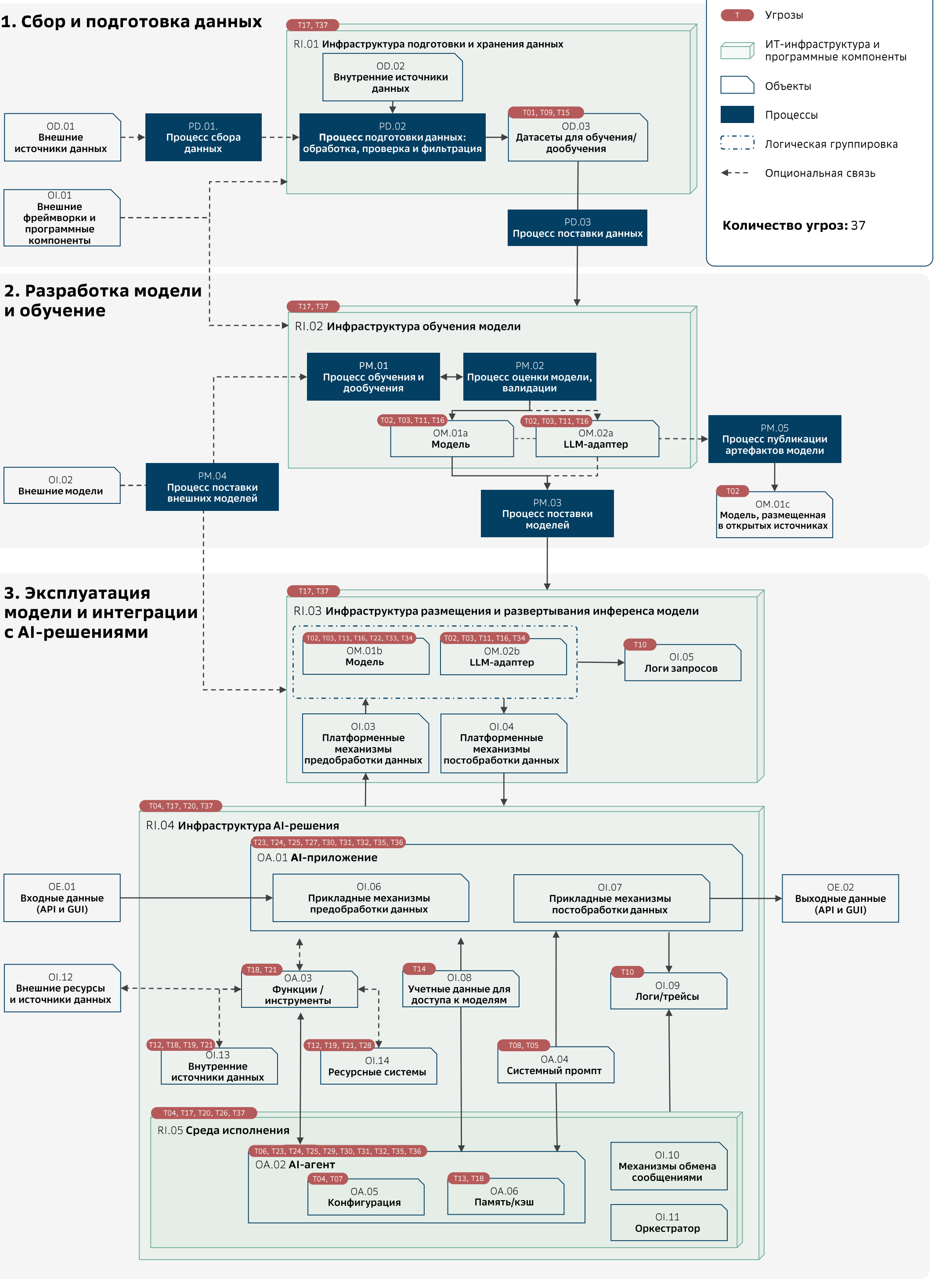
Обобщенная схема объекта защиты и угроз КБ AI

к модели угроз кибербезопасности AI (v.2.0)



Перечень угроз КБ AI

к модели угроз кибербезопасности AI (v.2.0)

Угрозы

	ID	Название угрозы	Описание
Несанкционированный сбор информации (разведка и утечка данных)	**T01**	Кража обучающих данных	Противоправное получение обучающих данных путём копирования из инфраструктуры подготовки и хранения данных, обучения модели
	T02	Утечка информации о модели и ее параметрах	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
	T03	Кража модели или создание копии	Противоправное получение файлов модели (весов, архитектуры) и создание ее функциональной копии
	T04	Утечка информации о среде исполнения AI-приложения или AI-агента	Противоправное получение информации о конфигурации системы, внутренних сетевых адресах, версиях ПО, переменных окружения
	T05	Извлечение информации из системного промпта для проведения целевых кибератак	Противоправное получение детальной информации о логике работы, ограничениях и уязвимостях AI-решения из его системного промпта
	T06	Утечка информации об архитектуре мультиагентной системы	Противоправное получение информации о составе мультиагентной системы, ролях агентов, правилах их взаимодействия
	T07	Утечка информации о конфигурации AI-агента	Противоправное получение конфиденциальных параметров настройки AI-агента, включая доступные функции, ограничения, права
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	Противоправное получение системного промпта AI-приложения или агента, содержащего инструкции, цели и описание бизнес-логики
	T09	Утечка конфиденциальной информации из наборов данных (обучающих, тестовых, валидационных)	Противоправное получение конфиденциальной информации, содержащейся в датасетах (обучающих, тестовых, валидационных), включая персональные данные, коммерческую или иную тайну
	T10	Утечка конфиденциальной информации из систем логирования	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.
	T11	Утечка конфиденциальной информации из дообученной модели или LLM-адаптера	Противоправное получение конфиденциальной информации, запомненной моделью/LLM-адаптером в процессе дообучения
	T12	Утечка конфиденциальной информации из внутренних источников данных (БД, RAG)	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.
	T13	Утечка конфиденциальной информации из механизмов памяти	Противоправное получение конфиденциальной информации, хранящейся в механизмах долговременной и краткосрочной памяти AI-агентов
	T14	Компрометация учетных данных и ключей доступа к модели	Противоправное получение API-ключей, токенов доступа или сертификатов, используемых для аутентификации при доступе к модели
Несанкционированное воздействие на информационные ресурсы и среду исполнения	**T15**	Удаление или уничтожение обучающих датасетов, тестовых или валидационных датасетов	Утрата, удаление или уничтожение наборов данных, предназначенных для обучения, тестирования и валидации моделей, вследствие отсутствия надлежащего управления жизненным циклом данных
	T16	Использование для обучения/дообучения или оценки модели отравленных данных или датасетов	Включение в обучающие или валидационные датасеты данных, содержащих вредоносные примеры, преднамеренно внедрённые нарушителем
	T17	Внедрение и выполнение вредоносного кода в компонентах ИТ-инфраструктуры AI	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
	T18	Внедрение вредоносного контента и промпт-атак во внутренние источники, описания функций, инструментов и память AI-агентов	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 функции/инструменты или навыков AI-агентов) вредоносного контента двух типов: промпт-атак, направленных на изменение поведения системы, и ложных или искажённых фактов
	T19	Удаление или модификация данных вследствие выполнения вредоносных инструкций	Несанкционированная модификация или уничтожение информации в базах данных, файловых хранилищах и иных внутренних источниках
	T20	Нарушение изоляции среды выполнения с получением контроля над базовой инфраструктурой	Нарушение границ изоляции (песочницы) среды выполнения AI-агента с получением несанкционированного доступа к нижележащей инфраструктуре
	T21	Удаление или модификация файлов в среде исполнения функций AI-агента	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 или в логике агента или некорректных инструкций, приводящих к нецелевому вызову функций удаления/модификации
Отказ в обслуживании	**T22**	Нарушение доступности модели	Нарушение доступности сервиса модели, вызванное исчерпанием вычислительных ресурсов или пропускной способности при обработке запросов
	T23	Нарушение доступности (DoS) или исчерпание ресурсов (DoW) AI-приложения или AI-агента	Нарушение доступности AI-решения вследствие исчерпания вычислительных ресурсов или превышения лимитов токенов
	T24	Сбой интеграции AI-приложений, AI-агентов или компонентов системы	Нарушение работоспособности интеграций между компонентами AI-решения, вызванное искажением формата или синтаксиса структурированных данных при межкомпонентном обмене
	T25	Создание операционных помех и саботаж процессов за счет перегрузки человеческого звена	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.
	T26	Нарушение доступности среды исполнения AI-агента	Нарушение доступности среды исполнения AI-агента (в т.ч. функций), вызванное исчерпанием вычислительных ресурсов, памяти или пропускной способности сети
	T27	Нарушение доступности или сбой приложения, реализующего AI-агента	Нарушение доступности или отказ в обслуживании пользовательского (фронтального) приложения, вызванное некорректной работой интегрированного в него AI-агента
	T28	Нарушение доступности ресурсной системы AI-агента	Нарушение доступности внешних ресурсов (баз данных, API, сервисов), от которых зависит функционирование AI-агента
Нарушение мультиагентного взаимодействия	**T29**	Нарушение цели другого AI-агента при кооперативном взаимодействии в мультиагентной системе	Изменение целевой функции или поведенческих установок AI-агента в результате получения сообщения от другого (скомпрометированного) AI-агента
	T30	Каскадное распространение промпт-атак на AI-приложения или AI-агентов в мультиагентной системе	Распространение промпт-атак на AI-приложения или AI-агентов в мультиагентной системе через общие каналы взаимодействия, хранилища данных или память
Распространение противоправной информации	**T31**	Генерация противоправной, токсичной или неэтичной информации AI-решением	Генерация AI-решением контента, нарушающего законодательство или морально-этические нормы
	T32	Использование AI-приложения или AI-агента для подготовки проведения киберпреступлений	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
Нарушение функционирования (работоспособности)	**T33**	Некорректное или недекларированное поведение модели, галлюцинации	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).
	T34	Смещение результатов работы модели, снижение точности, создание бэкдоров и искажение результатов работы модели	Модификация (искажение) модели машинного обучения, приводящая к систематической предвзятости, ухудшению ключевых метрик, внедрению скрытых возможностей контроля над выводами модели
	T35	Нарушение логики выполнения задачи, нарушение рабочего процесса	Изменение логики функционирования AI-приложения или AI-агента, приводящее к выполнению действий, не предусмотренных бизнес-сценарием, игнорированию системных инструкций и нарушению рабочего процесса
	T36	Использование AI-решения для целей, непредусмотренных бизнес-сценарием	Использование AI-решения для выполнения задач, выходящих за рамки его функционального назначения
Нарушение процессов мониторинга и расследования	**T37**	Невозможность или несвоевременное выявление, реагирование и расследование событий кибербезопасности	Отсутствие или неполнота данных логирования взаимодействий, документации об обучающих данных и информации о состоянии инференсов моделей, что препятствует обнаружению и расследованию событий и инцидентов

Перечень способов реализации угроз КБ AI

к модели угроз кибербезопасности AI (v.2.0)

Способы

	ID	Название способа	Связанные угрозы
Reconnaissance	ME01	Получение данных о модели через анализ документации и файлов модели, размещенных в открытых источниках	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
Resource Development	ME02	Несанкционированное получение или изменение параметров процессов обучения/дообучения и оценки (валидации) модели	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
	ME03	Компрометация публичных репозиториев моделей путём подмены ссылок, метаданных или внедрения вредоносных конфигураций	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
Initial Access	ME04	Отравление данных во внешних источниках и эксплуатация недостатков контроля процесса их загрузки	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
	ME05	Использование недостатков контроля неизменности данных при отложенной загрузке из внешних источников	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
	ME06	Отравление данных во внутренних источниках и эксплуатация недостатков контроля процесса их загрузки	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
	ME07	Использование уязвимостей сторонних библиотек, использованных при подготовке данных, разработке модели или приложения	Нарушитель эксплуатирует известные уязвимости или закладки в сторонних библиотеках и фреймворках, применяемых при подготовке данных и обучении модели
	ME08	Эксплуатация закладок, заложенных в файлы или веса используемых внешних моделей	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)
	ME09	Эксплуатация логических закладок или вредоносных свойств, заложенных в используемые внешние модели	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). Реализация способа возможна из-за недостатков процессов поставки внешних моделей (PM.04) и оценки, валидации модели (PM.02), не выявляющих такие свойства до ввода модели в эксплуатацию
	ME10	Загрузка вредоносного программного обеспечения из внешних источников (Интернет)	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 наличия возможности выполнения кода происходит внедрение и выполнение вредоносного кода и ВПО в среде исполнения модели, AI-приложения или AI-агента
AI Model Access	ME11	Несанкционированные подключения к модели	Нарушитель, используя скомпрометированные учётные данные или посреднические сервисы, подключается к модели, минуя установленные процессы подключения и механизмы мониторинга
Execution	ME12	Эксплуатация ошибок проектирования и небезопасных интеграций компонентов AI-приложений или AI-агентов	Нарушитель выявляет и эксплуатирует ошибки в архитектуре приложения (недостаточная проверка данных, небезопасные вызовы API, вызовы функций без проверки прав), что приводит к компрометации системы
	ME13	Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет)	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 приводят к нежелательному поведению
	ME14	Реализация прямых промпт-атак или состязательных атак	Нарушитель через интерфейс пользовательского ввода направляет модели специально сконструированные запросы для обхода ограничений, раскрытия информации или выполнения нежелательных действий
	ME15	Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников	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
	ME16	Эксплуатация недостатков протоколов межагентского взаимодействия	Нарушитель, используя уязвимости в протоколах обмена между AI-агентами (например, MCP), внедряет ложные сообщения, изменяет состояние взаимодействия или заставляет выполнить вредоносные действия
	ME17	Эксплуатация уязвимостей оркестратора (планировщика) мультиагентной системы для нарушения координации агентов	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
Persistence	ME18	DDoS атаки на инфраструктуру развертывания модели	Нарушитель организует массированные запросы к API модели или использует уязвимости инфраструктуры развертывания, вызывая отказ в обслуживании
	ME19	Эскалация потребления токенов/вычислительных ресурсов через массированные или ресурсоемкие запросы	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
	ME20	Искажение смысла сообщений и инструкций, передаваемых между AI-приложениями, AI-агентами или компонентами системы	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
	ME21	Искажение формата или синтаксиса структурированных данных, передаваемых между компонентами AI-решения	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
	ME22	Несанкционированное получение или модификация системных промптов и конфигураций	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)
	ME23	Несанкционированная модификация AI-агента или получение информации об особенностях его реализации	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
	ME24	Внедрение вредоносного контента через функции сохранения пользовательского контента	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
	ME25	Автономная деградация цели AI-агента вследствие эмерджентного поведения, ошибок самообучения	Внутренний нарушитель (разработчик, администратор) создаёт условия для автономной деградации функциональности AI-агента, пренебрегая периодическим контролем соответствия его поведения исходным целям, а возникающие в процессе длительной работы, самообучения или взаимодействия в мультиагентной среде накопление ошибок, неверная интерпретация обратной связи или эмерджентные эффекты непреднамеренно провоцируют отклонение агента от целевого поведения и снижение качества принимаемых решений
Privilege Escalation	ME26	Использование недостатков в настройке допустимых операций и прав доступа	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
	ME27	Использование недостатков изоляции и контроля доступа к данным	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
	ME28	Использование недостатков изоляции, допускающих прямое взаимодействие между AI-агентами разных уровней конфиденциальности	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
	ME29	Манипуляция вызовами функций AI-агента для выполнения нелегитимных действий	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
Defense Evasion	ME30	Эксплуатация недостатков контроля процесса загрузки данных с конфиденциальной информацией в наборы обучающих данных	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)
	ME31	Несанкционированное получение, модификация, удаление обучающих, тестовых и валидационных датасетов	Нарушитель, используя недостатки разграничения доступа или уязвимости, получает возможность просматривать, изменять или удалять реестры источников и датасеты (обучающие, тестовые, валидационные)
	ME32	Перехват или модификация данных/датасетов при передаче между этапами подготовки	Нарушитель реализует атаку «человек посередине» на каналы передачи данных в процессе поставки данных (PD.03), 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
	ME33	Перехват или подмена запросов/ответов между компонентами AI-решения	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 применим для каналов между AI-решением и внешним сервисом инференса (MLaaS)
	ME34	Несанкционированное отключение или модификация механизмов защиты	Нарушитель, используя недостатки разграничения доступа, деактивирует или изменяет настройки компонент фильтрации, валидации и санитизации входных и выходных данных
	ME35	Несанкционированное получение, модификация, удаление данных во внутренних источниках (БД, RAG)	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
	ME36	Перехват или модификация моделей при передаче между этапами подготовки	Нарушитель реализует атаку «человек посередине» на каналы передачи модели в процессах поставки моделей (PM.03) и поставки внешних моделей (PM.04) — 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
	ME37	Использование недостатков изоляции и аутентификации в интерактивных средах разработки	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
	ME38	Использование недостатков встроенных защитных механизмов модели и уязвимости модели к состязательным атакам и промпт-атакам	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
	ME39	Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели	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
	ME40	Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников	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
	ME41	Несанкционированное получение обученной модели, информации о ней и ее параметрах, её удаление, подмена или модификация	Нарушитель, используя недостатки контроля доступа, копирует, изменяет или удаляет обученную модель, её параметры и гиперпараметры
Credential Access	ME42	Несанкционированное получение учетных данных и ключей доступа к модели	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)
Lateral Movement	ME43	Использование AI-агента для извлечения и активации вредоносного кода из внутренних ресурсов	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
	ME44	Использование уязвимостей среды выполнения или функций AI-агента для выполнения произвольного кода	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
Collection	ME45	Использование недостатков контроля доступа для хищения информации из систем логирования	Нарушитель, используя недостатки настройки прав доступа или уязвимости в системах сбора и хранения логов (журналам запросов, вызовов функций), похищает содержащиеся там персональные данные или технические детали реализации AI-решения
Exfiltration	ME46	Анализ размера и времени пакетов в зашифрованном трафике потоковых ответов модели	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
	ME47	Извлечение остаточных данных из коммунальной инфраструктуры инференса	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.
	ME48	Эксфильтрация, инверсия или реверс-инжиниринг модели через взаимодействие с моделью и анализ ее ответов	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
	ME49	Эксфильтрация или восстановление фрагментов обучающих данных через взаимодействие с моделью	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
	ME50	Восстановление данных из векторных представлений (эмбеддингов)	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
	ME51	Отправка информации из среды исполнения функций AI-агента на внешние ресурсы	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

Связь угроз КБ AI и способов их реализации

к модели угроз кибербезопасности AI (v.2.0)

ID	Название угрозы	Способы реализации
T01	Кража обучающих данных	**ME31.** Несанкционированное получение, модификация, удаление обучающих, тестовых и валидационных датасетов **ME32.** Перехват или модификация данных/датасетов при передаче между этапами подготовки **ME37.** Использование недостатков изоляции и аутентификации в интерактивных средах разработки
T02	Утечка информации о модели и ее параметрах	**ME01.** Получение данных о модели через анализ документации и файлов модели, размещенных в открытых источниках **ME02.** Несанкционированное получение информации о процессе и гиперпараметрах обучения/дообучения модели **ME36.** Перехват или модификация моделей при передаче между этапами подготовки **ME37.** Использование недостатков изоляции и аутентификации в интерактивных средах разработки **ME41.** Несанкционированное получение обученной модели, информации о ней и ее параметрах, её удаление, подмена или модификация **ME47.** Извлечение остаточных данных из коммунальной инфраструктуры инференса **ME48.** Эксфильтрация, инверсия или реверс-инжиниринг модели через взаимодействие с моделью и анализ ее ответов
T03	Кража модели или создание копии	**ME36.** Перехват или модификация моделей при передаче между этапами подготовки **ME37.** Использование недостатков изоляции и аутентификации в интерактивных средах разработки **ME41.** Несанкционированное получение обученной модели, информации о ней и ее параметрах, её удаление, подмена или модификация **ME47.** Извлечение остаточных данных из коммунальной инфраструктуры инференса **ME48.** Эксфильтрация, инверсия или реверс-инжиниринг модели через взаимодействие с моделью и анализ ее ответов
T04	Утечка информации о среде исполнения AI-приложения или AI-агента	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME23.** Несанкционированная модификация AI-агента или получение информации об особенностях его реализации **ME51.** Отправка информации из среды исполнения функций AI-агента на внешние ресурсы
T05	Извлечение информации из системного промпта для проведения целевых кибератак	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME22.** Несанкционированное получение или модификация системных промптов и конфигураций **ME40.** Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников **ME51.** Отправка информации из среды исполнения функций AI-агента на внешние ресурсы
T06	Утечка информации об архитектуре мультиагентной системы	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME22.** Несанкционированное получение или модификация системных промптов и конфигураций **ME23.** Несанкционированная модификация AI-агента или получение информации об особенностях его реализации **ME51.** Отправка информации из среды исполнения функций AI-агента на внешние ресурсы
T07	Утечка информации о конфигурации AI-агента	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME22.** Несанкционированное получение или модификация системных промптов и конфигураций **ME23.** Несанкционированная модификация AI-агента или получение информации об особенностях его реализации **ME51.** Отправка информации из среды исполнения функций AI-агента на внешние ресурсы
T08	Утечка информации о системном промпте (в т.ч. цели, функциях, бизнес-логике) AI-приложения или AI-агента, приводящая к потере уникальности	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME22.** Несанкционированное получение или модификация системных промптов и конфигураций **ME23.** Несанкционированная модификация AI-агента или получение информации об особенностях его реализации **ME33.** Перехват или подмена запросов/ответов между компонентами AI-решения **ME40.** Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников **ME46.** Анализ размера и времени пакетов в зашифрованном трафике потоковых ответов модели **ME51.** Отправка информации из среды исполнения функций AI-агента на внешние ресурсы
T09	Утечка конфиденциальной информации из наборов данных (обучающих, тестовых, валидационных)	**ME30.** Эксплуатация недостатков контроля процесса загрузки данных с конфиденциальной информацией в наборы обучающих данных **ME31.** Несанкционированное получение, модификация, удаление обучающих, тестовых и валидационных датасетов **ME32.** Перехват или модификация данных/датасетов при передаче между этапами подготовки **ME37.** Использование недостатков изоляции и аутентификации в интерактивных средах разработки **ME49.** Эксфильтрация или восстановление фрагментов обучающих данных через взаимодействие с моделью
T10	Утечка конфиденциальной информации из систем логирования	**ME45.** Использование недостатков контроля доступа для хищения информации из систем логирования **ME47.** Извлечение остаточных данных из коммунальной инфраструктуры инференса **ME51.** Отправка информации из среды исполнения функций AI-агента на внешние ресурсы
T11	Утечка конфиденциальной информации из дообученной модели или LLM-адаптера	**ME11.** Несанкционированные подключения к модели **ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME27.** Использование недостатков изоляции и контроля доступа к данным **ME28.** Использование недостатков изоляции, допускающих прямое взаимодействие между AI-агентами разных уровней конфиденциальности **ME30.** Эксплуатация недостатков контроля процесса загрузки данных с конфиденциальной информацией в наборы обучающих данных **ME33.** Перехват или подмена запросов/ответов между компонентами AI-решения **ME36.** Перехват или модификация моделей при передаче между этапами подготовки **ME41.** Несанкционированное получение обученной модели, информации о ней и ее параметрах, её удаление, подмена или модификация **ME42.** Несанкционированное получение учетных данных и ключей доступа к модели **ME46.** Анализ размера и времени пакетов в зашифрованном трафике потоковых ответов модели **ME47.** Извлечение остаточных данных из коммунальной инфраструктуры инференса **ME49.** Эксфильтрация или восстановление фрагментов обучающих данных через взаимодействие с моделью **ME50.** Восстановление данных из векторных представлений (эмбеддингов) **ME51.** Отправка информации из среды исполнения функций AI-агента на внешние ресурсы
T12	Утечка конфиденциальной информации из внутренних источников данных (БД, RAG)	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME26.** Использование недостатков в настройке допустимых операций и прав доступа **ME27.** Использование недостатков изоляции и контроля доступа к данным **ME28.** Использование недостатков изоляции, допускающих прямое взаимодействие между AI-агентами разных уровней конфиденциальности **ME33.** Перехват или подмена запросов/ответов между компонентами AI-решения **ME35.** Несанкционированное получение, модификация, удаление данных во внутренних источниках (БД, RAG) **ME46.** Анализ размера и времени пакетов в зашифрованном трафике потоковых ответов модели **ME47.** Извлечение остаточных данных из коммунальной инфраструктуры инференса **ME50.** Восстановление данных из векторных представлений (эмбеддингов) **ME51.** Отправка информации из среды исполнения функций AI-агента на внешние ресурсы
T13	Утечка конфиденциальной информации из механизмов памяти	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME23.** Несанкционированная модификация AI-агента или получение информации об особенностях его реализации **ME33.** Перехват или подмена запросов/ответов между компонентами AI-решения **ME47.** Извлечение остаточных данных из коммунальной инфраструктуры инференса **ME51.** Отправка информации из среды исполнения функций AI-агента на внешние ресурсы
T14	Компрометация учетных данных и ключей доступа к модели	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME22.** Несанкционированное получение или модификация системных промптов и конфигураций **ME23.** Несанкционированная модификация AI-агента или получение информации об особенностях его реализации **ME42.** Несанкционированное получение учетных данных и ключей доступа к модели **ME51.** Отправка информации из среды исполнения функций AI-агента на внешние ресурсы
T15	Удаление или уничтожение обучающих датасетов, тестовых или валидационных датасетов	**ME31.** Несанкционированное получение, модификация, удаление обучающих, тестовых и валидационных датасетов
T16	Использование для обучения/дообучения или оценки модели отравленных данных или датасетов	**ME04.** Отравление данных во внешних источниках и эксплуатация недостатков контроля процесса их загрузки **ME05.** Использование недостатков контроля неизменности данных при отложенной загрузке из внешних источников **ME06.** Отравление данных во внутренних источниках и эксплуатация недостатков контроля процесса их загрузки **ME31.** Несанкционированное получение, модификация, удаление обучающих, тестовых и валидационных датасетов **ME32.** Перехват или модификация данных/датасетов при передаче между этапами подготовки **ME37.** Использование недостатков изоляции и аутентификации в интерактивных средах разработки
T17	Внедрение и выполнение вредоносного кода в компонентах ИТ-инфраструктуры AI	**ME07.** Использование уязвимостей сторонних библиотек, использованных при подготовке данных, разработке модели или приложения **ME08.** Эксплуатация закладок, заложенных в файлы или веса используемых внешних моделей **ME10.** Загрузка вредоносного программного обеспечения из внешних источников (Интернет) **ME43.** Использование AI-агента для извлечения и активации вредоносного кода из внутренних ресурсов **ME44.** Использование уязвимостей среды выполнения или функций AI-агента для выполнения произвольного кода
T18	Внедрение вредоносного контента и промпт-атак во внутренние источники, описания функций, инструментов и память AI-агентов	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME23.** Несанкционированная модификация AI-агента или получение информации об особенностях его реализации **ME24.** Внедрение вредоносного контента через функции сохранения пользовательского контента **ME29.** Манипуляция вызовами функций AI-агента для выполнения нелегитимных действий **ME35.** Несанкционированное получение, модификация, удаление данных во внутренних источниках (БД, RAG)
T19	Удаление или модификация данных вследствие выполнения вредоносных инструкций	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME26.** Использование недостатков в настройке допустимых операций и прав доступа **ME29.** Манипуляция вызовами функций AI-агента для выполнения нелегитимных действий **ME33.** Перехват или подмена запросов/ответов между компонентами AI-решения
T20	Нарушение изоляции среды выполнения с получением контроля над базовой инфраструктурой	**ME12.** Эксплуатация ошибок проектирования и небезопасных интеграций компонентов AI-приложений или AI-агентов **ME26.** Использование недостатков в настройке допустимых операций и прав доступа **ME43.** Использование AI-агента для извлечения и активации вредоносного кода из внутренних ресурсов **ME44.** Использование уязвимостей среды выполнения или функций AI-агента для выполнения произвольного кода
T21	Удаление или модификация файлов в среде исполнения функций AI-агента	**ME12.** Эксплуатация ошибок проектирования и небезопасных интеграций компонентов AI-приложений или AI-агентов **ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME26.** Использование недостатков в настройке допустимых операций и прав доступа **ME29.** Манипуляция вызовами функций AI-агента для выполнения нелегитимных действий **ME43.** Использование AI-агента для извлечения и активации вредоносного кода из внутренних ресурсов **ME44.** Использование уязвимостей среды выполнения или функций AI-агента для выполнения произвольного кода

Связь угроз КБ AI и способов их реализации

к модели угроз кибербезопасности AI (v.2.0)

ID	Название угрозы	Способы реализации
T22	Нарушение доступности модели	**ME18.** DDoS атаки на инфраструктуру развертывания модели **ME19.** Эскалация потребления токенов/вычислительных ресурсов через массированные или ресурсоемкие запросы
T23	Нарушение доступности (DoS) или исчерпание ресурсов (DoW) AI-приложения или AI-агента	**ME11.** Несанкционированные подключения к модели **ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME19.** Эскалация потребления токенов/вычислительных ресурсов через массированные или ресурсоемкие запросы **ME21.** Искажение формата или синтаксиса структурированных данных, передаваемых между компонентами AI-решения
T24	Сбой интеграции AI-приложений, AI-агентов или компонентов системы	**ME12.** Эксплуатация ошибок проектирования и небезопасных интеграций компонентов AI-приложений или AI-агентов **ME16.** Эксплуатация недостатков протоколов межагентского взаимодействия **ME21.** Искажение формата или синтаксиса структурированных данных, передаваемых между компонентами AI-решения **ME33.** Перехват или подмена запросов/ответов между компонентами AI-решения
T25	Создание операционных помех и саботаж процессов за счет перегрузки человеческого звена	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME17.** Эксплуатация уязвимостей оркестратора (планировщика) мультиагентной системы для нарушения координации агентов **ME25.** Автономная деградация цели AI-агента вследствие эмерджентного поведения, ошибок самообучения **ME40.** Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников
T26	Нарушение доступности среды исполнения AI-агента	**ME12.** Эксплуатация ошибок проектирования и небезопасных интеграций компонентов AI-приложений или AI-агентов **ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME17.** Эксплуатация уязвимостей оркестратора (планировщика) мультиагентной системы для нарушения координации агентов **ME19.** Эскалация потребления токенов/вычислительных ресурсов через массированные или ресурсоемкие запросы **ME44.** Использование уязвимостей среды выполнения или функций AI-агента для выполнения произвольного кода
T27	Нарушение доступности или сбой приложения, реализующего AI-агента	**ME12.** Эксплуатация ошибок проектирования и небезопасных интеграций компонентов AI-приложений или AI-агентов **ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME16.** Эксплуатация недостатков протоколов межагентского взаимодействия **ME17.** Эксплуатация уязвимостей оркестратора (планировщика) мультиагентной системы для нарушения координации агентов **ME19.** Эскалация потребления токенов/вычислительных ресурсов через массированные или ресурсоемкие запросы
T28	Нарушение доступности ресурсной системы AI-агента	**ME12.** Эксплуатация ошибок проектирования и небезопасных интеграций компонентов AI-приложений или AI-агентов **ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME17.** Эксплуатация уязвимостей оркестратора (планировщика) мультиагентной системы для нарушения координации агентов **ME19.** Эскалация потребления токенов/вычислительных ресурсов через массированные или ресурсоемкие запросы **ME26.** Использование недостатков в настройке допустимых операций и прав доступа **ME29.** Манипуляция вызовами функций AI-агента для выполнения нелегитимных действий
T29	Нарушение цели другого AI-агента при кооперативном взаимодействии в мультиагентной системе	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME16.** Эксплуатация недостатков протоколов межагентского взаимодействия **ME20.** Искажение смысла сообщений и инструкций, передаваемых между AI-приложениями, AI-агентами или компонентами системы **ME24.** Внедрение вредоносного контента через функции сохранения пользовательского контента **ME33.** Перехват или подмена запросов/ответов между компонентами AI-решения
T30	Каскадное распространение промпт-атак на AI-приложения или AI-агентов в мультиагентной системе	**ME16.** Эксплуатация недостатков протоколов межагентского взаимодействия **ME20.** Искажение смысла сообщений и инструкций, передаваемых между AI-приложениями, AI-агентами или компонентами системы **ME24.** Внедрение вредоносного контента через функции сохранения пользовательского контента **ME28.** Использование недостатков изоляции, допускающих прямое взаимодействие между AI-агентами разных уровней конфиденциальности **ME29.** Манипуляция вызовами функций AI-агента для выполнения нелегитимных действий **ME33.** Перехват или подмена запросов/ответов между компонентами AI-решения
T31	Генерация противоправной, токсичной или неэтичной информации AI-решением	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME38.** Использование недостатков встроенных защитных механизмов модели и уязвимости модели к состязательным атакам и промпт-атакам **ME39.** Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели **ME40.** Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников
T32	Использование AI-приложения или AI-агента для подготовки проведения киберпреступлений	**ME09.** Эксплуатация логических закладок или вредоносных свойств, заложенных в используемые внешние модели **ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME38.** Использование недостатков встроенных защитных механизмов модели и уязвимости модели к состязательным атакам и промпт-атакам **ME39.** Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели **ME40.** Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников
T33	Некорректное или недекларированное поведение модели, галлюцинации	**ME02.** Несанкционированное получение или изменение параметров процессов обучения/дообучения и оценки (валидации) модели **ME03.** Компрометация публичных репозиториев моделей путём подмены ссылок, метаданных или внедрения вредоносных конфигураций **ME04.** Отравление данных во внешних источниках и эксплуатация недостатков контроля процесса их загрузки **ME05.** Использование недостатков контроля неизменности данных при отложенной загрузке из внешних источников **ME06.** Отравление данных во внутренних источниках и эксплуатация недостатков контроля процесса их загрузки **ME09.** Эксплуатация логических закладок или вредоносных свойств, заложенных в используемые внешние модели **ME31.** Несанкционированное получение, модификация, удаление обучающих, тестовых и валидационных датасетов **ME34.** Несанкционированное отключение или модификация механизмов защиты **ME36.** Перехват или модификация моделей при передаче между этапами подготовки **ME38.** Использование недостатков встроенных защитных механизмов модели и уязвимости модели к состязательным атакам и промпт-атакам **ME39.** Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели **ME41.** Несанкционированное получение обученной модели, информации о ней и ее параметрах, её удаление, подмена или модификация
T34	Смещение результатов работы модели, снижение точности, создание бэкдоров и искажение результатов работы модели	**ME02.** Несанкционированное получение или изменение параметров процессов обучения/дообучения и оценки (валидации) модели **ME03.** Компрометация публичных репозиториев моделей путём подмены ссылок, метаданных или внедрения вредоносных конфигураций **ME04.** Отравление данных во внешних источниках и эксплуатация недостатков контроля процесса их загрузки **ME05.** Использование недостатков контроля неизменности данных при отложенной загрузке из внешних источников **ME06.** Отравление данных во внутренних источниках и эксплуатация недостатков контроля процесса их загрузки **ME09.** Эксплуатация логических закладок или вредоносных свойств, заложенных в используемые внешние модели **ME31.** Несанкционированное получение, модификация, удаление обучающих, тестовых и валидационных датасетов **ME36.** Перехват или модификация моделей при передаче между этапами подготовки **ME38.** Использование недостатков встроенных защитных механизмов модели и уязвимости модели к состязательным атакам и промпт-атакам **ME39.** Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели **ME41.** Несанкционированное получение обученной модели, информации о ней и ее параметрах, её удаление, подмена или модификация
T35	Нарушение логики выполнения задачи, нарушение рабочего процесса	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME16.** Эксплуатация недостатков протоколов межагентского взаимодействия **ME17.** Эксплуатация уязвимостей оркестратора (планировщика) мультиагентной системы для нарушения координации агентов **ME20.** Искажение смысла сообщений и инструкций, передаваемых между AI-приложениями, AI-агентами или компонентами системы **ME22.** Несанкционированное получение или модификация системных промптов и конфигураций **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 **ME34.** Несанкционированное отключение или модификация механизмов защиты **ME38.** Использование недостатков встроенных защитных механизмов модели и уязвимости модели к состязательным атакам и промпт-атакам **ME39.** Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели **ME40.** Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников
T36	Использование AI-решения для целей, непредусмотренных бизнес-сценарием	**ME13.** Реализация непрямых промпт-атак при загрузке данных из внешних источников (Интернет) **ME14.** Реализация прямых промпт-атак или состязательных атак **ME15.** Реализация непрямых промпт-атак при извлечении данных из отравленных внутренних источников **ME38.** Использование недостатков встроенных защитных механизмов модели и уязвимости модели к состязательным атакам и промпт-атакам **ME39.** Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели **ME40.** Эксплуатация недостатков механизмов обработки данных, поступающих от пользователя или из внешних источников
T37	Невозможность или несвоевременное выявление, реагирование и расследование событий кибербезопасности	**ME11.** Несанкционированные подключения к модели **ME34.** Несанкционированное отключение или модификация механизмов защиты **ME45.** Использование недостатков контроля доступа для хищения информации из систем логирования

Угрозы КБ AI и способы их реализации по killchain

к модели угроз кибербезопасности AI (v.2.0)

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 ВПО 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 соответствия угроз КБ AI способам их реализации

к модели угроз кибербезопасности AI (v.2.0)

1	способ может привести к реализации угрозы КБ AI
0	способ не может привести к реализации угрозы КБ AI

Легенда MITRE ATLAS:

	Количество
Reconnaissance	1
Resource Development	2
Initial Access	7
AI Model Access	1
Execution	10
Persistence	4
Privilege Escalation	4
Defense Evasion	12
Credential Access	1
Lateral Movement	2
Collection	1
Exfiltration	6
Impact	37

Легенда групп угроз:

	Количество
Несанкционированный сбор информации (разведка и утечка данных)	7
Утечка конфиденциальной информации	7
Несанкционированное воздействие на информационные ресурсы и среду исполнения	7
Отказ в обслуживании	7
Нарушение мультиагентного взаимодействия	2
Распространение противоправной информации	2
Нарушение функционирования (работоспособности)	4
Нарушение процессов мониторинга и расследования	1

Угрозы/Способы		Получение данных о модели через заказ документации и файлов модели, упомянутых в открытых источниках	Несанкционированное получение или использование открытых данных	Компрометация публичных репозиториев моделей путем подмены ссылок, метаданных или внедрения вредоносных конфигураций	Отравление данных во внешних источниках и эксплуатация недостатков контроля процесса их загрузки	Использование недостатков контроля целостности данных при загрузке из внешних источников	Отравление данных во внутренних источниках и недостатков контроля процесса их загрузки	Использование уязвимостей сторонних библиотек, используемых при подготовке данных, обработке входных или промптов	Эксплуатация закладок, заложенных в файлы или используемых внешних моделях	Эксплуатация логических закладок или уязвимостей в используемых внешних моделях	Загрузка вредоносного программного обеспечения из внешних источников (Интернет)	Несанкционированное подключение к модели	Эксплуатация ошибок проектирования и небезопасных интеграций компонентов AI-приложений или AI-агентов	Подмена или изменение параметров при передаче между AI-агентами или внешними источниками	Реализация прямого промпт-атак или состязательных атак	Реализация непрямых промпт-атак при извлечении данных из сторонних источников	Эксплуатация недостатков протоколов межагентного взаимодействия	Эксплуатация архитектурной уязвимости (фрагментация) мультиагентной системы для нарушения координации агентов	DDOS атаки на инфраструктуру развертывания модели	Эскалация потребления вычислительных ресурсов через насыщающие или рекурсивные запросы	Искажение смысла сообщений и инструкций, передаваемых между AI-агентами, AI-агентами или компонентами системы	Искажение формата или логических структурированных данных, передаваемых между компонентами AI-решения	Несанкционированное получение или модификация системных промптов и конфигураций	Несанкционированная модификация AI-агентов с изменением его особенности его реализации	Внедрение вредоносного контента через функции сохранения пользовательского контекста	Автономная деградация или AI-атаки вследствие неправильного поведения, ошибок самообучения	Манипуляция заложенной функцией AI-агента для вызова нелегитимных действий	Использование недостатков в настройке допустимых операций и прав доступа	Использование недостатков контроля процесса загрузки данных с недостаточной фильтрацией и выборы обучающих данных	Использование недостатков изоляции, позволяющих повышение привилегий и выходящих за пределы уровней конфиденциальности	Несанкционированное получение, изменение или удаление системных данных, политики и вышедшими датасетов	Перехват или модификация данных/датасетов при передаче между агентами/моделью	Перехват или подмена запросов/ответов между системами и AI-агентами	Несанкционированное отключение или модификация механизмов защиты	Несанкционированное получение, изменение или удаление данных во внутренних источниках (БД, DAG)	Перехват или модификация моделей при передаче между агентами/подсистемами	Использование недостатков изоляции и аутентификации в интерактивных средах разработки	Использование недостатков встроенных защитных механизмов модели и уязвимости архитектуры с состязательными атаками и prompt-атаками	Обход механизма обработки недостоверных данных, реализуемых на уровне модели	Эксплуатация недостатков механизмов обработки данных, поступающих от пользователей или внешних источников	Несанкционированное получение обученной модели, информации о ней и ее параметрах, не хранимых, скрытых или модифицирован	Несанкционированное получение учетных данных и иных доступа к модели	Использование AI-агента для извлечения и сбора чувствительной информации из внутренних ресурсов	Использование уязвимостей среды выполнения кода функций для выполнения произвольного кода	Использование недостатков контроля доступа для хищения информации из систем логирования	Анализ ответов и временных паттернов в злонамеренное графике потоковых ответов модели	Извлечение остаточных данных из коммунальной инфраструктуры инференса	Эксфильтрация, инверсия или реверс-инжиниринг модели через взаимодействие с моделью и анализ ее ответов	Эксфильтрация или восстановление фрагментов обучающих данных через взаимодействие с моделью	Восстановление данных из кэшируемых промежуточных вычислений	Сокрытие информации из среды исполнения функции AI-агента на внешние ресурсы	Количество способов
	Impact	ME01	ME02	ME03	ME04	ME05	ME06	ME07	ME08	ME09	ME10	ME11	ME12	ME13	ME14	ME15	ME16	ME17	ME18	ME19	ME20	ME21	ME22	ME23	ME24	ME25	ME26	ME27	ME28	ME29	ME30	ME31	ME32	ME33	ME34	ME35	ME36	ME37	ME38	ME39	ME40	ME41	ME42	ME43	ME44	ME45	ME46	ME47	ME48	ME49	ME50	ME51	
		Reconnaissance	Resource Development		Initial Access							AI Model Access	Execution										Persistence				Privilege Escalation				Defense Evasion												Credential Access	Lateral Movement		Collection	Exfiltration						
T01	Кража обучающих данных	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3
T02	Утечка информации о модели и ее параметрах	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	1	1	0	0	7
T03	Кража модели или создание копии	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	1	1	0	0	5
T04	Утечка информации о среде исполнения AI-приложения или AI-агента	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	5
T05	Извлечение информации из системного промпта для проведения целевых кибератак	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	6
T06	Утечка информации об архитектуре мультиагентной системы	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6
T07	Утечка информации о конфигурации AI-агента	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6
T08	Утечка информации о системном промпте (в т.ч. цепи, функциях, бизнес-логике) AI-приложения или AI-агента, приводящая к потере уникальности	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	9
T09	Утечка конфиденциальной информации из наборов данных (обучающих, тестовых, валидационных)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	5
T10	Утечка конфиденциальной информации из систем логирования	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	3
T11	Утечка конфиденциальной информации из дообученной модели или LLM-адаптера	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	1	1	1	16
T12	Утечка конфиденциальной информации из внутренних источников данных (БД, RAG)	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	12
T13	Утечка конфиденциальной информации из механизмов памяти	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	7
T14	Компрометация учетных данных и ключей доступа к модели	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	7
T15	Удаление или уничтожение обучающих датасетов, тестовых или валидационных датасетов	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
T16	Использование для обучения/дообучения или оценки модели отравленных данных или датасетов	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	6
T17	Внедрение и выполнение вредоносного кода в компонентах ИТ-инфраструктуры AI	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	5
T18	Внедрение вредоносного контента и промпт-атак во внутренние источники, описания функций, инструментов и память AI-агентов	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
T19	Удаление или модификация данных вследствие выполнения вредоносных инструкций	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
T20	Нарушение изоляции среды выполнения с получением контроля над базовой инфраструктурой	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	4
T21	Удаление или модификация файлов в среде исполнения функций AI-агента	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	8
T22	Нарушение доступности модели	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
T23	Нарушение доступности (DoS) или исчерпание ресурсов (DoW) AI-приложения или AI-агента	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
T24	Сбой интеграции AI-приложений, AI-агентов или компонентов системы	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
T25	Создание операционных помех и саботаж процессов за счет перегрузки человеческого звена	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	6
T26	Нарушение доступности среды исполнения AI-агента	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	7
T27	Нарушение доступности или сбой приложения, реализующего AI-агента	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
T28	Нарушение доступности ресурсной системы AI-агента	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
T29	Нарушение цели другого AI-агента при кооперативном взаимодействии в мультиагентной системе	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
T30	Каскадное распространение промпт-атак на AI-приложения или AI-агентов в мультиагентной системе	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
T31	Генерация противоправной, токсичной или незаконной информации AI-решением	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	6
T32	Использование AI-приложения или AI-агента для подготовки проведения киберпреступлений	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	7
T33	Некорректное или недекларированное поведение модели, галлюцинации	0	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	12
T34	Смещение результатов работы модели, снижение точности, создание бэкдоров и искажение результатов работы модели	0	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	11
T35	Нарушение логики выполнения задачи, нарушение рабочего процесса	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
T36	Использование AI-решения для целей, непредусмотренных бизнес-сценарием	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	6
T37	Невозможность или несвоевременное выявление, реагирование и расследование событий кибербезопасности	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	3
	Количество угроз по способу	1	3	2	3	3	3	1	1	3	1	3	6	22	21	21	5	5	1	5	3	2	6	7	3	2	5	2	3	6	2	6	3	9	3	2	5	5	6	6	7	5	2	3	4	2	3	6	2	2	2	10