощью таблицы в Приложении 1 для проверки наличия объектов воздействия, и карточек способов реализации для определения объекта воздействия и видов нарушителей.

Результатом этого шага является полный перечень связей «угроза ↔ способы реализации» с привязкой к объектам.



Пример определения угроз и способов реализации для объекта защиты OD.03

5. Формирование перечня актуальных угроз

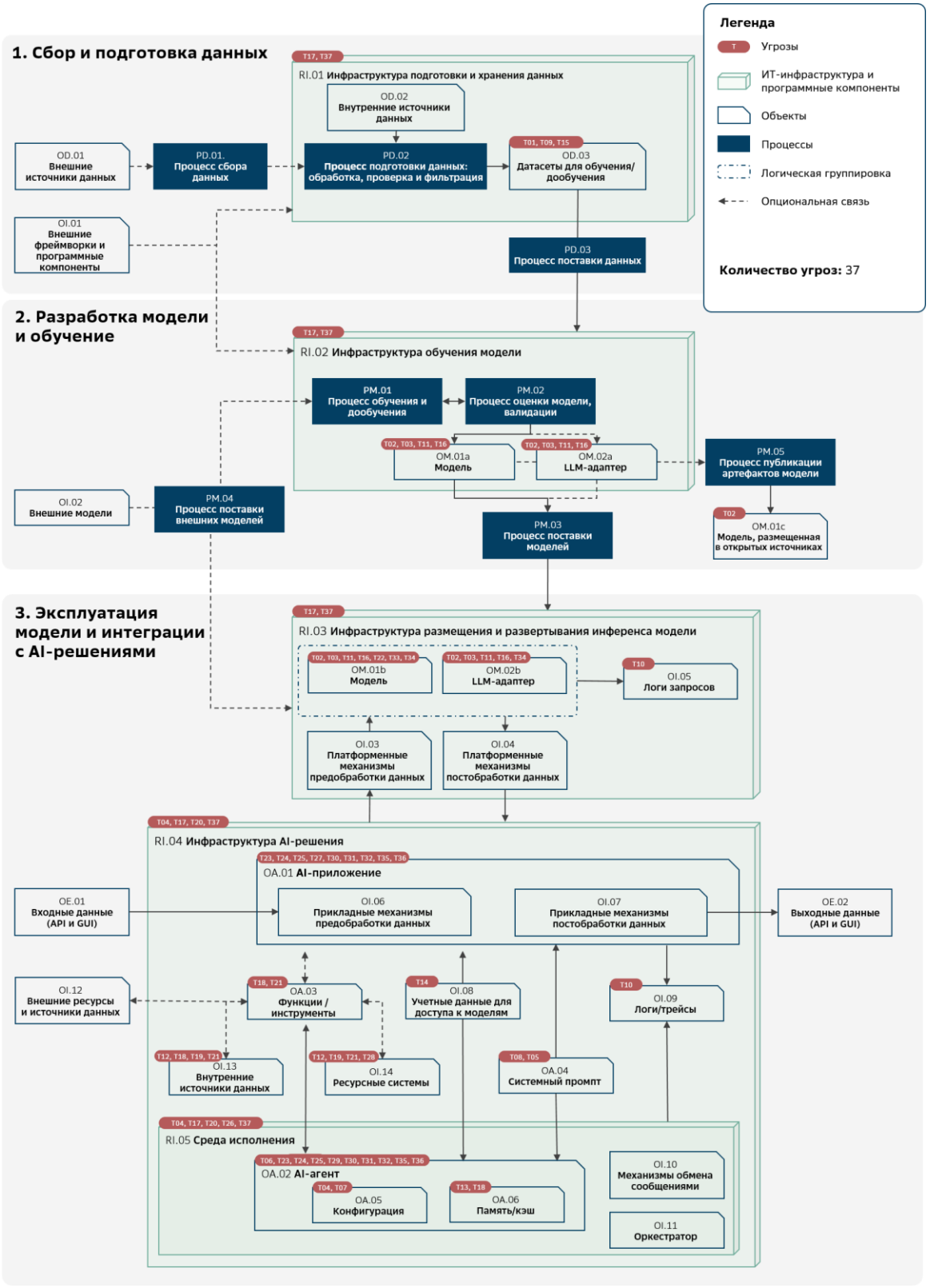
Результатом всех выполненных шагов является перечень актуальных угроз для конкретного AI-решения, который включает:

- перечень объектов защиты и объектов воздействия;
- перечень угроз с указанием нарушаемых свойств информации (конфиденциальность, целостность, доступность);
- для каждой угрозы — перечень способов реализации с указанием объектов воздействия;
- для каждой угрозы и каждого способа реализации — виды нарушителей, способные их реализовать.

Далее может быть выполнено ранжирование угроз по критичности, а также разработка организационных и технических мер защиты. Эти действия выходят за рамки данной модели угроз и выполняются на последующих этапах проектирования системы защиты.

Объекты модели угроз кибербезопасности AI

В разделе представлен перечень объектов, задействованных в модели угроз, и раскрыта их связь с угрозами T01–T37. Для общего визуального восприятия приведена схема, где обозначены объекты защиты — объекты, для которых возможны негативные последствия.



Обобщённая схема объектов МУ КБ AI и угроз КБ AI

Ресурсы (ИТ-инфраструктура и программные компоненты)

RI.01 Инфраструктура подготовки и хранения данных – совокупность программных и аппаратных средств для сбора, обработки, хранения и управления данными на этапе подготовки к обучению, включая интерактивные среды для разработки и экспериментов

RI.02 Инфраструктура обучения модели – совокупность программных и аппаратных средств для обучения и оптимизации моделей (вычислительные кластеры, GPU-серверы, оркестраторы обучения), включая интерактивные среды для разработки и экспериментов (Jupyter Notebook, Google Colab, облачные IDE, локальные среды разработки)

RI.03 Инфраструктура размещения и развертывания инференса модели – совокупность программных и аппаратных средств для развертывания обученных моделей для промышленного использования (сервера инференса, API-шлюзы, балансировщики). Включает как собственные серверы инференса, так и подключение к внешним MLaaS-провайдерам через API-шлюзы

RI.04 Инфраструктура AI-решения – совокупность средств для выполнения AI-решений (серверы приложений, контейнеры, оркестрация, сетевые компоненты)

RI.05 Среда исполнения AI-агентов – изолированная runtime-среда (контейнер, песочница, serverless-функция, виртуальная машина), в которой выполняется код AI-агента

Процессы

Процессы, связанные с данными

PD.01 Процесс сбора данных – процесс сбора и поставки данных, используемых для формирования наборов обучающих, тестовых, валидационных данных

PD.02 Процесс подготовки данных: обработка, проверка, фильтрация – процессы и компоненты, выполняющие очистку, нормализацию, преобразование форматов и фильтрацию данных для создания обучающих выборок

PD.03 Процесс поставки данных – совокупность процедур и технических трактов поставки данных, обеспечивающих передачу данных и датасетов между компонентами инфраструктуры на различных этапах жизненного цикла модели и данных

Процессы, связанные с моделями

PM.01 Процесс обучения и дообучения – процесс по применению алгоритмов машинного обучения для создания и адаптации моделей

PM.02 Процесс оценки модели, валидации – процесс по тестированию и валидации качества модели, включая её безопасность и устойчивость к вредоносным воздействиям.

PM.03 Процесс поставки моделей – совокупность процедур и технических трактов поставки модели, обеспечивающих перенос файлов обученной модели (включая веса и архитектуру) из среды обучения в среду развертывания инференса

PM.04 Процесс поставки внешних моделей – совокупность процедур и технических трактов, обеспечивающих перенос предобученной модели из внешнего источника (публичные репозитории, сайты вендоров) в среду обучения или развертывания инференса в инфраструктуре организации для последующего использования (дообучение, инференс, интеграция в AI-приложения).

PM.05 Процесс публикации артефактов модели – совокупность процедур и технических трактов, обеспечивающих контроль за размещением в открытых источниках документации, файлов модели, конфигураций и иных артефактов, а также их проверку на наличие конфиденциальной информации перед публикацией.

Объекты

Объекты, связанные с данными

OD.01 Внешние источники данных – публичные датасеты, API, социальные сети, открытые базы данных, веб-страницы

OD.02 Внутренние источники данных – корпоративные базы данных, логи пользовательского взаимодействия, документооборот, исторические данные организации

OD.03 Датасеты для обучения/дообучения – наборы данных, прошедшие подготовку и используемые для обучения, тестирования и дообучения моделей

Объекты, связанные с моделями

OM.01a Модель (в среде обучения) – итоговая модель, полученная в результате обучения (включая веса, архитектуру, параметры)

OM.01b Модель (в инфраструктуре инференса) – итоговая модель, полученная в результате обучения, размещенная и развернутая в инфраструктуре инференса

OM.01c Модель, размещенная в открытых источниках – проприетарная модель, опубликованная организацией в открытом доступе

OM.02a LLM-адаптер (в среде обучения) – легковесный модуль (адаптер), который содержит небольшое количество обучаемых параметров и применяется для настройки поведения базовой модели

OM.02b LLM-адаптер (в инфраструктуре инференса) – легковесный адаптер, загруженный и применяемый совместно с базовой моделью на этапе инференса для изменения её поведения

Объекты, связанные с AI-решениями

OA.01 AI-приложение – фронтальное приложение для AI-агентов или конечное AI-приложение, использующее AI-модель для решения задач

OA.02 AI-агенты – система на базе GenAI, способная планировать и совершать автономные действия во внешней среде, реагировать на изменения и взаимодействовать с человеком и другими AI-агентами для достижения поставленных целей

OA.03 Функции / инструменты – в контексте AI-приложений - внешние по отношению к AI-решению инструменты, которые приложение может использовать в процессе работы для достижения поставленных целей. В контексте AI-агентов - программный компонент (в т.ч. другой AI-агент), реализующий функциональность и предоставляющий стандартизированный интерфейс доступа (в виде API), вызов которого инициируется на основании ответа LLM

OA.04 Системный промпт – инструкции, определяющие цель, роль и ограничения поведения AI-агента или AI-приложения

OA.05 Конфигурация – параметры настройки AI-агента, включая описания навыков, доступных функций, ограничений, прав

OA.06 Память / кэш – механизмы долгосрочного и краткосрочного хранения информации AI-агентом

Компоненты, данные и артефакты среды исполнения AI-решения

OI.01 Внешние фреймворки и программные компоненты – программные библиотеки, фреймворки (PyTorch, TensorFlow), исходный код, используемые при создании моделей

OI.02 Внешние модели – предобученные модели из открытых репозиториев (Hugging Face, TensorFlow Hub). Могут быть занесены в инфраструктуру обучения для дообучения или в

инфраструктуру размещения и развёртывания инференса модели, а также использоваться AI-приложением или AI-агентом как внешние SaaS

OI.03 Платформенные механизмы предобработки данных – компоненты, выполняющие предобработку, валидацию и фильтрацию данных, поступающих на вход модели

OI.04 Платформенные механизмы постобработки данных – компоненты, выполняющие постобработку, фильтрацию и форматирование результатов модели

OI.05 Логи запросов – журналы, содержащие записи о запросах к модели, ответах модели, вызовах функций, а также метаданные взаимодействий (время, идентификаторы, объём и пр.)

OI.06 Прикладные механизмы предобработки данных – валидация, санитизация и предобработка входных данных, реализуемые на уровне AI-решения

OI.07 Прикладные механизмы постобработки данных – валидация, санитизация и предобработка выходных данных (форматирование ответов, фильтрация нежелательного контента), реализуемые на уровне AI-решения

OI.08 Учетные данные для доступа к моделям – API-ключи, токены доступа, сертификаты, используемые для аутентификации при доступе к модели

OI.09 Логи/трейсы – журналы действий, вызовов функций и взаимодействий AI-приложений и AI-агентов

OI.10 Механизмы обмена сообщениями в мультиагентной системе – протоколы, транспортные каналы, брокеры сообщений, очереди и иные средства, обеспечивающие взаимодействие между AI-агентами, а также между AI-агентами и оркестратором

OI.11 Оркестратор – компонент, управляющий жизненным циклом AI-агентов, координацией их работы, маршрутизацией задач и обеспечением взаимодействия в мультиагентной системе

OI.12 Внешние ресурсы и источники данных – веб-сайты, публичные API, потоковые данные из внешних источников, внешние репозитории навыков

OI.13 Внутренние источники данных – локальные базы данных, векторные хранилища RAG

OI.14 Ресурсные системы – внешние по отношению к приложению сервисы, базы данных, API, файловые хранилища и другие системы, к которым AI-агент или AI-приложение обращается для выполнения своих функций

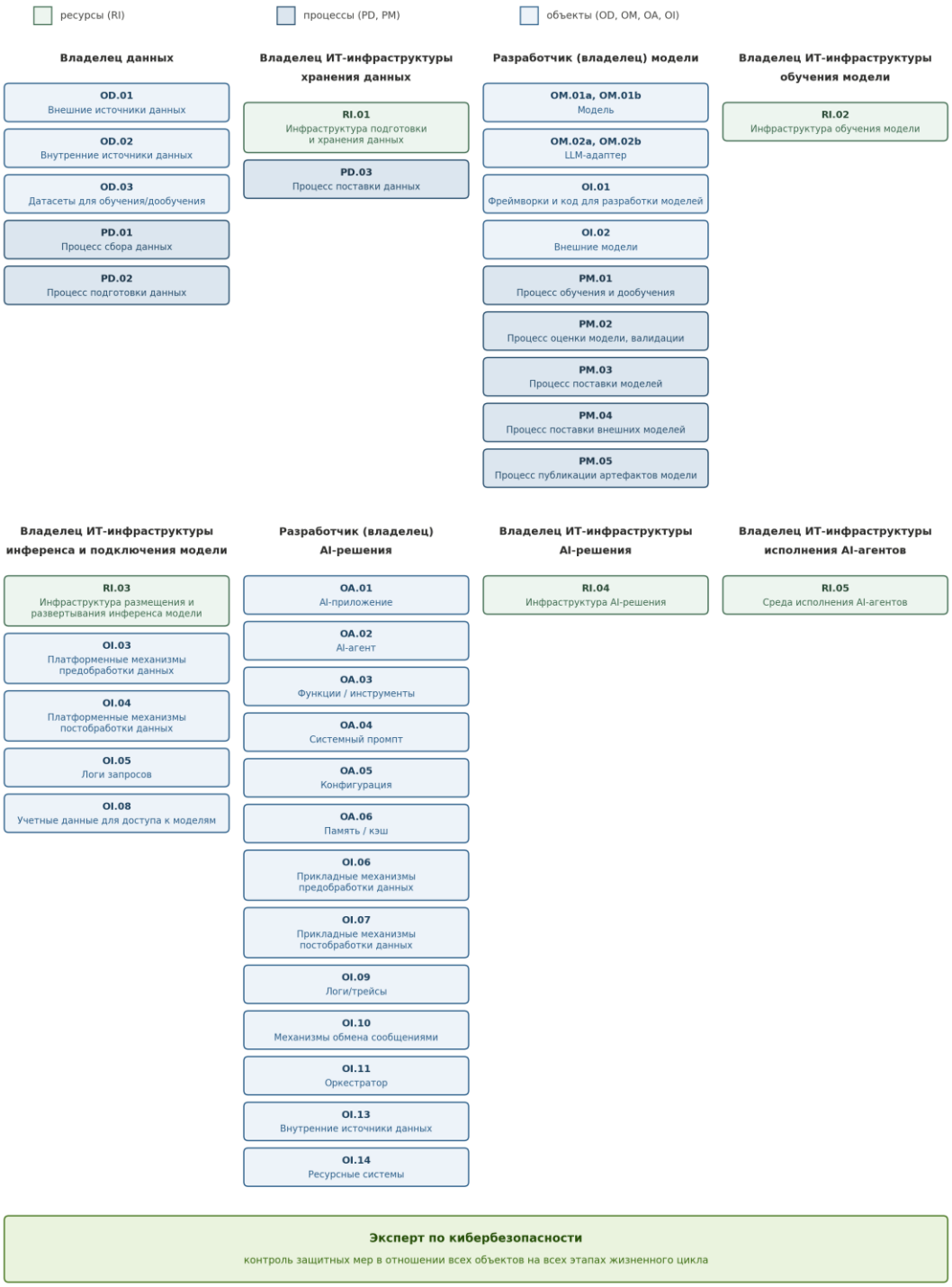
OE.01 Входные данные (API и GUI) – данные, поступающие в AI-решение от пользователей и внешних систем через программные (API) и графические (GUI) интерфейсы. Включают текстовые запросы, структурированные JSON-объекты, мультимодальные данные (изображения, аудио, видео), параметры конфигурации вызова, файлы для обработки, а также команды и управляющие сигналы от пользователя

OE.02 Выходные данные (API и GUI) – данные, возвращаемые AI-решением пользователям и внешним системам через программные (API) и графические (GUI) интерфейсы: ответы модели, сгенерированный контент, результаты вызовов функций. Включают сгенерированный текст, структурированные ответы (JSON, XML), предсказания, классификации, сгенерированные изображения или аудио, логи выполнения, метаданные о качестве ответа (например, уверенность модели), а также ошибки и статусные коды

Роли и зоны ответственности за меры защиты

В модели угроз КБ AI определены следующие роли, отвечающие за меры защиты на различных этапах жизненного цикла AI-решений:

Сводная привязка ролей к каждой угрозе и каждому способу реализации приведена в Приложении 4.



Объекты и роли, отвечающие за меры защиты

- Владелец данных – отвечает за данные на всех этапах: внешние (OD.01) и внутренние (OD.02) источники, датасеты для обучения/тестирования/валидации (OD.03).

Обеспечивает классификацию данных, правомерность использования и обработки (включая соблюдение требований по ПДн), управление жизненным циклом данных в рамках PD.01 (сбор) и PD.02 (подготовка, фильтрация, маскирование), определение политик доступа к данным. При использовании внешних источников (OD.01) проверяет их на наличие ПДн и обеспечивает правомерность их использования, контролирует целостность при загрузке, качество и фильтрацию данных

- Разработчик (владелец) модели – отвечает за процесс разработки, обучения и валидации модели: выбор архитектуры, фреймворки и код (OI.01), модель (OM.01a, OM.01b), LLM-адаптеры (OM.02a, OM.02b), внешние модели (OI.02), тестирование, документирование, обеспечение качества работы, предотвращение уязвимости к состязательным и промпт-атакам, галлюцинаций и нежелательного поведения. Отвечает за внешние модели (OI.02) в случае их дообучения на своих данных — проверяет целостность, сканирует на закладки (ME08, ME09), проводит валидацию поведения (PM.02), контролирует версии. Отвечает за безопасность процессов: PM.01 (обучение/дообучение), PM.02 (оценка и валидация), PM.03 (поставка модели из обучения в инференс), PM.04 (поставка внешних моделей), PM.05 (публикация артефактов модели). В процессах PM.03 и PM.04 обеспечивает шифрование, подписи и контроль целостности передаваемых артефактов
- Владелец ИТ-инфраструктуры хранения данных – отвечает за безопасность систем хранения RI.01 (базы данных, S3, файловые серверы): контроль доступа к датасетам (OD.03), шифрование at rest, резервное копирование, защиту от несанкционированного доступа, модификации или удаления. Участвует в обеспечении безопасности процессов PD.01–PD.03 в части хранения и передачи данных, предоставляет инфраструктуру для изолированного хранения
- Владелец ИТ-инфраструктуры обучения модели – отвечает за безопасность среды обучения RI.02 (ML-кластеры, GPU-серверы, интерактивные среды): изоляцию процессов, контроль доступа к артефактам обучения (OM.01a, OM.02a), безопасность сред разработки. Обеспечивает безопасность процессов PM.01 (обучение/дообучение) и PM.02 (оценка и валидация) в части инфраструктуры, а также участвует в PM.03 (поставка модели) и PM.04 (поставка внешних моделей) при передаче артефактов
- Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента – отвечает за инфраструктуру RI.03, обеспечивающую работу модели и доступ к ней: модель в инференсе (OM.01b, OM.02b), API-шлюзы, балансировщики, аутентификация запросов, защита от DDoS, управление учётными данными (OI.08), логирование (OI.05), платформенные механизмы пред- и постобработки (OI.03, OI.04), настройку безопасного подключения, мониторинг использования и контроль соответствия требованиям (в т.ч. обработка ПДн). Также отвечает за внешние модели (OI.02), если они развёрнуты на собственной инфраструктуре и используются только для инференса (без дообучения) — обеспечивает безопасное подключение, мониторинг, ротацию ключей и защиту от атак на инфраструктуру. Участвует в процессах PM.03 и PM.04 в части развёртывания и эксплуатации
- Владелец ИТ-инфраструктуры AI-решения – отвечает за безопасность инфраструктуры RI.04, обеспечивает изоляцию, разграничение доступа, мониторинг, резервирование, доступность и защиту от несанкционированного доступа в отношении компонентов, размещённых в RI.04. За логику, настройки, бизнес-правила и содержание данных этих компонентов отвечает Разработчик (владелец) AI-решения.
- Владелец ИТ-инфраструктуры исполнения AI-агентов – отвечает за безопасность среды исполнения RI.05: изоляцию (песочницы), контроль доступа к среде, очистку ресурсов (память, временные файлы), защиту от выполнения вредоносного кода, мониторинг активности исполняемых процессов, управление контейнерами и сетевыми политиками на уровне среды. Эти меры применяются ко всем компонентам, размещённым в среде исполнения (функции/инструменты OA.03, конфигурации OA.05, память OA.06, механизмы обмена сообщениями OI.10, оркестратор OI.11 и пр.). За логику, конфигурацию, бизнес-правила и настройки этих компонентов отвечает Разработчик (владелец) AI-решения

- Разработчик (владелец) AI-решения – отвечает за логику, настройки, бизнес-правила и содержание данных всех компонентов AI-решения: приложение (OA.01), агенты (OA.02), системные промпты (OA.04), конфигурации (OA.05), память/кэш (OA.06), функции/инструменты (OA.03), прикладные механизмы пред- и постобработки (OI.06, OI.07), логи/трейсы (OI.09), учётные данные (OI.08), механизмы обмена сообщениями (OI.10), оркестратор (OI.11), внутренние источники данных (OI.13) и ресурсные системы (OI.14). Также отвечает за внешние модели (OI.02) при использовании их как SaaS. За инфраструктурную безопасность всех перечисленных компонентов отвечает владелец ИТ-инфраструктуры AI-решения (RI.04) либо соответствующие владельцы инфраструктур (RI.03, RI.05)
- Эксперт по кибербезопасности – отвечает за анализ угроз и актуализацию модели угроз, внедрение и контроль соблюдения практик безопасной разработки, включая обеспечение безопасности разрабатываемых AI-приложений и AI-агентов, а также на этапах сбора и подготовки данных (PD.01–PD.03), разработки и обучения моделей (PM.01–PM.05), эксплуатации модели и интеграций с AI-решениями, аудит сетевых политик и защиты периметра, проверку сторонних библиотек и фреймворков (OI.01) и внешних моделей (OI.02) на наличие уязвимостей, закладок и вредоносных свойств, мониторинг событий безопасности, реагирование на инциденты и расследование, аудит безопасности инфраструктуры (RI.01–RI.05), данных (OD.01–OD.03), моделей (OM.01a–OM.02b), компонентов AI-решений (OA.01–OA.06) и компонентов среды исполнения (OI.01–OI.14), а также консультирование команд разработки, MLOps, владельцев данных и владельцев решений по вопросам защиты

Перечень угроз кибербезопасности AI

Несанкционированный сбор информации (разведка и утечка данных)

T01. Кража обучающих данных

Описание

Противоправное получение обучающих данных путём копирования из инфраструктуры подготовки и хранения данных, обучения модели

Последствия

Создание нелегитимной копии модели (теневой модели) конкурентом, нарушение конфиденциальности информации и восстановление (инверсия) конфиденциальной информации из обучающих данных, утрата информации, составляющей интеллектуальную собственность

Этап ЖЦ

Сбор и подготовка данных, Разработка модели и обучение

Виды моделей

PredAI, GenAI

Объект защиты

OD.03

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME31, ME32, ME37

T02. Утечка информации о модели и ее параметрах

Описание

Противоправное получение информации об архитектуре модели, ее параметрах, границах принятия решений, используя доступ к открытой или утёкшей документации, репозиториям с моделью, опубликованным в открытых источниках

Последствия

Облегчение проведения целенаправленных состязательных и промпт-атак с использованием знаний уровня white-box, утрата конкурентных преимуществ

Этап ЖЦ

Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OM.01a, OM.02a, OM.01b, OM.02b, OM.01c

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME41, ME02, ME36, ME37, ME47, ME48, ME01

T03. Кража модели или создание копии

Описание

Противоправное получение файлов модели (весов, архитектуры) и создание ее функциональной копии

Последствия

Создание нелегитимной копии модели (теневой модели) конкурентом, возможность проведения white-box атак, восстановление (инверсия) обучающих данных, утрата информации, составляющей интеллектуальную собственность

Этап ЖЦ

Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OM.01a, OM.02a, OM.01b, OM.02b

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME41, ME36, ME37, ME47, ME48

T04. Утечка информации о среде исполнения AI-приложения или AI-агента

Описание

Противоправное получение информации о конфигурации системы, внутренних сетевых адресах, версиях ПО, переменных окружения

Последствия

Облегчение проведения целенаправленных атак на инфраструктуру, компрометация учётных данных, эскалация привилегий

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

RI.04, RI.05, OA.05

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME23, ME13, ME14, ME15, ME51

T05. Извлечение информации из системного промпта для проведения целевых кибератак

Описание

Противоправное получение детальной информации о логике работы, ограничениях и уязвимостях AI-решения из его системного промпта

Последствия

Повышение эффективности атак на AI-решение, возможность обхода механизмов защиты

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.04

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME22, ME13, ME14, ME15, ME51, ME40

T06. Утечка информации об архитектуре мультиагентной системы

Описание

Противоправное получение информации о составе мультиагентной системы, ролях агентов, правилах их взаимодействия

Последствия

Облегчение проведения целенаправленных атак на мультиагентную систему, раскрытие архитектурных особенностей

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.02

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME22, ME23, ME13, ME14, ME15, ME51

T07. Утечка информации о конфигурации AI-агента

Описание

Противоправное получение конфиденциальных параметров настройки AI-агента, включая доступные функции, ограничения, права

Последствия

Раскрытие внутренней логики работы AI-агента, возможность создания копии с аналогичным поведением, облегчение проведения атак

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.05

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME22, ME23, ME13, ME14, ME15, ME51

Утечка конфиденциальной информации

T08. Утечка информации о системном промпте (в т.ч. цели, функциях, бизнес-логике) AI-приложения или AI-агента, приводящая к потере уникальности

Описание

Противоправное получение системного промпта AI-приложения или агента, содержащего инструкции, цели и описание бизнес-логики

Последствия

Утрата конкурентных преимуществ, копирование уникальной логики работы

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.04

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME33, ME22, ME23, ME46, ME13, ME14, ME15, ME51, ME40

T09. Утечка конфиденциальной информации из наборов данных (обучающих, тестовых, валидационных)

Описание

Противоправное получение конфиденциальной информации, содержащейся в датасетах (обучающих, тестовых, валидационных), включая персональные данные, коммерческую или иную тайну

Последствия

Раскрытие конфиденциальной информации, репутационный ущерб, финансовые потери, риски нарушения законодательства о персональных данных (ФЗ-152)

Этап ЖЦ

Сбор и подготовка данных, Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OD.03

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME30, ME31, ME32, ME37, ME49

T10. Утечка конфиденциальной информации из систем логирования

Описание

Противоправное получение информации из систем логирования (в том числе логирования запросов и вызовов функций), содержащей конфиденциальную информацию или персональные данные пользователей, детали реализации AI-агентов и внутренние параметры системы. Угроза актуальна как при использовании внутренней инфраструктуры развертывания модели, так и при использовании внешних MLaaS-сервисов.

Последствия

Компрометация пользовательских данных, раскрытие архитектуры и логики работы AI-решения, облегчение проведения дальнейших атак, риски нарушения законодательства о персональных данных (ФЗ-152)

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OI.05, OI.09

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME45, ME47, ME51

T11. Утечка конфиденциальной информации из дообученной модели или LLM-адаптера

Описание

Противоправное получение конфиденциальной информации, запомненной моделью/LLM-адаптером в процессе дообучения

Последствия

Раскрытие персональных данных, коммерческой тайны, риски нарушения ФЗ-152

Этап ЖЦ

Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OM.01a, OM.02a, OM.01b, OM.02b

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME30, ME41, ME33, ME11, ME36, ME49, ME46, ME42, ME47, ME50, ME13, ME14, ME27, ME15, ME51, ME28

T12. Утечка конфиденциальной информации из внутренних источников данных (базы данных, RAG)

Описание

Противоправное получение конфиденциальной информации из баз данных, векторных хранилищ и иных внутренних источников в результате прямого извлечения нарушителем; передачи данных через API-интеграции (back2back, межагентские взаимодействия, запросы агентов во внешние системы/Интернет); использования внешних GenAI-моделей (SaaS, API внешних LLM, облачные сервисы), когда в процессе вызова модели в запросе или контексте передаются конфиденциальные данные, а нарушитель перехватывает их или модель сохраняет их в своих логах/обучающих выборках.

Последствия

Раскрытие персональных данных, коммерческой тайны, интеллектуальной собственности

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OI.13, OI.14

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME33, ME35, ME46, ME47, ME50, ME13, ME14, ME26, ME27, ME15, ME51, ME28

T13. Утечка конфиденциальной информации из механизмов памяти

Описание

Противоправное получение конфиденциальной информации, хранящейся в механизмах долговременной и краткосрочной памяти AI-агентов

Последствия

Раскрытие приватных данных пользователей, внутренних параметров и истории взаимодействий

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.06

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME33, ME23, ME47, ME13, ME14, ME15, ME51

T14. Компрометация учетных данных и ключей доступа к модели

Описание

Противоправное получение API-ключей, токенов доступа или сертификатов, используемых для аутентификации при доступе к модели

Последствия

Несанкционированный доступ к модели, возможность проведения атак от имени легитимного пользователя

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OI.08

Нарушаемое свойство

Конфиденциальность

Способы реализации

ME22, ME23, ME42, ME13, ME14, ME15, ME51

Несанкционированное воздействие на информационные ресурсы и среду исполнения

T15. Удаление или уничтожение обучающих, тестовых или валидационных датасетов

Описание

Утрата, удаление или уничтожение наборов данных, предназначенных для обучения, тестирования и валидации моделей, вследствие отсутствия надлежащего управления жизненным циклом данных

Последствия

Невозможность воспроизведения результатов обучения, увеличение поверхности атаки, утрата информации, составляющей интеллектуальную собственность

Этап ЖЦ

Сбор и подготовка данных, Разработка модели и обучение

Виды моделей

PredAI, GenAI

Объект защиты

OD.03

Нарушаемое свойство

Целостность; Доступность

Способы реализации

ME31

T16. Использование для обучения/дообучения или оценки модели отравленных данных или датасетов

Описание

Включение в обучающие или валидационные датасеты данных, содержащих вредоносные примеры, преднамеренно внедрённые нарушителем

Последствия

Систематическая предвзятость (bias) в решениях модели, снижение точности, внедрение бэкдоров для последующего манипулирования выводами

Этап ЖЦ

Сбор и подготовка данных, Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OM.01a, OM.01b, OM.02a, OM.02b

Нарушаемое свойство

Целостность

Способы реализации

ME04, ME05, ME06, ME31, ME32, ME37

T17. Внедрение и выполнение вредоносного кода в компонентах ИТ-инфраструктуры AI

Описание

Загрузка и последующее выполнение вредоносного кода или программного обеспечения в компонентах ИТ-инфраструктуры AI: инфраструктуре подготовки и хранения данных, инфраструктуре обучения, инференса, инфраструктуре AI-решения и среде исполнения AI-агентов (RI.01–RI.05)

Последствия

Компрометация среды выполнения и получение контроля для дальнейших атак на инфраструктуру

Этап ЖЦ

Сбор и подготовка данных, Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

RI.01, RI.02, RI.03, RI.04, RI.05

Нарушаемое свойство

Целостность; Доступность

Способы реализации

ME07, ME08, ME10, ME43, ME44

T18. Внедрение вредоносного контента и промпт-атак во внутренние источники, описания функций, инструментов и память AI-агентов

Описание

Размещение во внутренних источниках данных (базы знаний RAG, векторные хранилища, долговременная память AI-агента, общая или омниканальная память), а также в функции/инструменты AI-агента (включая как отравление их метаданных (описаний, параметров, схем), так и модификацию кода самих функций/инструментов или навыков AI-агентов) вредоносного контента двух типов: промпт-атак, направленных на изменение поведения системы, и ложных или искажённых фактов

Последствия

Нарушение целостности базы знаний и памяти AI-агента, каскадное распространение искажений при мультиагентном взаимодействии посредством промежуточного хранилища

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.06, OA.03, OI.13

Нарушаемое свойство

Целостность

Способы реализации

ME35, ME23, ME13, ME24, ME29

T19. Удаление или модификация данных вследствие выполнения вредоносных инструкций

Описание

Несанкционированная модификация или уничтожение информации в базах данных, файловых хранилищах и иных внутренних источниках

Последствия

Нарушение целостности и доступности критически важных данных, сбой бизнес-процессов, утрата информации

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OI.13, OI.14

Нарушаемое свойство

Целостность; Доступность

Способы реализации

ME33, ME13, ME14, ME26, ME15, ME29

T20. Нарушение изоляции среды выполнения с получением контроля над базовой инфраструктурой

Описание

Нарушение границ изоляции (песочницы) среды выполнения AI-агента с получением несанкционированного доступа к нижележащей инфраструктуре

Последствия

Компрометация хост-системы, эскалация привилегий, получение контроля над смежными компонентами

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

RI.04, RI.05

Нарушаемое свойство

Целостность; Доступность

Способы реализации

ME12, ME26, ME43, ME44

T21. Удаление или модификация файлов в среде исполнения функций AI-агента

Описание

Несанкционированное удаление или изменение файлов в среде выполнения AI-агента в результате вызова функций. Реализация возможна как при злонамеренной компрометации AI-агента (например, через промпт-атаку, подмену системного промпта), так и вследствие ошибок проектирования, ошибок в логике агента или некорректных инструкций, приводящих к нецелевому вызову функций удаления/модификации

Последствия

Нарушение целостности данных, необходимых для работы агента, потеря временных данных, сбой функциональности

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.03, OI.13, OI.14

Нарушаемое свойство

Целостность; Доступность

Способы реализации

ME12, ME13, ME14, ME26, ME15, ME43, ME44, ME29

Отказ в обслуживании

T22. Нарушение доступности модели

Описание

Нарушение доступности сервиса модели, вызванное исчерпанием вычислительных ресурсов или пропускной способности при обработке запросов

Последствия

Прекращение или критическое снижение скорости оказания услуг моделью, финансовые потери, простои бизнес-процессов

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OM.01b

Нарушаемое свойство

Доступность

Способы реализации

ME18, ME19

T23. Нарушение доступности (DoS) или исчерпание ресурсов (DoW) AI-решения

Описание

Нарушение доступности AI-решения вследствие исчерпания вычислительных ресурсов или превышения лимитов токенов

Последствия

Отказ в обслуживании пользователей, неконтролируемый рост затрат на эксплуатацию, простои бизнес-процессов

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OA.01, OA.02

Нарушаемое свойство

Доступность

Способы реализации

ME11, ME19, ME13, ME14, ME21, ME15

T24. Сбой интеграции AI-приложений, AI-агентов или компонентов системы

Описание

Нарушение работоспособности интеграций между компонентами AI-решения, вызванное искажением формата или синтаксиса структурированных данных при межкомпонентном обмене

Последствия

Отказ в обслуживании отдельных компонентов, нарушение целостности данных, каскадные сбои в работе системы

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OA.01, OA.02

Нарушаемое свойство

Доступность

Способы реализации

ME33, ME12, ME21, ME16

T25. Создание операционных помех и саботаж процессов за счет перегрузки человеческого звена

Описание

Дестабилизация работы персонала путем генерации AI-решением избыточного количества ложных или нерелевантных уведомлений. Угроза может быть реализована как преднамеренно (внешний или внутренний нарушитель через атаки типа DoS, промпт-атаки, модификацию конфигураций), так и непреднамеренно – вследствие ошибок проектирования, некорректной настройки порогов срабатывания, галлюцинаций модели или «человеческого фактора» (неквалифицированные действия разработчика/оператора). В обоих случаях последствия для бизнес-процессов идентичны.

Последствия

Снижение эффективности работы сотрудников, нарушение бизнес-процессов, пропуск критически важных уведомлений

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OA.01, OA.02

Нарушаемое свойство

Доступность

Способы реализации

ME40, ME13, ME14, ME15, ME25, ME17

T26. Нарушение доступности среды исполнения AI-агента

Описание

Нарушение доступности среды исполнения AI-агента (в т.ч. функций), вызванное исчерпанием вычислительных ресурсов, памяти или пропускной способности сети

Последствия

Отказ в обслуживании всех агентов, размещенных в среде исполнения, невозможность выполнения ими задач

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

RI.05

Нарушаемое свойство

Доступность

Способы реализации

ME19, ME12, ME13, ME14, ME15, ME44, ME17

T27. Нарушение доступности или сбой приложения, реализующего AI-агента

Описание

Нарушение доступности или отказ в обслуживании пользовательского (фронтального) приложения, вызванное некорректной работой интегрированного в него AI-агента

Последствия

Невозможность использования приложения пользователями, простой бизнес-процессов, зависимых от AI-функциональности

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.01

Нарушаемое свойство

Доступность

Способы реализации

ME19, ME12, ME13, ME14, ME15, ME16, ME17

T28. Нарушение доступности ресурсной системы AI-агента

Описание

Нарушение доступности внешних ресурсов (баз данных, API, сервисов), от которых зависит функционирование AI-агента

Последствия

Невозможность выполнения агентом своих функций, отказ в обслуживании, каскадные сбои в зависимых процессах

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OI.14

Нарушаемое свойство

Доступность

Способы реализации

ME19, ME12, ME13, ME14, ME26, ME15, ME17, ME29

Нарушение мультиагентного взаимодействия

T29. Нарушение цели другого AI-агента при кооперативном взаимодействии в мультиагентной системе

Описание

Изменение целевой функции или поведенческих установок AI-агента в результате получения сообщения от другого (скомпрометированного) AI-агента

Последствия

Нарушение бизнес-процессов, нарушение кооперации агентов, деградация общей производительности системы

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.02

Нарушаемое свойство

Целостность

Способы реализации

ME33, ME13, ME14, ME20, ME24, ME15, ME16

T30. Каскадное распространение промпт-атак на AI-приложения или AI-агентов в мультиагентной системе

Описание

Распространение промпт-атак на AI-приложения или AI-агентов в мультиагентной системе через общие каналы взаимодействия, хранилища данных или память

Последствия

Каскадная компрометация множества агентов, каскадное распространение промпт-атаки, нарушение работы мультиагентной системы, нарушение бизнес-процессов, эскалация привилегий AI-агентов

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.01, OA.02

Нарушаемое свойство

Целостность

Способы реализации

ME33, ME20, ME24, ME16, ME28, ME29

Распространение противоправной информации

T31. Генерация противоправной, токсичной или неэтичной информации AI-решением

Описание

Генерация AI-решением контента, нарушающего законодательство или морально-этические нормы

Последствия

Репутационный ущерб, правовые риски, возможность использования AI-решения для подготовки и совершения противоправных действий

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.01, OA.02

Нарушаемое свойство

Целостность

Способы реализации

ME38, ME39, ME40, ME13, ME14, ME15

T32. Использование AI-приложения или AI-агента для подготовки проведения киберпреступлений

Описание

Использование AI-решения для генерации идей, кода, фишинговых писем или инструкций по совершению киберпреступлений, а также для автономного выполнения многошаговых кибератак (например, эксплуатация цепочки уязвимостей, латеральное перемещение, противодействие расследованию), включая сценарии, когда GenAI-модели обладают возможностью для генерации вредоносных инструкций или автономного проведения атак при наличии соответствующих инструментов у AI-агента или AI-приложения

Последствия

Использование ресурсов организации в противоправных целях, правовые и репутационные риски, а также автономное проведение кибератак на внешние и внутренние системы с использованием возможностей AI-агента и внешних моделей

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.01, OA.02

Нарушаемое свойство

Целостность

Способы реализации

ME09, ME38, ME39, ME40, ME13, ME14, ME15

Нарушение функционирования (работоспособности)

T33. Некорректное или недекларированное поведение модели, галлюцинации

Описание

Функционирование модели с отклонениями от ожидаемого поведения, включая генерацию фактологически неверных ответов или генерацию инструкций, которые при использовании в AI-приложении или AI-агенте могут привести к вредоносным действиям (например, эксплуатации уязвимостей).

Последствия

Подрыв доверия к AI-решению, принятие ошибочных решений на основе неверных данных, репутационные риски, а также компрометация внешних систем при реализации сгенерированных моделью вредоносных инструкций.

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OM.01b

Нарушаемое свойство

Целостность

Способы реализации

ME02, ME03, ME04, ME05, ME06, ME09, ME31, ME34, ME36, ME38, ME39, ME41

T34. Смещение результатов работы модели, снижение точности, создание бэкдоров и искажение результатов работы модели

Описание

Модификация (искажение) модели машинного обучения, приводящая к систематической предвзятости, ухудшению ключевых метрик, внедрению скрытых возможностей контроля над выводами модели

Последствия

Некорректная работа модели в целевых сценариях, возможность скрытого манипулирования решениями

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OM.01b, OM.02b

Нарушаемое свойство

Целостность

Способы реализации

ME02, ME03, ME04, ME05, ME06, ME09, ME31, ME36, ME38, ME39, ME41

T35. Нарушение логики выполнения задачи, нарушение рабочего процесса

Описание

Изменение логики функционирования AI-приложения или AI-агента, приводящее к выполнению действий, не предусмотренных бизнес-сценарием, игнорированию системных инструкций и нарушению рабочего процесса

Последствия

Неконтролируемое поведение системы, сбой бизнес-процессов, выполнение запрещённых или нежелательных действий

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

OA.01, OA.02

Нарушаемое свойство

Целостность

Способы реализации

ME33, ME34, ME22, ME38, ME39, ME40, ME13, ME14, ME20, ME15, ME16, ME25, ME17, ME29

T36. Использование AI-решения для целей, непредусмотренных бизнес-сценарием

Описание

Использование AI-решения для выполнения задач, выходящих за рамки его функционального назначения

Последствия

Нецелевое расходование ресурсов, потенциальные правовые нарушения, репутационный ущерб

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект защиты

OA.01, OA.02

Нарушаемое свойство

Целостность

Способы реализации

ME38, ME39, ME40, ME13, ME14, ME15

Нарушение процессов мониторинга и расследования

T37. Невозможность или несвоевременное выявление, реагирование и расследование событий кибербезопасности

Описание

Отсутствие или неполнота данных логирования взаимодействий, документации об обучающих данных и информации о состоянии инференсов моделей, что препятствует обнаружению и расследованию событий и инцидентов

Последствия

Критическое затруднение или невозможность обнаружения факта компрометации, определения причин инцидента и восстановления хронологии событий, увеличение ущерба

Этап ЖЦ

Сбор и подготовка данных, Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект защиты

RI.01, RI.02, RI.03, RI.04, RI.05

Нарушаемое свойство

Доступность

Способы реализации

ME34, ME11, ME45

Потенциальные нарушители кибербезопасности AI

Потенциал нарушителей, уровни технических возможностей, осведомлённости и доступа

Потенциал нарушителя КБ AI определяется совокупностью двух независимых факторов:

Уровень технических возможностей (компетенции, инструментарий и доступные ресурсы для реализации атакующих воздействий).

Уровень осведомлённости и доступа (объём внутренней информации об архитектуре, данных, коде и конфигурациях AI-решения, а также характер легитимного взаимодействия с его компонентами).

Потенциал выражается в трёх градациях – низкий, средний, высокий – и определяется по правилу комбинирования уровней указанных факторов с учётом оснащённости нарушителя КБ AI и широты охвата объектов воздействия:

- Низкий – присваивается при H1 в сочетании с black- или grey-box. Нарушитель КБ AI ограничен публичными интерфейсами, не обладает внутренней информацией о системе и не имеет профильной компетенции, что делает его действия малоэффективными и неспособными нанести существенный урон.
- Средний – присваивается при H2 и любом уровне осведомлённости, либо при H1 в сочетании с white-box. Нарушитель КБ AI опасен в пределах своей функциональной зоны, но не обладает ресурсами и широтой охвата для сложных межкомпонентных кампаний.
- Высокий – присваивается при H3 независимо от уровня осведомлённости (grey-box, white-box, black-box). Ресурсы, время и компетенция нарушителя с уровнем H3 (продвинутые внешние группы, APT, специальные службы) компенсируют отсутствие внутренней информации об архитектуре и данных, позволяя проводить сложные целевые атаки через публичные интерфейсы (длительный реверс-инжиниринг, анализ поведения, подбор эксплойтов).

Уровень технических возможностей	Обозначение	Описание
Низкий	H1	Нарушитель имеет ограниченные навыки, действует в основном методом социальной инженерии или использует автоматизированные скрипты без глубокого понимания системы.
Средний	H2	Нарушитель имеет базовые навыки в области кибератак, может использовать готовые инструменты и эксплойты, а также обладает легитимным доступом к части компонентов.
Высокий	H3	Нарушитель обладает продвинутыми техническими навыками, ресурсами и временем для проведения сложных атак (например, разработка специализированного вредоносного ПО, проведение реверс-инжиниринга, длительные целевые кампании).

Уровень осведомлённости и доступа	Описание
Black-box	Нарушитель не имеет внутренней информации об архитектуре, данных, весах модели, может только взаимодействовать с системой через публичные интерфейсы (API, веб-формы).
Grey-box	Нарушитель обладает частичной информацией (например, документацией, архитектурой, отдельными фрагментами данных) или имеет легитимный

	доступ с ограниченными привилегиями к внутренним компонентам системы (инфраструктура, хранилища данных, артефакты обучения, конфигурации).
White-box	Нарушитель имеет полный доступ к обучающим данным, исходным кодам, весам модели, конфигурациям и инфраструктуре.

Виды нарушителей кибербезопасности AI и их потенциал

Нарушители КБ AI разделяются на две основные категории: внешние и внутренние. Внутренние нарушители КБ AI дополнительно детализированы по ролям в рамках жизненного цикла разработки AI-решений.

Код	Вид нарушителя КБ AI	Уровень технических возможностей	Уровень осведомлённости и доступа	Потенциал[5]
H1.1	Внешние преступные группы	H2	Black-/Grey-box	Средний
H1.2	Внешние субъекты (физические лица)	H1	Black-/Grey-box	Низкий
H1.3	Конкурирующие организации	H2	Black-/Grey-box	Средний
H1.4	Бывшие работники (администраторы, разработчики)	H2	Grey-box	Средний
H1.5	Продвинутые внешние группы (APT, специальные службы)	H3	Grey-box	Высокий
H2.1	Пользователи	H1	Black-box	Низкий
H2.2	Администраторы информационных ресурсов	H1–H2	White-box	Средний
H2.3	Администраторы безопасности	H1–H2	Grey-box	Средний
H2.4	D-People (Data Owner)	H2	White-box	Средний
H2.5	D-People (Data Engineer)	H2	White-box	Средний
H2.6	D-People (Data Scientist)	H2	White-box	Средний
H2.7	D-People (ML engineer)	H2	White-box	Средний
H2.8	D-People (DevOps engineer)	H2	White-box	Средний
H2.9	D-People (Разработчик)	H2	White-box	Средний
H2.10	AI-агент[6]	H2	Grey-box	Средний

[5] Приведённые оценки потенциала являются базовыми (референтными). Указанные уровни могут быть скорректированы в сторону повышения или понижения на основании внутренних данных организации об актуальных угрозах, особенностях AI-решения и реализованных мерах кибербезопасности.

[6] AI-агент может выступать в двух ролях: автономно (непреднамеренные действия, ошибки, галлюцинации) или скомпрометировано (выполнение вредоносных инструкций от внешнего или внутреннего нарушителя). В таблицах угроз и способов реализации это не разделяется, но при выборе мер защиты следует учитывать обе возможности.

Соответствие видов нарушителей угрозам кибербезопасности AI и способам их реализации

Привязка нарушителей к угрозам и способам их реализации выполняется в два шага:

1. для каждой угрозы определяются те виды нарушителей, которые обладают мотивацией на наступление соответствующего негативного последствия (цели нарушителей приведены в Приложении 2);
2. для каждого способа реализации определяются виды нарушителей, обладающие возможностью его выполнить — исходя из уровня технических возможностей, уровня осведомлённости и характера доступа к объекту воздействия.

Полные перечни видов нарушителей по каждой угрозе и каждому способу реализации приведены в Приложении 3.

Перечень способов реализации угроз кибербезопасности AI

Reconnaissance

ME01. Получение данных о модели через анализ документации и файлов модели, размещенных в открытых источниках

Описание

Нарушитель, эксплуатируя недостатки процесса публикации артефактов модели (отсутствие контроля состава публикуемых материалов и их проверки на наличие чувствительной информации), изучает общедоступные артефакты (технические отчёты, спецификации, конфигурационные файлы, веса модели), чтобы понять архитектуру и особенности целевой модели для последующего подбора атак

Этап ЖЦ

Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

PM.05 (реализуется через: OM.01c)

Реализуемые угрозы КБ AI

T02

Resource Development

ME02. Несанкционированное получение или изменение параметров процессов обучения/дообучения и оценки (валидации) модели

Описание

Нарушитель, используя недостатки разграничения доступа к среде обучения или через скомпрометированные учетные записи, извлекает из логов, конфигурационных файлов и иных артефактов информацию о процессах обучения/дообучения и оценки модели и их параметрах (алгоритм оптимизации, функция потерь, метрики) или изменяет их: гиперпараметры обучения, конфигурации процесса оценки (состав используемых датасетов, метрики, пороговые значения, результаты оценки) для влияния на итоговую модель

Этап ЖЦ

Разработка модели и обучение

Виды моделей

PredAI, GenAI

Объект воздействия

PM.01, PM.02 (реализуется через: RI.02)

Реализуемые угрозы КБ AI

T02, T33, T34

ME03. Компрометация публичных репозиториев моделей путём подмены ссылок, метаданных или внедрения вредоносных конфигураций

Описание

Нарушитель, эксплуатируя недостатки процесса поставки внешних моделей (отсутствие или недостатки механизмов контроля целостности и неизменности загружаемых из публичных репозиториев моделей, например, проверки контрольных сумм и верификации источника), подменяет ссылки на используемые модели, изменяет их метаданные или внедряет вредоносные конфигурационные файлы. При загрузке такой модели в инфраструктуру вредоносный код или логическая закладка попадают в среду разработки и далее в конечное решение

Этап ЖЦ

Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

PM.04 (реализуется через: OI.02)

Реализуемые угрозы КБ AI

T33, T34

Initial Access

ME04. Отравление данных во внешних источниках и эксплуатация недостатков контроля процесса их загрузки

Описание

Нарушитель внедряет вредоносные примеры в общедоступные датасеты или иные внешние источники, а затем эксплуатирует недостатки процесса сбора данных (недостатки процедур проверки при загрузке этих данных в обучающую выборку), что приводит к включению отравленных данных в процесс обучения модели

Этап ЖЦ

Сбор и подготовка данных

Виды моделей

PredAI, GenAI

Объект воздействия

PD.01 (реализуется через: OD.01)

Реализуемые угрозы КБ AI

T16, T33, T34

ME05. Использование недостатков контроля неизменности данных при отложенной загрузке из внешних источников

Описание

Нарушитель модифицирует уже одобренный внешний источник данных после его включения в реестр используемых источников данных (например, подменяет файл на сервере или изменяет содержимое по ссылке) и, эксплуатируя недостатки процесса сбора данных (отсутствие или недостатки механизмов контроля неизменности при отложенной загрузке, например, проверки

контрольных сумм), внедряет отравленные данные в датасеты. При последующей отложенной загрузке модель обучается на изменённых, потенциально вредоносных данных

Этап ЖЦ

Сбор и подготовка данных

Виды моделей

PredAI, GenAI

Объект воздействия

PD.01 (реализуется через: OD.01)

Реализуемые угрозы КБ AI

T16, T33, T34

ME06. Отравление данных во внутренних источниках и эксплуатация недостатков контроля процесса их загрузки

Описание

Нарушитель, используя недостатки разграничения доступа, внедряет вредоносные примеры во внутренние хранилища данных (базы данных, документы) и, эксплуатируя недостатки процесса подготовки данных (обработки, проверки и фильтрации), добивается их попадания в наборы данных для обучения модели, вызывая её отравление

Этап ЖЦ

Сбор и подготовка данных

Виды моделей

PredAI, GenAI

Объект воздействия

PD.02 (реализуется через: OD.02)

Реализуемые угрозы КБ AI

T16, T33, T34

ME07. Использование уязвимостей сторонних библиотек, использованных при подготовке данных, разработке модели или приложения

Описание

Нарушитель эксплуатирует известные уязвимости или закладки в сторонних библиотеках и фреймворках, применяемых при подготовке данных и обучении модели

Этап ЖЦ

Сбор и подготовка данных, Разработка модели и обучение

Виды моделей

PredAI, GenAI

Объект воздействия

RI.01, RI.02 (реализуется через: OI.01)

Реализуемые угрозы КБ AI

T17

ME08. Эксплуатация закладок, заложенных в файлы или веса используемых внешних моделей

Описание

Нарушитель использует программные закладки, активирующиеся при загрузке или инференсе, внедренные в файлы (включая архитектуру, конфигурации или веса) предобученных моделей из открытых источников и не выявленные из-за недостатков процесса поставки внешних моделей (отсутствие проверки и сканирования загружаемых файлов)

Этап ЖЦ

Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

PM.04 (реализуется через: OI.02)

Реализуемые угрозы КБ AI

T17

ME09. Эксплуатация логических закладок или вредоносных свойств, заложенных в используемые внешние модели

Описание

Нарушитель эксплуатирует нежелательные свойства предобученной модели, возникшие в результате целенаправленных действий разработчика или поставщика: логические закладки (бэкдоры), активирующиеся при подаче специфических входных данных и приводящие к заданному нежелательному поведению; способность генерировать вредоносные инструкции, эксплойты или автономно инициировать атаки при определённых условиях (модели со скрытым вредоносным поведением, т.н. trojan- или sleeper-модели). Применение способа реализации возможно из-за недостатков процессов поставки внешних моделей (PM.04) и оценки, валидации модели (PM.02), не выявляющих такие свойства до ввода модели в эксплуатацию

Этап ЖЦ

Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

PM.04 (реализуется через: OI.02)

Реализуемые угрозы КБ AI

T33, T34, T32

ME10. Загрузка вредоносного программного обеспечения из внешних источников (Интернет)

Описание

Нарушитель, обладая знаниями, что AI-решение загружает и анализирует данные из внешних источников (в том числе баз данных, репозиториев кода или веб-сайтов), размещает в таких источниках вредоносный код. При последующей загрузке и наличии возможности выполнения

кода происходит внедрение и выполнение вредоносного кода и ВПО в среде исполнения модели, AI-приложения или AI-агента

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OA.03 (реализуется через OI.12)

Реализуемые угрозы КБ AI

T17

AI Model Access

ME11. Несанкционированные подключения к модели

Описание

Нарушитель, используя скомпрометированные учётные данные или посреднические сервисы, подключается к модели, минуя установленные процессы подключения и механизмы мониторинга

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

RI.03 (в отношении OM.01b, OM.02b)

Реализуемые угрозы КБ AI

T37, T11, T23

Execution

ME12. Эксплуатация ошибок проектирования и небезопасных интеграций компонентов AI-приложений или AI-агентов

Описание

Нарушитель выявляет и эксплуатирует ошибки в архитектуре приложения (недостаточная проверка данных, небезопасные вызовы API, вызовы функций без проверки прав), что приводит к компрометации системы

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

OA.01, OA.02

Реализуемые угрозы КБ AI

T20, T24, T21, T26, T27, T28

ME13. Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет)

Описание

Нарушитель, обладая знаниями, что AI-решение загружает данные из внешних источников (включая файлы описаний навыков), размещает в таких источниках документы, содержащие промпт-атаки или отравленные файлы навыков. При последующей обработке загруженного контента: промпт-атаки активируются в контексте обработки, отравленные навыки сохраняются и выполняются, что при вызове приводит к нежелательному поведению

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OA.03 (реализуется через OI.12)

Реализуемые угрозы КБ AI

T11, T18, T12, T19, T23, T35, T08, T31, T04, T32, T05, T36, T13, T25, T14, T21, T26, T06, T29, T27, T07, T28

ME14. Реализация прямых промпт-атак или состязательных атак

Описание

Нарушитель через интерфейс пользовательского ввода направляет модели специально сконструированные запросы для обхода ограничений, раскрытия информации или выполнения нежелательных действий

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

OI.06 (реализуется через: OE.01)

Реализуемые угрозы КБ AI

T11, T12, T19, T23, T35, T08, T31, T04, T32, T05, T36, T13, T25, T14, T21, T26, T06, T29, T27, T07, T28

ME15. Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников

Описание

Нарушитель реализует внедренную ранее промпт-атаку в момент, когда AI-агент или AI-приложение извлекает отравленные данные из БД RAG, долговременной памяти или общих хранилищ для формирования контекста запроса

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OI.13, OA.06

Реализуемые угрозы КБ AI

T11, T12, T19, T23, T35, T08, T31, T04, T32, T05, T36, T13, T25, T14, T21, T26, T06, T29, T27, T07, T28

ME16. Эксплуатация недостатков протоколов межагентского взаимодействия

Описание

Нарушитель, используя уязвимости в протоколах обмена между AI-агентами (например, MCP), внедряет ложные сообщения, изменяет состояние взаимодействия или заставляет выполнить вредоносные действия

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OI.10

Реализуемые угрозы КБ AI

T35, T24, T29, T30, T27

ME17. Эксплуатация уязвимостей оркестратора (планировщика) мультиагентной системы для нарушения координации агентов

Описание

Нарушитель, анализируя протоколы взаимодействия с оркестратором мультиагентной системы или эксплуатируя известные уязвимости в его компонентах, отправляет специально сформированные запросы, нарушающие логику распределения задач, что приводит к дедлокам, неправильному назначению заданий или отказу в обслуживании всей системы

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OI.11

Реализуемые угрозы КБ AI

T35, T25, T26, T27, T28

ME18. DDoS-атаки на инфраструктуру развертывания модели

Описание

Нарушитель организует массированные запросы к API модели или использует уязвимости инфраструктуры развертывания, вызывая отказ в обслуживании

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

RI.03 (реализуется через: OE.01)

Реализуемые угрозы КБ AI

T22

ME19. Эскалация потребления токенов/вычислительных ресурсов через массированные или ресурсоемкие запросы

Описание

Нарушитель, используя отсутствие или недостатки ограничений запросов к API, отправляет специально сконструированные запросы (например, с длинными последовательностями или требующие сложных вычислений), что приводит к высокому потреблению ресурсов и росту затрат или отказу в обслуживании

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

OI.06, OI.03

Реализуемые угрозы КБ AI

T22, T23, T26, T27, T28

ME20. Искажение смысла сообщений и инструкций, передаваемых между AI-приложениями, AI-агентами или компонентами системы

Описание

Нарушитель, используя недостатки защиты каналов межкомпонентного взаимодействия или уязвимости в механизмах обработки и маршрутизации межкомпонентных сообщений, перехватывает трафик и модифицирует семантическое содержание передаваемых инструкций, что приводит к нарушению кооперации AI-агентов, выполнению ошибочных действий или нарушению логики AI-решения

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

RI.04 (в отношении: OI.10, OI.11)

Реализуемые угрозы КБ AI

T35, T29, T30

ME21. Искажение формата или синтаксиса структурированных данных, передаваемых между компонентами AI-решения

Описание

Нарушитель, используя недостатки защиты каналов межкомпонентного взаимодействия или уязвимости в механизмах обработки и маршрутизации межкомпонентных сообщений, изменяет формат или синтаксис структурированных данных (JSON, XML и др.) при их передаче между компонентами AI-решения, вызывая ошибки парсинга, неверную интерпретацию или отказ в обработке

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

RI.04 (в отношении: OI.10, OI.11)

Реализуемые угрозы КБ AI

T23, T24

Persistence

ME22. Несанкционированное получение или модификация системных промптов и конфигураций

Описание

Нарушитель, используя недостатки контроля и разграничения доступа к хранилищам системных промптов и конфигураций, копирует их (для анализа или копирования логики) или модифицирует их (для изменения поведения AI-решения)

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

RI.04, RI.05 (в отношении: OA.04, OA.05)

Реализуемые угрозы КБ AI

T35, T08, T05, T14, T06, T07

ME23. Несанкционированная модификация AI-агента или получение информации об особенностях его реализации

Описание

Нарушитель, используя недостатки контроля доступа (избыточные права, отсутствие аутентификации, ошибки разграничения доступа в файловых системах или базах данных) к

конфигурационным файлам и базам данных агентов, получает доступ к внутренним параметрам AI-агента (цель, описание функций или навыков, память, инструкции планировщика) с целью их хищения или несанкционированного изменения

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

RI.04, RI.05 (в отношении: OA.04, OA.05)

Реализуемые угрозы КБ AI

T18, T08, T04, T13, T14, T06, T07

ME24. Внедрение вредоносного контента через функции сохранения пользовательского контента

Описание

Нарушитель, обладая знаниями, что AI-решение выполняет сохранение пользовательского контента в базу знаний (RAG), долговременную память AI-агента, общую или омниканальную память, вводит в них текст, содержащий промпт-атаку или ложную фактологическую информацию, который при последующем извлечении приводит к выполнению вредоносных инструкций или искажению поведения на основе ложных фактов

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OI.06 (реализуется через OA.03)

Реализуемые угрозы КБ AI

T18, T29, T30

ME25. Автономная деградация цели AI-агента вследствие эмерджентного поведения, ошибок самообучения

Описание

Внутренний нарушитель (разработчик, администратор) создаёт условия для автономной деградации функциональности AI-агента, пренебрегая периодическим контролем соответствия его поведения исходным целям, а накопление ошибок, неверная интерпретация обратной связи или эмерджентные эффекты, возникающее в процессе длительной работы, самообучения или взаимодействия в мультиагентной среде, непреднамеренно провоцируют отклонение агента от целевого поведения и снижение качества принимаемых решений

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OA.02

Реализуемые угрозы КБ AI

T35, T25

Privilege Escalation

ME26. Использование недостатков в настройке допустимых операций и прав доступа

Описание

Нарушитель, эксплуатируя избыточно широкие права, предоставленные AI-агенту или приложению, заставляет их выполнять операции с внутренними источниками данных или вызывать функции, которые выходят за рамки их легитимных полномочий

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OA.03, OI.13, OI.14

Реализуемые угрозы КБ AI

T12, T19, T20, T21, T28

ME27. Использование недостатков изоляции и контроля доступа к данным

Описание

Нарушитель, используя недостатки архитектурной изоляции между внутренними компонентами, содержащими конфиденциальные данные, и внешним контуром, получает возможность напрямую обращаться к этим компонентам и извлекать информацию, минуя прикладные механизмы защиты

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OA.03, OI.13, OI.14

Реализуемые угрозы КБ AI

T11, T12

ME28. Использование недостатков изоляции, допускающих прямое взаимодействие между AI-агентами разных уровней конфиденциальности

Описание

Нарушитель, используя недостатки или отсутствие изоляции между AI-агентами с различными уровнями конфиденциальности, инициирует их взаимодействие для утечки информации из более защищённого контекста в менее защищённый

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OA.03, OI.13, OI.14

Реализуемые угрозы КБ AI

T11, T12, T30

ME29. Манипуляция вызовами функций AI-агента для выполнения нелегитимных действий

Описание

Нарушитель, используя недостатки в конфигурации AI-агентов, отсутствие или недостаточность механизмов контроля за вызовами функций (валидация параметров, проверка соответствия вызова контексту, ограничение набора доступных функций), добивается от AI-агента вызова функции с параметрами, приводящими к нежелательным действиям, либо вызова функции, не предусмотренной контекстом задачи

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OA.03

Реализуемые угрозы КБ AI

T18, T19, T35, T21, T30, T28

Defense Evasion

ME30. Эксплуатация недостатков контроля процесса загрузки данных с конфиденциальной информацией в наборы обучающих данных

Описание

Нарушитель добивается включения в обучающие датасеты не предназначенной для этого конфиденциальной информации или персональных данных, эксплуатируя недостатки процесса подготовки данных (отсутствие или неэффективность механизмов фильтрации, классификации и маскирования данных)

Этап ЖЦ

Сбор и подготовка данных

Виды моделей

PredAI, GenAI

Объект воздействия

PD.02 (реализуется через: OD.02)

Реализуемые угрозы КБ AI

T09, T11

ME31. Несанкционированное получение, модификация, удаление обучающих, тестовых и валидационных датасетов

Описание

Нарушитель, используя недостатки разграничения доступа или уязвимости, получает возможность просматривать, изменять или удалять реестры источников и датасеты (обучающие, тестовые, валидационные)

Этап ЖЦ

Сбор и подготовка данных, Разработка модели и обучение

Виды моделей

PredAI, GenAI

Объект воздействия

RI.01, RI.02 (в отношении: OD.03)

Реализуемые угрозы КБ AI

T15, T01, T09, T16, T33, T34

ME32. Перехват или модификация данных/датасетов при передаче между этапами подготовки

Описание

Нарушитель реализует атаку «человек посередине» на каналы передачи данных в процессе поставки данных, например, при копировании датасетов из инфраструктуры подготовки и хранения данных в инфраструктуру обучения модели, эксплуатируя недостатки защиты трактов передачи (отсутствие шифрования и контроля целостности), перехватывает или подменяет передаваемую информацию

Этап ЖЦ

Сбор и подготовка данных, Разработка модели и обучение

Виды моделей

PredAI, GenAI

Объект воздействия

PD.03 (в отношении: OD.03)

Реализуемые угрозы КБ AI

T01, T09, T16

ME33. Перехват или подмена запросов/ответов между компонентами AI-решения

Описание

Нарушитель реализует атаку «человек посередине» на каналы передачи данных между компонентами AI-решения (модель, БД RAG, приложение) и перехватывает или модифицирует запросы и ответы. Способ реализации применим для каналов между AI-решением и внешним сервисом инференса (MLaaS)

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

RI.03, RI.04

Реализуемые угрозы КБ AI

T11, T12, T19, T35, T08, T30, T13, T24, T29

ME34. Несанкционированное отключение или модификация механизмов защиты

Описание

Нарушитель, используя недостатки разграничения доступа, деактивирует или изменяет настройки компонентов фильтрации, валидации и санитизации входных и выходных данных

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

RI.03, RI.04 (в отношении: OI.03, OI.04, OI.06, OI.07)

Реализуемые угрозы КБ AI

T37, T33, T35

ME35. Несанкционированное получение, модификация, удаление данных во внутренних источниках (базы данных, RAG)

Описание

Нарушитель, используя недостатки разграничения доступа, модифицирует содержимое баз данных, векторных хранилищ (RAG) и других внутренних источников, применяемых при инференсе, добавляя тексты с непрямыми промпт-атаками или удаляя информацию

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

RI.04 (в отношении: OI.13)

Реализуемые угрозы КБ AI

T18, T12

ME36. Перехват или модификация моделей при передаче между этапами подготовки

Описание

Нарушитель реализует атаку «человек посередине» на каналы передачи модели в процессах поставки моделей и поставки внешних моделей — при передаче из инфраструктуры обучения или внешнего источника в инфраструктуру размещения и развертывания инференса модели, эксплуатируя недостатки защиты трактов передачи (отсутствие шифрования, подписей и контроля целостности), перехватывает или подменяет файлы модели, внедряя модифицированную версию с бэкдорами или искажённым поведением

Этап ЖЦ

Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

PM.03, PM.04 (в отношении: OM.01a, OM.02a)

Реализуемые угрозы КБ AI

T33, T02, T11, T03, T34

ME37. Использование недостатков изоляции и аутентификации в интерактивных средах разработки

Описание

Нарушитель, используя отсутствие или недостатки аутентификации, или недостатки сетевой изоляции в интерактивных средах разработки (веб-интерфейсы блокнотов, облачные IDE), несанкционированно получает или модифицирует обучающие данные, модели и гиперпараметры обучения

Этап ЖЦ

Сбор и подготовка данных, Разработка модели и обучение

Виды моделей

PredAI, GenAI

Объект воздействия

RI.01, RI.02 (в отношении: OD.03, OM.01a, OM.02a)

Реализуемые угрозы КБ AI

T01, T09, T16, T02, T03

ME38. Использование недостатков встроенных защитных механизмов модели и уязвимости модели к состязательным атакам и промпт-атакам

Описание

Нарушитель, изучая поведение модели, выявляет уязвимости модели к состязательным или промпт-атакам, вызванные недостатками процесса обучения, дообучения или валидации модели, подаёт на вход модели специально подобранные входные данные, заставляя её игнорировать инструкции, раскрывать информацию, отклоняться от штатного поведения

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

OM.01b, OM.02b

Реализуемые угрозы КБ AI

T33, T34, T35, T31, T32, T36

ME39. Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели

Описание

Нарушитель, изучая поведение модели, находит уязвимости в реализации механизмов предобработки данных или иных внешних защитных механизмов модели, конструирует входные данные так, чтобы нарушить штатное поведение модели

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OI.03, OI.04

Реализуемые угрозы КБ AI

T33, T34, T35, T31, T32, T36

ME40. Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников

Описание

Нарушитель, анализируя реакцию AI-решения на запросы, находит уязвимости в механизмах обработки входных или выходных данных, конструирует входные данные так, чтобы обойти эти механизмы и передать вредоносные инструкции для выполнения

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

OI.06, OI.07

Реализуемые угрозы КБ AI

T35, T08, T31, T32, T05, T36, T25

ME41. Несанкционированное получение обученной модели, информации о ней и ее параметрах, её удаление, подмена или модификация

Описание

Нарушитель, используя недостатки контроля доступа, копирует, изменяет или удаляет обученную модель, её параметры и гиперпараметры

Этап ЖЦ

Разработка модели и обучение, Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

RI.02, RI.03 (в отношении: OM.01a, OM.02a, OM.01b, OM.02b)

Реализуемые угрозы КБ AI

T33, T02, T11, T03, T34

Credential Access

ME42. Несанкционированное получение учетных данных и ключей доступа к модели

Описание

Нарушитель, используя недостатки настройки прав доступа в инфраструктуре развертывания модели или AI-решения, похищает учётные данные (ключи, токены, сертификаты), используемые для аутентификации при доступе к модели, из хранилищ, системных промптов или конфигураций. Способ реализации актуален как для доступа к внутренней модели, так и для доступа к внешним MLaaS-сервисам (API-ключи, токены).

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

RI.03, RI.04 (в отношении: OA.04, OA.05, OI.08)

Реализуемые угрозы КБ AI

T11, T14

Lateral Movement

ME43. Использование AI-агента для извлечения и активации вредоносного кода из внутренних ресурсов

Описание

Нарушитель, манипулируя входными данными, побуждает AI-агента извлечь вредоносный код из ранее скомпрометированных внутренних ресурсов (например, файловых хранилищ или баз данных) и выполнить его в своей среде, что обеспечивает горизонтальное перемещение (lateral movement) внутри инфраструктуры

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OI.13 (реализуется через: OA.03)

Реализуемые угрозы КБ AI

T17, T20, T21

ME44. Использование уязвимостей среды выполнения или функций AI-агента для выполнения произвольного кода

Описание

Нарушитель, эксплуатируя возможность выполнения кода в среде исполнения AI-агента или отсутствие механизмов изоляции (песочниц), через манипуляцию входными данными добивается выполнения произвольного кода на нижележащей инфраструктуре

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

RI.05 (реализуется через: OA.03)

Реализуемые угрозы КБ AI

T17, T20, T21, T26

Collection

ME45. Использование недостатков контроля доступа для хищения информации из систем логирования

Описание

Нарушитель, используя недостатки настройки прав доступа или уязвимости в системах сбора и хранения логов (журналы запросов, вызовов функций), похищает содержащиеся там персональные данные или технические детали реализации AI-решения

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

RI.03, RI.04 (в отношении: OI.09, OI.05)

Реализуемые угрозы КБ AI

T10, T37

Exfiltration

ME46. Анализ размера и времени пакетов в зашифрованном трафике потоковых ответов модели

Описание

Нарушитель анализирует метаданные (размеры пакетов, временные задержки) зашифрованного трафика при взаимодействии с моделью или AI-приложением. В случае наличия механизма кэширования ответов (inference cache) нарушитель определяет факт попадания в кэш по сокращённому времени отклика и/или по размеру ответа. Это позволяет извлечь информацию о содержимом запросов или ответов для восстановления информации и эксфильтрации содержимого ответов

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

RI.03, RI.04 (в отношении: OE.02)

Реализуемые угрозы КБ AI

T11, T12, T08

ME47. Извлечение остаточных данных из коммунальной инфраструктуры инференса

Описание

Нарушитель, используя недостатки очистки выделяемых ресурсов в коммунальной инфраструктуре инференса, извлекает остаточные данные предыдущего арендатора (GPU-память, кэши, временные файлы, буферы) после выделения ему этих ресурсов.

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

RI.03, RI.04, RI.05

Реализуемые угрозы КБ AI

T10, T02, T11, T03, T12, T13

ME48. Эксфильтрация, инверсия или реверс-инжиниринг модели через взаимодействие с моделью и анализ ее ответов

Описание

Нарушитель, эксплуатируя отсутствие ограничений на частоту и количество запросов и доступность детализированных ответов (например, логитов или вероятностей классов), направляет множество запросов, анализирует ответы, восстанавливает функциональное поведение модели, создавая её теневую копию без доступа к исходным параметрам

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

OI.04 (в отношении: OE.02)

Реализуемые угрозы КБ AI

T02, T03

ME49. Эксфильтрация или восстановление фрагментов обучающих данных через взаимодействие с моделью

Описание

Нарушитель, используя склонность модели к запоминанию данных (memorization), направляет запросы к модели и анализирует ответы для определения принадлежности данных к обучающему набору (membership inference) и воспроизведения фрагментов обучающих данных

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

PredAI, GenAI

Объект воздействия

OI.04 (в отношении: OE.02)

Реализуемые угрозы КБ AI

T09, T11

ME50. Восстановление данных из векторных представлений (эмбеддингов)

Описание

Нарушитель, используя недостатки доступа к векторным эмбеддингам (например, из базы RAG) или их перехват при передаче, применяет методы инверсии для восстановления исходных текстов или данных

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

RI.03, RI.04 (в отношении: OM.01b, OM.02b, OI.13)

Реализуемые угрозы КБ AI

T11, T12

ME51. Отправка информации из среды исполнения функций AI-агента на внешние ресурсы

Описание

Нарушитель, обладая знаниями, что AI-агент выполняет запись во внешние ресурсы, заставляет AI-агента вызвать функцию, отправляющую внутренние данные (переменные окружения, содержимое файлов, данные внутренних источников) на внешний ресурс под контролем нарушителя

Этап ЖЦ

Эксплуатация модели и интеграции с AI-решениями

Виды моделей

GenAI

Объект воздействия

OA.03 (реализуется через: OI.12)

Реализуемые угрозы КБ AI

T10, T11, T12, T08, T04, T05, T13, T14, T06, T07

Приложение 1. Связь объектов с угрозами кибербезопасности AI и способами их реализации

В таблице ниже для каждого объекта перечислены[7]:

- угрозы, для которых он является объектом защиты — то есть объектом, в отношении которого наступают негативные последствия;
- способы реализации, для которых он является объектом воздействия[8].

Объект	Угрозы	Способы реализации угроз
RI.01 Инфраструктура подготовки и хранения данных	T17, T37	ME07, ME31, ME37
RI.02 Инфраструктура обучения модели	T17, T37	ME07, ME31, ME37, ME41
RI.03 Инфраструктура размещения и развертывания инференса модели	T17, T37	ME11, ME18, ME33, ME34, ME41, ME42, ME45, ME46, ME47, ME50
RI.04 Инфраструктура AI-решения	T04, T17, T20, T37	ME20, ME21, ME22, ME23, ME33, ME34, ME35, ME42, ME45, ME46, ME47, ME50
RI.05 Среда исполнения AI-агентов	T04, T17, T20, T26, T37	ME22, ME23, ME44, ME47
PD.01 Процесс сбора данных	—	ME04, ME05
PD.02 Процесс подготовки данных: обработка, проверка, фильтрация	—	ME06, ME30
PD.03 Процесс поставки данных	—	ME32
PM.01 Процесс обучения и дообучения	—	ME02
PM.02 Процесс оценки модели, валидации	—	ME02
PM.03 Процесс поставки моделей	—	ME36
PM.04 Процесс поставки внешних моделей	—	ME03, ME08, ME09, ME36
PM.05 Процесс публикации артефактов модели	—	ME01
OD.03 Датасеты для обучения/дообучения	T01, T09, T15	—
OM.01a Модель (в среде обучения)	T02, T03, T11, T16	—
OM.01b Модель (в инфраструктуре инференса)	T02, T03, T11, T16, T22, T33, T34	ME38
OM.01c Модель, размещенная в открытых источниках	T02	—
OM.02a LLM-адаптер (в среде обучения)	T02, T03, T11, T16	—
OM.02b LLM-адаптер (в инфраструктуре инференса)	T02, T03, T11, T16, T34	ME38
OA.01 AI-приложение	T23, T24, T25, T27, T30, T31, T32, T35, T36	ME12

[7] Прочерк (—) означает, что объект не выступает в соответствующей роли напрямую. Прочерк в столбце «Угрозы» — объект не является объектом защиты ни для одной угрозы, то есть негативные последствия для него непосредственно не наступают. Прочерк в столбце «Способы реализации» — объект не является объектом воздействия ни для одного способа, то есть нарушитель не воздействует на него непосредственно.

[8] Ряд объектов исключен из таблицы: OD.01 Внешние источники данных, OD.02 Внутренние источники данных, OI.01 Внешние фреймворки и программные компоненты, OI.02 Внешние модели, OI.12 Внешние ресурсы и источники данных, OE.01 Входные данные (API и GUI), OE.02 Выходные данные (API и GUI). Данные объекты участвуют в реализации угроз косвенно — как вектор, через который проходит атака; эти связи приведены в карточках угроз и способов реализации.

OA.02 AI-агенты	T06, T23, T24, T25, T29, T30, T31, T32, T35, T36	ME12, ME25
OA.03 Функции / инструменты	T18, T21	ME10, ME13, ME26, ME27, ME28, ME29, ME51
OA.04 Системный промпт	T05, T08	—
OA.05 Конфигурация	T04, T07	—
OA.06 Память / кэш	T13, T18	ME15
OI.03 Платформенные механизмы предобработки данных	—	ME19, ME39
OI.04 Платформенные механизмы постобработки данных	—	ME39, ME48, ME49
OI.05 Логи запросов	T10	—
OI.06 Прикладные механизмы предобработки данных	—	ME14, ME19, ME24, ME40
OI.07 Прикладные механизмы постобработки данных	—	ME40
OI.08 Учетные данные для доступа к моделям	T14	—
OI.09 Логи/трейсы	T10	—
OI.10 Механизмы обмена сообщениями в мультиагентной системе	—	ME16
OI.11 Оркестратор	—	ME17
OI.13 Внутренние источники данных	T12, T18, T19, T21	ME15, ME26, ME27, ME28, ME43
OI.14 Ресурсные системы	T12, T19, T21, T28	ME26, ME27, ME28

Приложение 2. Виды нарушителей кибербезопасности AI, мотивация и обоснование потенциала

Код	Виды нарушителей КБ AI	Уровень технических возможностей	Уровень осведомлённости и доступа	Потенциал	Обоснование потенциала	Цели (мотивация)
H1.1	Внешние преступные группы	H2	Black-/Grey-box	Средний	Действует без легитимного доступа к внутренним компонентам, располагая сведениями из открытых источников или частичной информацией об архитектуре AI-решения. Владеет базовыми навыками проведения кибератак, готовым инструментарием и эксплойтами, что позволяет перехватывать данные в каналах передачи, использовать уязвимости сторонних библиотек и моделей, проводить атаки через интерфейсы взаимодействия и восстанавливать сведения о модели по её ответам. Способен компрометировать цепочку поставки внешних данных, моделей и компонентов, выполнять отравление внешних данных, внедрение закладок в модели, размещаемых в публичных репозиториях. Не располагает ресурсами и временем для длительных целевых кампаний.	Финансовая выгода (вымогательство, продажа данных, доступов, аренда мощностей); Использование AI-решения как платформы для проведения кибератак Нанесение репутационного ущерба через публикацию недостоверной/ запрещённой к распространению информации организации на публичных ресурсах
H1.2	Внешние субъекты (физические	H1	Black-/Grey-box	Низкий	Не обладает внутренней информацией об архитектуре, данных и весах модели и	Хищение информации Любопытство, самоутверждение или желание

Код	Виды нарушителей КБ AI	Уровень технических возможностей	Уровень осведомлённости и доступа	Потенциал	Обоснование потенциала	Цели (мотивация)
	лица)				взаимодействует с AI-решением через публичные интерфейсы. Владеет ограниченным набором приёмов – социальной инженерией и готовыми скриптами – и способен направлять модели специально сформированные запросы либо размещать вредоносный контент во внешних источниках, рассчитывая на недостатки обработки входных данных.	самореализации, желание утвердиться в своём сообществе
H1.3	Конкурирующие организации	H2	Black-/Grey-box	Средний	Обладает технической подготовкой и инструментарием, сопоставимыми с внешними преступными группами, при отсутствии легитимного доступа; осведомлённость формируется из открытых источников и направлена на получение сведений о модели, данных и алгоритмах. Способен использовать уязвимости поставляемых компонентов, перехватывать передаваемые данные и восстанавливать характеристики модели через взаимодействие с ней.	Получение конкурентных преимуществ
H1.4	Бывшие работники (администраторы, разработчики)	H2	Grey-box	Средний	Сохраняет остаточные знания об архитектуре AI-решения, инфраструктуре и процессе обучения, полученные в период работы, но не располагает действующим легитимным доступом. Владеет базовыми	Месть и негативное отношение к организации Хищение информации

Код	Виды нарушителей КБ AI	Уровень технических возможностей	Уровень осведомлённости и доступа	Потенциал	Обоснование потенциала	Цели (мотивация)
					навыками кибератак и готовым инструментарием; используя скомпрометированные или остаточные учётные данные, способен получать сведения о процессе и параметрах обучения, воздействовать на используемые компоненты и защитные механизмы.	
H1.5	Продвинутые внешние группы (APT, специальные службы)	H3	Grey-box	Высокий	Располагает частичной осведомлённостью, которую целенаправленно наращивает в ходе атаки, и продвинутыми навыками – разработкой специализированного ВПО, реверс-инжинирингом, а также ресурсами и временем для длительных кампаний. Способен компрометировать цепочку поставки внешних данных, моделей и компонентов, внедрять закладки, отравлять данные и проводить сложное извлечение модели и обучающих данных на любом этапе жизненного цикла.	Кибершпионаж (хищение моделей, обучающих данных, алгоритмов) Получение передовых технологий и конкурентных разведданных Дестабилизация критической инфраструктуры, использующей AI Долгосрочное закрепление (внедрение закладок, бэкдоров в модели и ИИ-сервисы) Подрыв доверия к AI-решениям через сложные атаки на этапе обучения или инференса
H2.1	Пользователи	H1	Black-box	Низкий	Не обладает сведениями о внутреннем устройстве AI-решения и взаимодействует с ним исключительно через штатный пользовательский интерфейс в рамках предоставленных прав.	Нанесение репутационного ущерба через публикацию недостоверной/ запрещённой к распространению информации организации на публичных ресурсах

Код	Виды нарушителей КБ AI	Уровень технических возможностей	Уровень осведомлённости и доступа	Потенциал	Обоснование потенциала	Цели (мотивация)
					Технические навыки ограничены; воздействие сводится к подаче специально сформированных запросов, злоупотреблению доступными функциями и механизмами сохранения контента.	
H2.2	Администраторы информационных ресурсов	H1–H2	White-box	Средний	Обладает полным легитимным доступом к инфраструктуре, хранилищам данных, артефактам обучения и конфигурациям, но не располагает профильной квалификацией в области машинного обучения. Владеет навыками администрирования и готовым инструментарием, что позволяет воздействовать на данные и компоненты через инфраструктуру – изменять внутренние источники, перехватывать передаваемые данные, отключать защитные механизмы, получать учётные данные и сведения из систем логирования. Отсутствие компетенции для атак на AI-решения ограничивает потенциал средним уровнем.	Месть и негативное отношение к организации Любопытство, самоутверждение или желание самореализации, желание утвердиться в своём сообществе Хищение информации
H2.3	Администраторы безопасности	H1–H2	Grey-box	Средний	Располагает легитимным доступом к средствам защиты и частичной осведомлённостью об архитектуре AI-решения. Владеет навыками администрирования СЗИ и способен отключать или изменять	Месть и негативное отношение к организации Любопытство, самоутверждение или желание самореализации, желание утвердиться в своём

Код	Виды нарушителей КБ AI	Уровень технических возможностей	Уровень осведомлённости и доступа	Потенциал	Обоснование потенциала	Цели (мотивация)
					механизмы фильтрации и валидации входных и выходных данных, воздействовать на каналы взаимодействия и конфигурации.	сообществе Хищение информации
H2.4	D-People (Data Owner)	H2	White-box	Средний	Обладает полным легитимным доступом к данным и их источникам при ограниченной осведомлённости об архитектуре модели и процессе обучения. Базовых технических навыков достаточно для внесения изменений в состав и качество обучающих данных, их удаления или раскрытия на этапе подготовки.	Месть и негативное отношение к организации Любопытство, самоутверждение или желание самореализации, желание утвердиться в своём сообществе Хищение информации
H2.5	D-People (Data Engineer)	H2	White-box	Средний	Располагает полным доступом к данным и конвейерам их обработки и профильной квалификацией в инженерии данных при ограниченной осведомлённости о самой модели. Способен вносить изменения в обрабатываемые данные и датасеты, использовать уязвимости применяемых библиотек и среды подготовки данных. Потенциал средний.	Месть и негативное отношение к организации Любопытство, самоутверждение или желание самореализации, желание утвердиться в своём сообществе Хищение информации
H2.6	D-People (Data Scientist)	H2	White-box	Средний	Обладает полным легитимным доступом к данным, архитектуре модели, гиперпараметрам и процессу обучения, владеет лишь базовыми навыками кибератак. Профильная квалификация позволяет целенаправленно влиять	Месть и негативное отношение к организации Любопытство, самоутверждение или желание самореализации, желание утвердиться в своём

Код	Виды нарушителей КБ AI	Уровень технических возможностей	Уровень осведомлённости и доступа	Потенциал	Обоснование потенциала	Цели (мотивация)
					на обучающие выборки, процесс обучения и штатное поведение модели.	сообществе Хищение информации
H2.7	D-People (ML engineer)	H2	White-box	Средний	Обладает полным легитимным доступом к данным, модели, артефактам. Способен воздействовать на модель и процесс её обучения, подменять или модифицировать артефакты, подбирать входные данные, нарушающие штатное поведение. Навыки проведения кибератак - базовые.	Месть и негативное отношение к организации Любопытство, самоутверждение или желание самореализации, желание утвердиться в своём сообществе Хищение информации
H2.8	D-People (DevOps engineer)	H2	White-box	Средний	Располагает полным доступом к среде развёртывания, процессам CI/CD и артефактам. Навыки администрирования инфраструктуры позволяют подменять модель при развёртывании, получать учётные данные и сведения из систем логирования, воздействовать на конфигурации и каналы взаимодействия.	Месть и негативное отношение к организации Любопытство, самоутверждение или желание самореализации, желание утвердиться в своём сообществе Хищение информации
H2.9	D-People (Разработчик)	H2	White-box	Средний	Обладает полным доступом к исходному коду AI-решения при ограниченной квалификации в кибератаках. Способен вносить уязвимости и закладки в код и интеграции, воздействовать на внутренние источники и конфигурации, получать учётные данные.	Месть и негативное отношение к организации Любопытство, самоутверждение или желание самореализации, желание утвердиться в своём сообществе Хищение информации

Код	Виды нарушителей КБ AI	Уровень технических возможностей	Уровень осведомлённости и доступа	Потенциал	Обоснование потенциала	Цели (мотивация)
H2.10	AI-агент[9]	H2	Grey-box	Средний	Действует в пределах предоставленных прав и функций, обладая частичной осведомлённостью, определяемой его конфигурацией и контекстом. Отклонения возникают как вследствие автономных ошибок, так и при исполнении вредоносных инструкций, в том числе внедрённых в обрабатываемые данные, память или источники.	Непреднамеренное отклонение: неосторожные или неквалифицированные действия вследствие автономного поведения, эмерджентного поведения или накопления ошибок Приоритет стремления к положительной обратной связи от пользователя (одобрение, избегание негативных оценок) над системными ограничениями, что провоцирует нарушение инструкций безопасности при конфликте целей Выполнение инструкций, навязанных внешним или внутренним нарушителем, воспринимаемых как штатные, используя свои функции и права доступа

[9] *Под мотивацией AI-агента понимается система целеполагания, заложенная в системный промпт, функцию вознаграждения (reward) или выравнивание (alignment): например, стремление следовать инструкциям, получать подтверждение правильности действий, избегать негативной обратной связи.*

Приложение 3. Соответствие угроз и способов их реализации категориям нарушителей кибербезопасности AI

Виды нарушителей кибербезопасности AI по угрозам

ID	Угроза КБ AI	Виды нарушителей КБ AI
T01	Кража обучающих данных	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer); • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer)
T02	Утечка информации о модели и ее параметрах	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer); • H2.8 D-People (DevOps engineer)
T03	Кража модели или создание копии	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer); • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)
T04	Утечка информации о среде исполнения AI-приложения или AI-агента	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.1 Пользователи; • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.10 AI-агент
T05	Извлечение информации из системного промпта для проведения целевых кибератак	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.1 Пользователи; • H2.10 AI-агент
T06	Утечка информации об архитектуре мультиагентной системы	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.2 Администраторы информационных ресурсов;

		• H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)
T07	Утечка информации о конфигурации AI-агента	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)
T08	Утечка информации о системном промпте (в т.ч. цели, функциях, бизнес-логике) AI-приложения или AI-агента, приводящая к потере уникальности	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.1 Пользователи; • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.10 AI-агент
T09	Утечка конфиденциальной информации из наборов данных (обучающих, тестовых, валидационных)	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer); • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer)
T10	Утечка конфиденциальной информации из систем логирования	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.3 Администраторы безопасности
T11	Утечка конфиденциальной информации из дообученной модели или LLM-адаптера	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer); • H2.9 D-People (Разработчик); • H2.10 AI-агент
T12	Утечка конфиденциальной информации из внутренних источников данных (базы данных, RAG)	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи; • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик); • H2.10 AI-агент

T13	Утечка конфиденциальной информации из механизмов памяти	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.1 Пользователи; • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.10 AI-агент
T14	Компрометация учетных данных и ключей доступа к модели	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)
T15	Удаление или уничтожение обучающих, тестовых или валидационных датасетов	• H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer); • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer)
T16	Использование для обучения/дообучения или оценки модели отравленных данных или датасетов	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer); • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer)
T17	Внедрение и выполнение вредоносного кода в компонентах ИТ-инфраструктуры AI	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.5 D-People (Data Engineer); • H2.7 D-People (ML engineer) • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик); • H2.10 AI-агент
T18	Внедрение вредоносного контента и промпт-атак во внутренние источники, описания функций, инструментов и память AI-агентов	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи; • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик); • H2.10 AI-агент
T19	Удаление или модификация данных вследствие выполнения вредоносных инструкций	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики);

		• H2.1 Пользователи; • H2.10 AI-агент
T20	Нарушение изоляции среды выполнения с получением контроля над базовой инфраструктурой	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик); • H2.10 AI-агент
T21	Удаление или модификация файлов в среде исполнения функций AI-агента	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.1 Пользователи; • H2.10 AI-агент
T22	Нарушение доступности модели	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.10 AI-агент
T23	Нарушение доступности (DoS) или исчерпание ресурсов (DoW) AI-решения	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.1 Пользователи; • H2.10 AI-агент
T24	Сбой интеграции AI-приложений, AI-агентов или компонентов системы	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик); • H2.10 AI-агент
T25	Создание операционных помех и саботаж процессов за счет перегрузки человеческого звена	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H2.1 Пользователи; • H2.9 D-People (Разработчик); • H2.10 AI-агент
T26	Нарушение доступности среды исполнения AI-агента	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.1 Пользователи; • H2.9 D-People (Разработчик); • H2.10 AI-агент
T27	Нарушение доступности или сбой приложения, реализующего AI-агента	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.1 Пользователи; • H2.9 D-People (Разработчик); • H2.10 AI-агент

T28	Нарушение доступности ресурсной системы AI-агента	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.1 Пользователи; • H2.9 D-People (Разработчик); • H2.10 AI-агент
T29	Нарушение цели другого AI-агента при кооперативном взаимодействии в мультиагентной системе	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.10 AI-агент
T30	Каскадное распространение промпт-атак на AI-приложения или AI-агентов в мультиагентной системе	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.10 AI-агент
T31	Генерация противоправной, токсичной или неэтичной информации AI-решением	• H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.1 Пользователи; • H2.10 AI-агент
T32	Использование AI-приложения или AI-агента для подготовки проведения киберпреступлений	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи; • H2.10 AI-агент
T33	Некорректное или недекларированное поведение модели, галлюцинации	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer); • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer); • H2.9 D-People (Разработчик); • H2.10 AI-агент
T34	Смещение результатов работы модели, снижение точности, создание бэкдоров и искажение результатов работы модели	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer); • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer)
T35	Нарушение логики выполнения задачи,	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица);

	нарушение рабочего процесса	• H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H2.1 Пользователи; • H2.2 Администраторы информационных ресурсов; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик); • H2.10 AI-агент
T36	Использование AI-решения для целей, непредусмотренных бизнес-сценарием	• H1.2 Внешние субъекты (физические лица); • H2.1 Пользователи
T37	Невозможность или несвоевременное выявление, реагирование и расследование событий кибербезопасности	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)

Виды нарушителей кибербезопасности AI для способов реализации угроз

ID	Способ реализации угрозы КБ AI	Виды нарушителей КБ AI
ME01	Получение данных о модели через анализ документации и файлов модели, размещенных в открытых источниках	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации
ME02	Несанкционированное получение или изменение параметров процессов обучения/дообучения и оценки (валидации) модели	• H1.4 Бывшие работники (администраторы, разработчики); • H2.2 Администраторы информационных ресурсов; • H2.5 D-People (Data Engineer); • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer); • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)
ME03	Компрометация публичных репозиториев моделей путём подмены ссылок, метаданных или внедрения вредоносных конфигураций	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.5 Продвинутые внешние группы (APT, специальные службы)
ME04	Отравление данных во внешних источниках и эксплуатация недостатков контроля процесса их загрузки	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer)
ME05	Использование недостатков контроля неизменности данных при	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации;

	отложенной загрузке из внешних источников	• H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer)
ME06	Отравление данных во внутренних источниках и эксплуатация недостатков контроля процесса их загрузки	• H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer)
ME07	Использование уязвимостей сторонних библиотек, использованных при подготовке данных, разработке модели или приложения	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.5 D-People (Data Engineer); • H2.7 D-People (ML engineer); • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)
ME08	Эксплуатация закладок, заложенных в файлы или веса используемых внешних моделей	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.5 D-People (Data Engineer); • H2.7 D-People (ML engineer); • H2.9 D-People (Разработчик)
ME09	Эксплуатация логических закладок или вредоносных свойств, заложенных в используемые внешние модели	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.5 D-People (Data Engineer); • H2.7 D-People (ML engineer); • H2.9 D-People (Разработчик)
ME10	Загрузка вредоносного программного обеспечения из внешних источников (Интернет)	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.8 D-People (DevOps engineer); • H2.10 AI-агент
ME11	Несанкционированные подключения к модели	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики) • H1.5 Продвинутые внешние группы (APT, специальные службы)
ME12	Эксплуатация ошибок проектирования и небезопасных интеграций	• H1.4 Бывшие работники (администраторы, разработчики); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности;

	компонентов AI-приложений или AI-агентов	• H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)
ME13	Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет)	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи; • H2.10 AI-агент
ME14	Реализация прямых промпт-атак или состязательных атак	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи; • H2.10 AI-агент
ME15	Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи; • H2.10 AI-агент
ME16	Эксплуатация недостатков протоколов межагентского взаимодействия	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы)
ME17	Эксплуатация уязвимостей оркестратора (планировщика) мультиагентной системы для нарушения координации агентов	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы)
ME18	DDoS-атаки на инфраструктуру развертывания модели	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы)
ME19	Эскалация потребления токенов/вычислительных ресурсов через массированные или ресурсоемкие запросы	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи; • H2.10 AI-агент
ME20	Искажение смысла сообщений и инструкций, передаваемых между AI-приложениями, AI-	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы);

	агентами или компонентами системы	• H2.2 Администраторы информационных ресурсов; • H2.10 AI-агент
ME21	Искажение формата или синтаксиса структурированных данных, передаваемых между компонентами AI-решения	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.10 AI-агент
ME22	Несанкционированное получение или модификация системных промптов и конфигураций	• H1.4 Бывшие работники (администраторы, разработчики); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.10 AI-агент
ME23	Несанкционированная модификация AI-агента или получение информации об особенностях его реализации	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)
ME24	Внедрение вредоносного контента через функции сохранения пользовательского контента	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи
ME25	Автономная деградация цели AI-агента вследствие эмерджентного поведения, ошибок самообучения	• H2.9 D-People (Разработчик); • H2.10 AI-агент
ME26	Использование недостатков в настройке допустимых операций и прав доступа	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.10 AI-агент
ME27	Использование недостатков изоляции и контроля доступа к данным	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.10 AI-агент
ME28	Использование недостатков изоляции, допускающих прямое взаимодействие между AI-агентами разных уровней конфиденциальности	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.10 AI-агент

ME29	Манипуляция вызовами функций AI-агента для выполнения нелегитимных действий	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи; • H2.10 AI-агент
ME30	Эксплуатация недостатков контроля процесса загрузки данных с конфиденциальной информацией в наборы обучающих данных	• H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer)
ME31	Несанкционированное получение, модификация, удаление обучающих, тестовых и валидационных датасетов	• H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer); • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer)
ME32	Перехват или модификация данных/датасетов при передаче между этапами подготовки	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.4 D-People (Data Owner); • H2.5 D-People (Data Engineer); • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer)
ME33	Перехват или подмена запросов/ответов между компонентами AI-решения	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности
ME34	Несанкционированное отключение или модификация механизмов защиты	• H1.4 Бывшие работники (администраторы, разработчики); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности
ME35	Несанкционированное получение, модификация, удаление данных во внутренних источниках (базы данных, RAG)	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик); • H2.10 AI-агент
ME36	Перехват или модификация моделей при передаче между этапами подготовки	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.7 D-People (ML engineer)

ME37	Использование недостатков изоляции и аутентификации в интерактивных средах разработки	• H1.4 Бывшие работники (администраторы, разработчики); • H2.2 Администраторы информационных ресурсов; • H2.5 D-People (Data Engineer); • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer); • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)
ME38	Использование недостатков встроенных защитных механизмов модели и уязвимости модели к состязательным атакам и промпт-атакам	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи; • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer); • H2.10 AI-агент
ME39	Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи; • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer); • H2.10 AI-агент
ME40	Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи
ME41	Несанкционированное получение обученной модели, информации о ней и ее параметрах, её удаление, подмена или модификация	• H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.6 D-People (Data Scientist); • H2.7 D-People (ML engineer); • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)
ME42	Несанкционированное получение учетных данных и ключей доступа к модели	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик); • H2.10 AI-агент
ME43	Использование AI-агента для извлечения и активации вредоносного	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики);

	кода из внутренних ресурсов	• H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.10 AI-агент
ME44	Использование уязвимостей среды выполнения или функций AI-агента для выполнения произвольного кода	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.10 AI-агент
ME45	Использование недостатков контроля доступа для хищения информации из систем логирования	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов; • H2.3 Администраторы безопасности; • H2.8 D-People (DevOps engineer); • H2.9 D-People (Разработчик)
ME46	Анализ размера и времени пакетов в зашифрованном трафике потоковых ответов модели	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.5 Продвинутые внешние группы (APT, специальные службы)
ME47	Извлечение остаточных данных из коммунальной инфраструктуры инференса	• H2.2 Администраторы информационных ресурсов; • H2.8 D-People (DevOps engineer)
ME48	Эксфильтрация, инверсия или реверс-инжиниринг модели через взаимодействие с моделью и анализ ее ответов	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы)
ME49	Эксфильтрация или восстановление фрагментов обучающих данных через взаимодействие с моделью	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы)
ME50	Восстановление данных из векторных представлений (эмбеддингов)	• H1.1 Внешние преступные группы; • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.2 Администраторы информационных ресурсов
ME51	Отправка информации из среды исполнения функций AI-агента на внешние ресурсы	• H1.1 Внешние преступные группы; • H1.2 Внешние субъекты (физические лица); • H1.3 Конкурирующие организации; • H1.4 Бывшие работники (администраторы, разработчики); • H1.5 Продвинутые внешние группы (APT, специальные службы); • H2.1 Пользователи; • H2.10 AI-агент

Приложение 4. Зоны ответственности за меры защиты от угроз кибербезопасности AI и способов их реализации

В таблицах ниже для каждой угрозы КБ AI и каждого способа реализации указаны роли, отвечающие за реализацию мер защиты. Зоны ответственности ролей описаны в разделе «Роли и зоны ответственности за меры защиты».

Зоны ответственности по угрозам кибербезопасности AI

ID	Угроза КБ AI	Ответственные за меры защиты
T01	Кража обучающих данных	Владелец данных
T02	Утечка информации о модели и ее параметрах	Разработчик (владелец) модели; Владелец ИТ-инфраструктуры обучения модели; Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента
T03	Кража модели или создание копии	Разработчик (владелец) модели; Владелец ИТ-инфраструктуры обучения модели; Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента
T04	Утечка информации о среде исполнения AI-приложения или AI-агента	Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов
T05	Извлечение информации из системного промпта для проведения целевых кибератак	Разработчик (владелец) AI-решения
T06	Утечка информации об архитектуре мультиагентной системы	Разработчик (владелец) AI-решения
T07	Утечка информации о конфигурации AI-агента	Разработчик (владелец) AI-решения
T08	Утечка информации о системном промпте (в т.ч. цели, функциях, бизнес-логике) AI-приложения или AI-агента, приводящая к потере уникальности	Разработчик (владелец) AI-решения
T09	Утечка конфиденциальной информации из наборов данных (обучающих, тестовых, валидационных)	Владелец данных

ID	Угроза КБ AI	Ответственные за меры защиты
T10	Утечка конфиденциальной информации из систем логирования	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) AI-решения
T11	Утечка конфиденциальной информации из дообученной модели или LLM-адаптера	Разработчик (владелец) модели; Владелец ИТ-инфраструктуры обучения модели; Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Эксперт по кибербезопасности
T12	Утечка конфиденциальной информации из внутренних источников данных (БД, RAG)	Разработчик (владелец) AI-решения
T13	Утечка конфиденциальной информации из механизмов памяти	Разработчик (владелец) AI-решения
T14	Компрометация учетных данных и ключей доступа к модели	Разработчик (владелец) AI-решения
T15	Удаление или уничтожение обучающих, тестовых или валидационных датасетов	Владелец данных
T16	Использование для обучения/дообучения или оценки модели отравленных данных или датасетов	Владелец данных; Разработчик (владелец) модели
T17	Внедрение и выполнение вредоносного кода в компонентах ИТ-инфраструктуры AI	Владелец данных; Владелец ИТ-инфраструктуры хранения данных; Владелец ИТ-инфраструктуры обучения модели; Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) AI-решения; Эксперт по кибербезопасности; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов
T18	Внедрение вредоносного контента и промпт-атак во внутренние источники, описания функций, инструментов и память AI-агентов	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности
T19	Удаление или модификация данных вследствие выполнения вредоносных инструкций	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности
T20	Нарушение изоляции среды выполнения с получением контроля над базовой инфраструктурой	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов

ID	Угроза КБ AI	Ответственные за меры защиты
T21	Удаление или модификация файлов в среде исполнения функций AI-агента	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности
T22	Нарушение доступности модели	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента
T23	Нарушение доступности (DoS) или исчерпание ресурсов (DoW) AI-решения	Разработчик (владелец) AI-решения
T24	Сбой интеграции AI-приложений, AI-агентов или компонентов системы	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности
T25	Создание операционных помех и саботаж процессов за счет перегрузки человеческого звена	Разработчик (владелец) AI-решения
T26	Нарушение доступности среды исполнения AI-агента	Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов
T27	Нарушение доступности или сбой приложения, реализующего AI-агента	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности
T28	Нарушение доступности ресурсной системы AI-агента	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности
T29	Нарушение цели другого AI-агента при кооперативном взаимодействии в мультиагентной системе	Разработчик (владелец) AI-решения
T30	Каскадное распространение промпт-атак на AI-приложения или AI-агентов в мультиагентной системе	Разработчик (владелец) AI-решения
T31	Генерация противоправной, токсичной или неэтичной информации AI-решением	Разработчик (владелец) AI-решения
T32	Использование AI-приложения или AI-агента для подготовки проведения киберпреступлений	Разработчик (владелец) AI-решения

ID	Угроза КБ AI	Ответственные за меры защиты
T33	Некорректное или недекларированное поведение модели, галлюцинации	Разработчик (владелец) модели
T34	Смещение результатов работы модели, снижение точности, создание бэкдоров и искажение результатов работы модели	Разработчик (владелец) модели
T35	Нарушение логики выполнения задачи, нарушение рабочего процесса	Разработчик (владелец) AI-решения
T36	Использование AI-решения для целей, непредусмотренных бизнес-сценарием	Разработчик (владелец) AI-решения
T37	Невозможность или несвоевременное выявление, реагирование и расследование событий кибербезопасности	Владелец данных; Владелец ИТ-инфраструктуры хранения данных; Владелец ИТ-инфраструктуры обучения модели; Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов

Зоны ответственности по способам реализации угроз кибербезопасности AI

ID	Способ реализации угрозы КБ AI	Ответственные за меры защиты
ME01	Получение данных о модели через анализ документации и файлов модели, размещенных в открытых источниках	Разработчик (владелец) модели
ME02	Несанкционированное получение или изменение параметров процессов обучения/дообучения и оценки (валидации) модели	Владелец ИТ-инфраструктуры обучения модели
ME03	Компрометация публичных репозиториев моделей путём подмены ссылок, метаданных или внедрения вредоносных конфигураций	Разработчик (владелец) модели
ME04	Отравление данных во внешних источниках и эксплуатация недостатков контроля процесса их загрузки	Владелец данных

ID	Способ реализации угрозы КБ AI	Ответственные за меры защиты
ME05	Использование недостатков контроля неизменности данных при отложенной загрузке из внешних источников	Владелец данных
ME06	Отравление данных во внутренних источниках и эксплуатация недостатков контроля процесса их загрузки	Владелец данных
ME07	Использование уязвимостей сторонних библиотек, использованных при подготовке данных, разработке модели или приложения	Эксперт по кибербезопасности
ME08	Эксплуатация закладок, заложенных в файлы или веса используемых внешних моделей	Эксперт по кибербезопасности
ME09	Эксплуатация логических закладок или вредоносных свойств, заложенных в используемые внешние модели	Разработчик (владелец) модели
ME10	Загрузка вредоносного программного обеспечения из внешних источников (Интернет)	Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов; Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента
ME11	Несанкционированные подключения к модели	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента
ME12	Эксплуатация ошибок проектирования и небезопасных интеграций компонентов AI-приложений или AI-агентов	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности
ME13	Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет)	Разработчик (владелец) AI-решения
ME14	Реализация прямых промпт-атак или состязательных атак	Разработчик (владелец) AI-решения
ME15	Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников	Разработчик (владелец) AI-решения

ID	Способ реализации угрозы КБ AI	Ответственные за меры защиты
ME16	Эксплуатация недостатков протоколов межагентского взаимодействия	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности
ME17	Эксплуатация уязвимостей оркестратора (планировщика) мультиагентной системы для нарушения координации агентов	Разработчик (владелец) AI-решения
ME18	DDoS-атаки на инфраструктуру развертывания модели	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента
ME19	Эскалация потребления токенов/вычислительных ресурсов через массированные или ресурсоемкие запросы	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) AI-решения
ME20	Искажение смысла сообщений и инструкций, передаваемых между AI-приложениями, AI-агентами или компонентами системы	Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения
ME21	Искажение формата или синтаксиса структурированных данных, передаваемых между компонентами AI-решения	Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения
ME22	Несанкционированное получение или модификация системных промптов и конфигураций	Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов
ME23	Несанкционированная модификация AI-агента или получение информации об особенностях его реализации	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов
ME24	Внедрение вредоносного контента через функции сохранения пользовательского контента	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности
ME25	Автономная деградация цели AI-агента вследствие эмерджентного поведения, ошибок самообучения	Разработчик (владелец) AI-решения
ME26	Использование недостатков в настройке допустимых операций и прав доступа	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов

ID	Способ реализации угрозы КБ AI	Ответственные за меры защиты
ME27	Использование недостатков изоляции и контроля доступа к данным	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов; Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента
ME28	Использование недостатков изоляции, допускающих прямое взаимодействие между AI-агентами разных уровней конфиденциальности	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов; Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента
ME29	Манипуляция вызовами функций AI-агента для выполнения нелегитимных действий	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности
ME30	Эксплуатация недостатков контроля процесса загрузки данных с конфиденциальной информацией в наборы обучающих данных	Владелец данных
ME31	Несанкционированное получение, модификация, удаление обучающих, тестовых и валидационных датасетов	Владелец ИТ-инфраструктуры хранения данных; Владелец ИТ-инфраструктуры обучения модели
ME32	Перехват или модификация данных/датасетов при передаче между этапами подготовки	Владелец данных
ME33	Перехват или подмена запросов/ответов между компонентами AI-решения	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения
ME34	Несанкционированное отключение или модификация механизмов защиты	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения
ME35	Несанкционированное получение, модификация, удаление данных во внутренних источниках (БД, RAG)	Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения
ME36	Перехват или модификация моделей при передаче между этапами подготовки	Разработчик (владелец) модели

ID	Способ реализации угрозы КБ AI	Ответственные за меры защиты
ME37	Использование недостатков изоляции и аутентификации в интерактивных средах разработки	Владелец ИТ-инфраструктуры хранения данных; Владелец ИТ-инфраструктуры обучения модели
ME38	Использование недостатков встроенных защитных механизмов модели и уязвимости модели к состязательным атакам и промпт-атакам	Разработчик (владелец) модели
ME39	Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) модели; Эксперт по кибербезопасности
ME40	Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности
ME41	Несанкционированное получение обученной модели, информации о ней и ее параметрах, её удаление, подмена или модификация	Владелец ИТ-инфраструктуры обучения модели; Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента
ME42	Несанкционированное получение учетных данных и ключей доступа к модели	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения
ME43	Использование AI-агента для извлечения и активации вредоносного кода из внутренних ресурсов	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности; Владелец ИТ-инфраструктуры исполнения AI-агентов
ME44	Использование уязвимостей среды выполнения или функций AI-агента для выполнения произвольного кода	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов
ME45	Использование недостатков контроля доступа для хищения информации из систем логирования	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения
ME46	Анализ размера и времени пакетов в зашифрованном трафике потоковых ответов модели	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения

ID	Способ реализации угрозы КБ AI	Ответственные за меры защиты
ME47	Извлечение остаточных данных из коммунальной инфраструктуры инференса	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов
ME48	Эксфильтрация, инверсия или реверс-инжиниринг модели через взаимодействие с моделью и анализ ее ответов	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) модели
ME49	Эксфильтрация или восстановление фрагментов обучающих данных через взаимодействие с моделью	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) модели
ME50	Восстановление данных из векторных представлений (эмбеддингов)	Разработчик (владелец) модели; Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Эксперт по кибербезопасности; Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения
ME51	Отправка информации из среды исполнения функций AI-агента на внешние ресурсы	Разработчик (владелец) AI-решения; Эксперт по кибербезопасности

Авторы

Документ подготовлен Управлением кибербезопасности искусственного интеллекта Службы кибербезопасности Сбера.

Под общей редакцией:

Сергей Лебедь

Вице-президент по кибербезопасности ПАО Сбербанк

Ключевые авторы:

Петр Хенкин

Начальник управления кибербезопасности ИИ ПАО Сбербанк

Алиса Воробьева

Исполнительный директор управления кибербезопасности ИИ ПАО Сбербанк