MOEX GROUP

MOEX.COM

ФРЕЙМВОРК ПО УПРАВЛЕНИЮ РИСКАМИ ИИ

Новые ключевые технологии

Облака

Инструмент масштабируемости в условиях импортозамещения

1

Open API

Историческое развитие инструментов работы с открытыми интерфейсами. Область расширения экспертизы – шеринг экспертизы вовне

2

3

Квантовая криптография

Готовность к квантовым угрозам – обеспечение защиты и хранение информации

4

AI / ML

Одновременно область развития бизнеса и инструмент оптимизации внутренних процессов

ЗНАЧИМОСТЬ УПРАВЛЕНИЯ РИСКАМИ

MOEX GROUP

Вызовы, которые стоят в эпоху распространения ИИ

Обработка больших данных

Использование искусственного интеллекта требует обработки огромных объемов данных, в том числе конфиденциальной информации и персональных данных

Усложнение кибератак

Искусственный интеллект активно развивается не только в деятельности коммерческих компаний, но и широко используется злоумышленниками для своих атак

Этические вызовы

ИИ может усугубить дискриминацию, если обучается на смещенных данных (например, непреднамеренное отторжение кандидатов из определенных демографических групп)

Усложнение моделей ИИ

Сложность понимания, как ИИ принимает решения, может привести к непредсказуемым последствиям (например, отказ в кредите без объяснения причин)

Терминология и область применения фреймворка

ТЕРМИНОЛОГИЯ

Искусственный интеллект (ИИ): 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.

Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, программное обеспечение (в том числе в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений.

Технологии искусственного интеллекта (технологии ИИ): совокупность технологий, включающая в себя компьютерное зрение, обработку естественного языка, распознавание и синтез речи, интеллектуальную поддержку принятия решений и перспективные методы искусственного интеллекта[1]

Система искусственного интеллекта (система ИИ): техническая система, использующая одну или несколько моделей ИИ, которая порождает такие конечные результаты, как контент, прогнозы, рекомендации или решения для заданного набора определенных человеком целей[2]

Модель искусственного интеллекта (модель ИИ): 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[1].

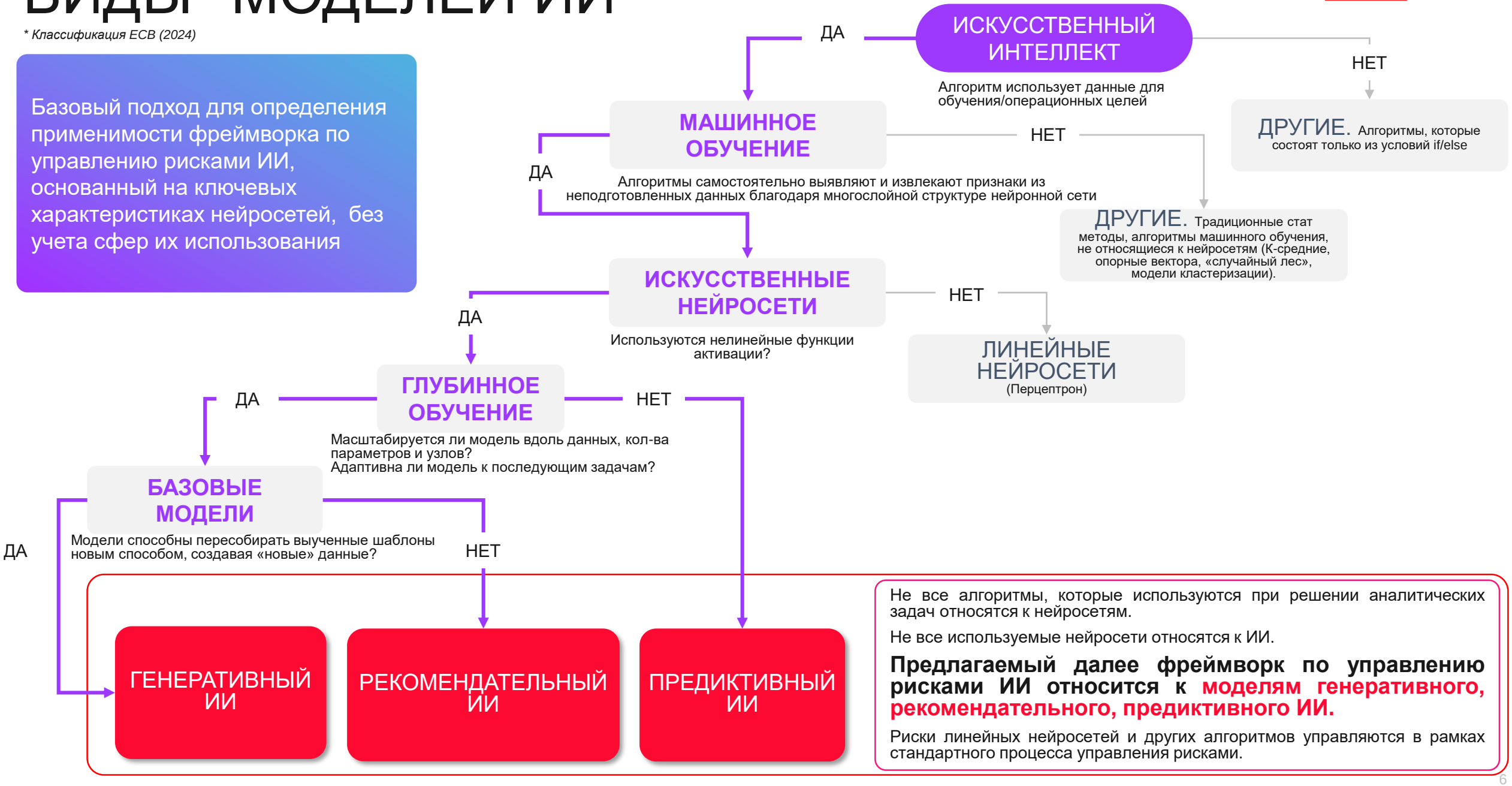
[1] Указ Президента РФ от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» (далее – Указ Президента РФ).

[2] ГОСТ Р 71476-2024 (ИСО/МЭК 22989:2022) «Искусственный интеллект. Концепции и терминология искусственного интеллекта» (утв. и введен в действие приказом Федерального агентства по техническому регулированию и метрологии от 28 октября 2024 г. № 1550-ст)

ВИДЫ* МОДЕЛЕЙ ИИ

** Классификация ЕСВ (2024)*

Базовый подход для определения применимости фреймворка по управлению рисками ИИ, основанный на ключевых характеристиках нейросетей, без учета сфер их использования



Не все алгоритмы, которые используются при решении аналитических задач относятся к нейросетям.

Не все используемые нейросети относятся к ИИ.

Предлагаемый далее фреймворк по управлению рисками ИИ относится к моделям генеративного, рекомендательного, предиктивного ИИ.

Риски линейных нейросетей и других алгоритмов управляются в рамках стандартного процесса управления рисками.

Управляемость ИИ и жизненный цикл систем ИИ

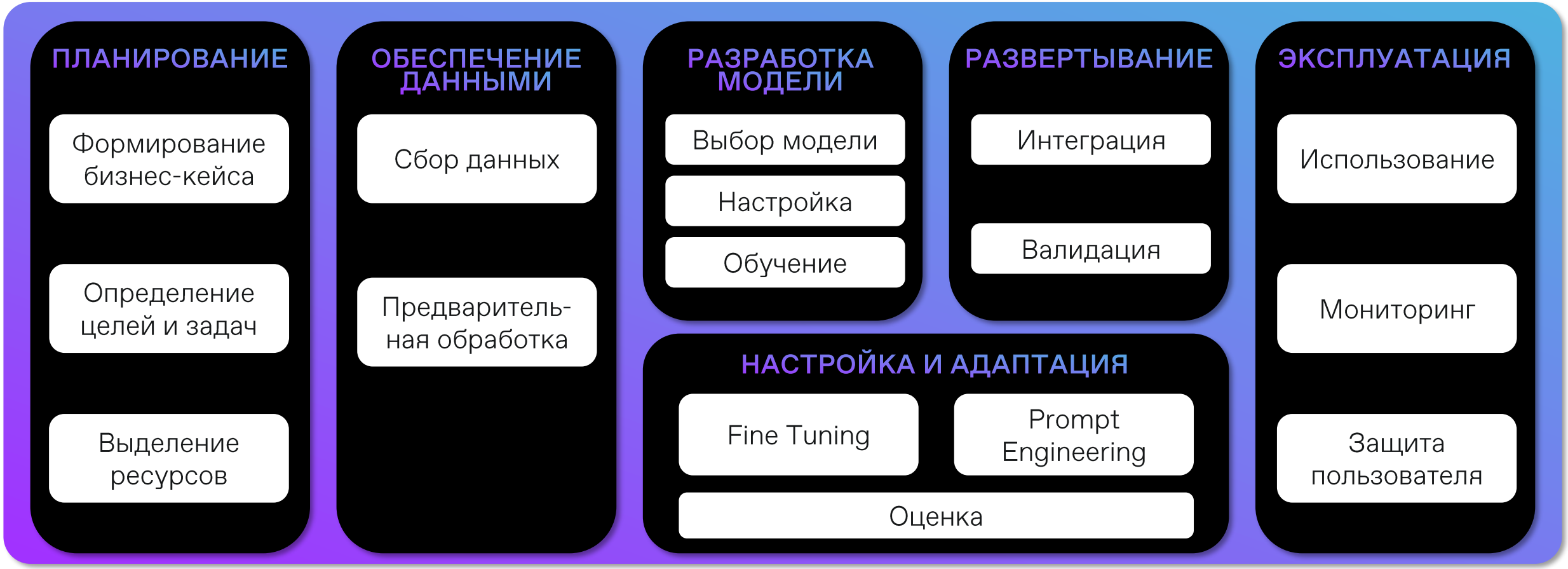
УРОВНИ УПРАВЛЯЕМОСТИ ИИ

Уровень управления	Тип управления	Область применения
«управление в курсе» ***(Governance-loop)*** ***Задает правила***	*Формирование и утверждение Правлением метрик, политик, аппетита, реперных точек*	*Политики, матрицы полномочий, Риск-аппетит, метрики риска, КИР, дашборды и тд*
«человек в курсе» ***(Human-on-the-Loop, HOTL)*** ***Модель принимает решение, человек наблюдает***	*Мониторинг человеком действий модели и контроль отклонений*	*Некритичные процессы, Процессы, требующие высокой скорости исполнения Системы со средним и низким уровнем риска*
«человек в контуре» ***(Human-in-the-Loop, HITL)*** ***Человек принимает решение на основе рекомендации модели***	*Ручная валидация и утверждение решения модели*	*Критичные процессы, системы с высоким риском ИИ*

ЖИЗНЕННЫЙ ЦИКЛ СИСТЕМЫ ИИ

концепция MLSecOps (включая специфику ГенИИ)

MOEX GROUP



MLSecOps — подход, который объединяет операционные аспекты машинного обучения с вопросами безопасности. Направлен на снижение рисков, которые могут принести модели AI/ML/GenAI в организацию.

Понятие риска ИИ и его место в системе управления рисками

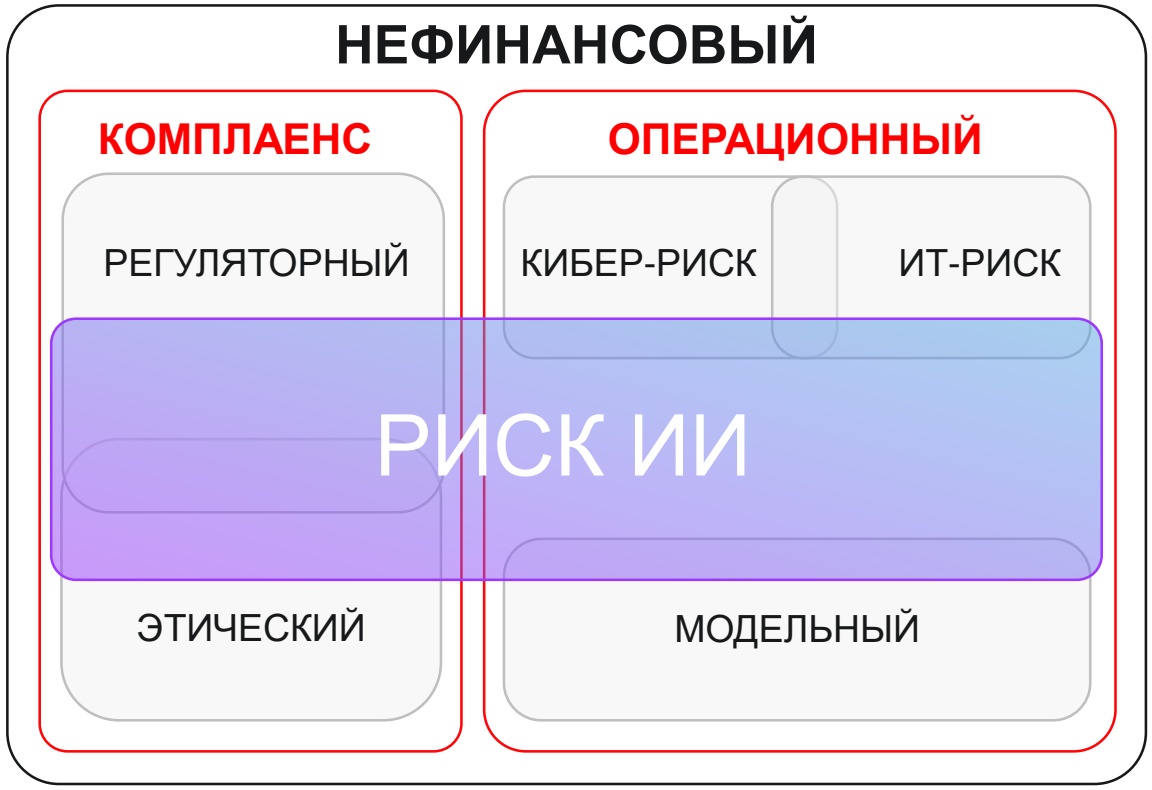
УПРАВЛЕНИЕ РИСКАМИ ИИ интегрировано в существующую СУР

MOEX GROUP

Риск ИИ – подвид нефинансового риска, находящийся на стыке:

- ИБ-рисков
- ИТ-рисков
- Модельных рисков
- Классических операционных рисков
- Этических
- Регуляторных

ВЗАИМОСВЯЗЬ РИСКА ИИ С ДРУГИМИ РИСКАМИ

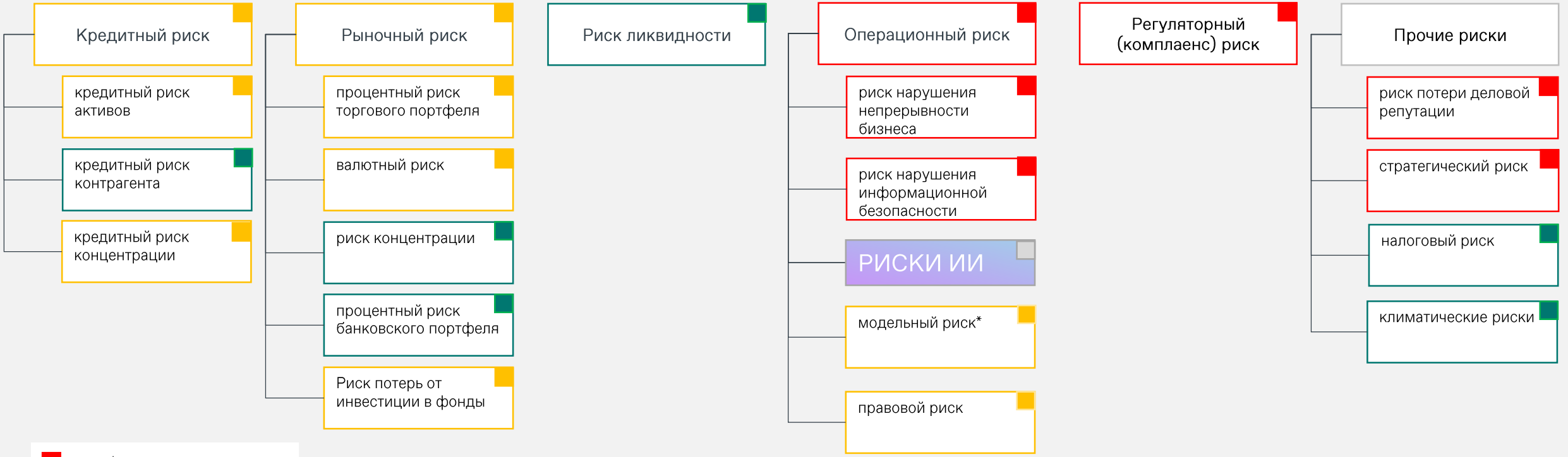


СТРУКТУРА РЕКОМЕНДАЦИЙ В СФЕРЕ ИИ

Управление рисками ИИ строится в соответствии с положениями ВнД по управлению рисками, утвержденных НС

КАРТА РИСКОВ МОСКОВСКОЙ БИРЖИ

КАРТА РИСКОВ МОСКОВСКАЯ БИРЖА



- *наиболее значимые риски*
- *риски средней значимости*
- *риски низкой значимости*
- *исследуется потребность выделения в отдельный риск. Решение будет принято в 2026*

Список рисков, управление которыми обеспечивается на уровне Группы:

1. Климатический риск – распространение практики управления, реализованной в Московской Бирже, на уровень Группы
2. Риски ESG - внедрение системы управление на уровне Группы
3. Коммерческий риск – управление на уровне Группы
4. Системный риск - управляется в рамках иных видов риска

*включая остаточные риски

ЦИКЛ ПРОЦЕССА УПРАВЛЕНИЯ РИСКАМИ ИИ

MOEX GROUP

ВЫЯВЛЕНИЕ РИСКОВ ИИ

- **Офис по развитию ИИ**
- **Управление операционных рисков**
- **УИБ**
- **Бизнес-владелец риска (пользователь сервиса на базе ИИ)**

- Идентификация рисков ИИ в моделях и процессах, в которых внедряется ИИ
- Составление матрицы рисков ИИ

ОЦЕНКА И ПРИСВОЕНИЕ УРОВНЯ РИСКА

Управление операционных рисков

- Оценка влияния реализации рисков ИИ
- Оценка негативных исходов от галлюцинаций моделей
- Анализ негативных исходов и оценка последствий в долгосрочной и краткосрочной перспективе на различные области риска

МИНИМИЗАЦИЯ РИСКОВ

Управление операционных рисков

- Определение мер по снижению уровня риска и лиц, ответственных за реализацию указанных мер

ВЕДЕНИЕ БАЗЫ ДАННЫХ РИСК-СОБЫТИЙ

Управление операционных рисков

- Выявление и регистрация риск-событий с описанием обстоятельств, повлекших его наступление, а также мероприятий, направленных на минимизацию последствий

УПРАВЛЕНИЕ РИСКАМИ ИИ ИНТЕГРИРОВАНО В ОБЩУЮ СИСТЕМУ СУР

МОНИТОРИНГ И КОНТРОЛЬ

- **Управление операционных рисков**
- **УИБ**

- Установка КИР и их порогов
- Мониторинг поведения моделей
- Мониторинг достаточности применяемых мер защиты

РЕАГИРОВАНИЕ НА СОБЫТИЯ РИСКА

Управление операционных рисков

- Определение типа риск-события
- Осуществление мер, направленных на минимизацию их последствий
- Предоставление возможности сотрудникам и клиентам направлять жалобы и сообщения о риск-событиях

УЧЕТ ПРИМЕНЯЕМЫХ МОДЕЛЕЙ ИИ

- **Управление операционных рисков**
- **Офис ИИ**

- Ведение реестра моделей
- Валидация высокорисковых моделей
- Регулярная инвентаризация моделей

РОЛИ И ОТВЕТСТВЕННОСТЬ

в управлении риском ИИ

Первая линия

Бизнес-владелец риска

Владелец модели/Пользователь

руководитель структурного подразделения, на бизнес-процессы которого реализация риска ИИ оказывает негативное влияние

- Соблюдение политик и правил использования ИИ-сервисов
- Ответственное обращение с данными, подаваемыми на вход
- Отслеживание и отчетность о подозрительных или аномальных результатах работы ИИ
- Участие в тестировании и обратной связи по функциональности ИИ
- Ознакомление с ограничениями ИИ

Владелец риска

Владелец модели/Разработчик

владелец бизнес-процесса или его этапа, в ходе осуществления которого появляются обстоятельства, обусловливающие возможную реализацию риска ИИ

- Обеспечение прозрачности и интерпретируемости алгоритмов ИИ
- Тестирование моделей на корректность и непредвзятость результата
- Разработка механизмов контроля качества
- Документирование процессов разработки, обучения моделей и источников данных
- Обеспечение соблюдения нормативных требований

Вторая линия

- Оценка и классификация рисков ИИ
- Разработка и внедрение политики управления рисками ИИ
- Мониторинг систем ИИ на соответствие требованиям безопасности и устойчивости к атакам
- Координация между разработчиками и пользователями для синхронизации мер по снижению рисков
- Реагирование на инциденты
- Учет применяемых моделей ИИ
- Выявление рисков ИИ
- Оценка и присвоение уровня риска ИИ
- Мониторинг и контроль рисков ИИ
- Минимизация выявленных рисков ИИ
- Реагирование на реализовавшиеся риски ИИ
- Ведение базы риск-событий

Оценка эффективности управления рисками в рамках выполнения первой и второй линиями задач в зоне своей ответственности (не позднее 3 месяцев после запуска)

Третья линия (внутренний аудит)

Ответственность за управление риском ИИ определена в соответствии с классической моделью трех линий на всем жизненном цикле модели

14

Определение уровня риска системы ИИ

ПАРАМЕТРЫ ПРИСВОЕНИЯ УРОВНЯ РИСКА СИСТЕМЕ ИИ

MOEX GROUP

ЗНАЧИМОСТЬ РИСКА

КРИТИЧНОСТЬ СИСТЕМЫ

01 **Критичность данных.** Категория данных, используемых при разработке модели ИИ

02 **Функциональная критичность**. Сфера применения модели ИИ, **критичность процессов**, для которых она используется

03 **Потенциальный ущерб**. Размер убытков, в тч регуляторных или ущерб деловой репутации, которые могут быть причинены в случае реализации риска

04 **Массовость потребления.** Количество клиентов/внутренних пользователей, при оказании услуг которым применяется ИИ.

ВЕРОЯТНОСТЬ РЕАЛИЗАЦИИ РИСКА

ОБЪЕМ ПОКРЫТИЯ СИСТЕМЫ

01 **Объяснимость решений**, принимаемых ИИ, наличие механизмов валидации и оценки качества выдаваемых результатов

02 **Чистота данных,** используемых в модели. Использование наборов данных, полученных от третьих лиц, или наборов данных, находящихся в открытом доступе в информационно-телекоммуникационной сети «Интернет»

03 **Точность модели** (комплексное определение на основе технико-статистических параметров Accuracy, Completeness, Specificity, Error Rate и тд)

04 **Наличие риск-событий, связанных с реализацией модельного риска**

ЗАЩИЩЕННОСТЬ СИСТЕМЫ

01 **Уровень зависимости от внешних факторов.** Участие третьих лиц в создании/эксплуатации модели ИИ, технологические ограничения (зависимость от данных, систем, оборудования)

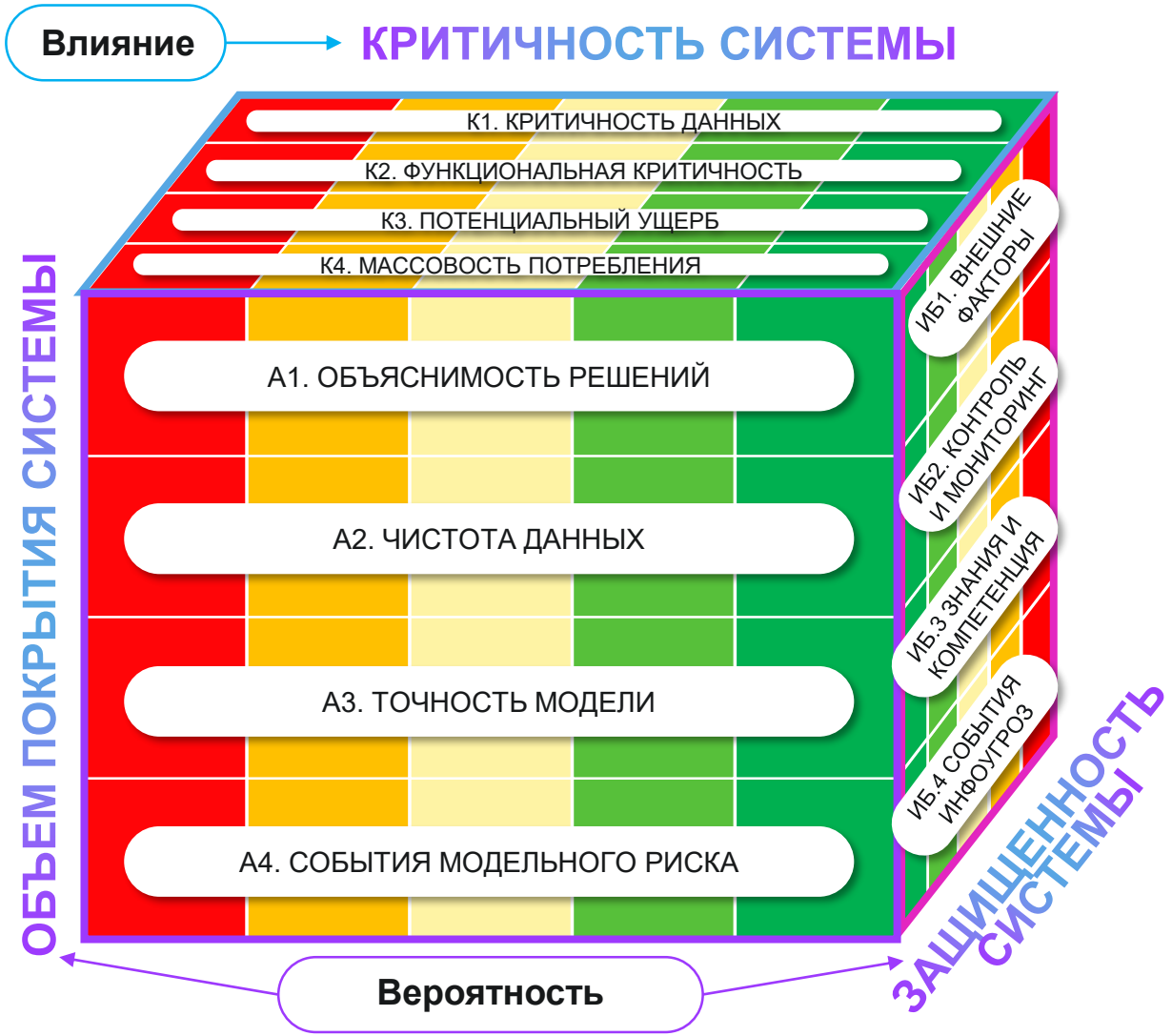
02 **Уровень контроля и мониторинга модели. Модели угроз.** Наличие в модели механизмов контроля и защиты от нарушителей (Защита от атак, защита данных, системная безопасность и тд). Их качество

03 **Уровень знаний и компетенций** владельцев модели, разработчиков, пользователей модели

04 **Наличие риск-событий, связанных с реализацией инфоугроз** в отношении модели ИИ

ПАРАМЕТРЫ ПРИСВОЕНИЯ УРОВНЯ РИСКА СИСТЕМЕ ИИ

Оценка производится по 12 критериям, позволяющим оценить совокупный уровень риска ИИ для каждой конкретной системы



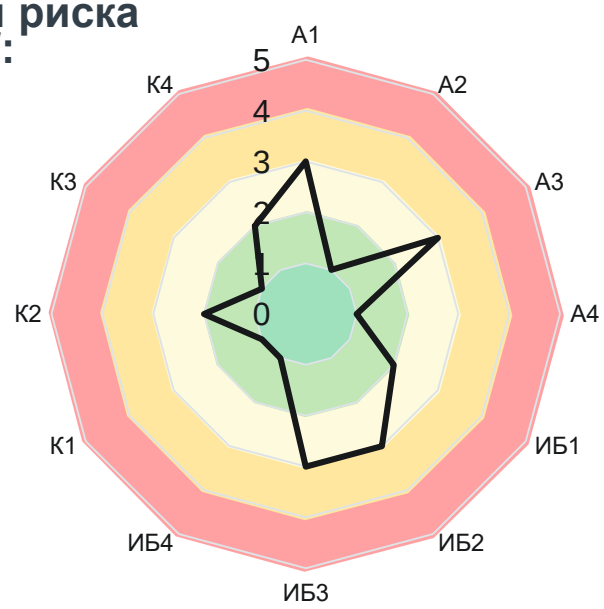
Распределение уровня риска по критериям /пример/:

Пример распределения критериев уровня риска для RAG- модели (RAG-RM), используемой для поиска информации по документам операционных рисков.

Для данной модели получена следующая бальная оценка:

Вероятность реализации риска (критерии А + критерии ИБ) = **17**

Влияние (критерии К) = **6**



Совокупный уровень риска системы ИИ:

Вероятность		Влияние: минимальное	низкое	среднее	значительное	высокое
		0-4	5-8	9-12	13-16	17-20
минимальная	0-8					
низкая	9-16					
средняя	17-24		RAG-RM			
значительная	25-32					
высокая	33-40					

ИСТОЧНИКИ ИНФОРМАЦИИ О РИСКАХ ИИ

ВНУТРЕННИЕ

ИТ-система

- Отчеты по результатам тестирований (DR, нагрузка, регрессионное, UAT и тд);
- Данные ИТ-мониторинга;
- Данные по результатам DQ-анализа.

Система ИИ

- Результаты валидации (производительность, корректность, устойчивость);
- Отчеты по аудиту модели;
- Данные базы событий модельного риска.

Объект СЗИ

- Результаты пентестов;
- Данные мониторинга ИБ;
- Данные базы инцидентов ИБ.

Объект СУР

- Результаты: RCSA, оценки новых продуктов, оценки рисков третьих сторон, анализа бизнес-процессов, сценарного анализа, стресс-теста;
- Результаты тренингов;
- Данные БДР, БДСОР, КИР, внутренних аудитов;
- Добровольные сообщения и тд

**перечень не является исчерпывающим и пополняется с появлением новых источников информации о рисках ИИ*

ВНЕШНИЕ

Отчеты и исследования Банка России

Российские НПА по теме рисков ИИ

Международные НПА по теме рисков ИИ

Международные исследования, отчеты и диссертации в области ИИ

Открытые базы атак и уязвимостей

Открытые базы этических принципов

Открытые проекты и стандарты

(АФТ, OWASP, NIST, MITRE и др.)

Результаты внешних аудитов

Открытые фреймворки

(AI Sec, AI TRISM, AI RMF, AI RedTeam, SAIF Map, MLDevSecOps, DASF и др)

Открытые форумы и сообщества

(Git, Habr, cyberorda и др)

Открытые рекомендации представителей сообщества

(Яндекс, Сбер, РТ и др.)

Управление рисками на этапах жизненного цикла систем ИИ

УПРАВЛЕНИЕ РИСКАМИ ИНТЕГРИРОВАНО В ЖЦ СУР

MOEX GROUP

Объем вовлечения второй линии

ПРОДУКТОВЫЙ РИСК-МЕНЕДЖМЕНТ – *заранее делаем все, чтобы минимизировать риски до момента публикации системы в прод*

КЛАССИЧЕСКИЙ РИСК-МЕНЕДЖМЕНТ

ПЛАНИРОВАНИЕ

- ВЫЯВЛЕНИЕ РИСКОВ ИИ
- ОЦЕНКА И ПРИСВОЕНИЕ УРОВНЯ РИСКА
- МИНИМИЗАЦИЯ РИСКОВ

ОБЕСПЕЧЕНИЕ ДАННЫМИ

- АКТУАЛИЗАЦИЯ РИСКОВ ИИ
- ОЦЕНКА И ПРИСВОЕНИЕ УРОВНЯ РИСКА
- МИНИМИЗАЦИЯ РИСКОВ

РАЗРАБОТКА МОДЕЛИ

- АКТУАЛИЗАЦИЯ РИСКОВ ИИ
- ОЦЕНКА И ПРИСВОЕНИЕ УРОВНЯ РИСКА
- МИНИМИЗАЦИЯ РИСКОВ

РАЗВЕРТЫВАНИЕ

- АКТУАЛИЗАЦИЯ РИСКОВ ИИ
- ОЦЕНКА И ПРИСВОЕНИЕ УРОВНЯ РИСКА
- МИНИМИЗАЦИЯ РИСКОВ

НАСТРОЙКА И АДАПТАЦИЯ

- АКТУАЛИЗАЦИЯ РИСКОВ ИИ
- ОЦЕНКА И ПРИСВОЕНИЕ УРОВНЯ РИСКА
- МИНИМИЗАЦИЯ РИСКОВ
- МОНИТОРИНГ РИСКОВ ИИ

ЭКСПЛУАТАЦИЯ

- АКТУАЛИЗАЦИЯ РИСКОВ ИИ
- ОЦЕНКА И ПРИСВОЕНИЕ УРОВНЯ РИСКА
- МИНИМИЗАЦИЯ РИСКОВ
- УЧЕТ ПРИМЕНЯЕМЫХ МОДЕЛЕЙ ИИ
- МОНИТОРИНГ РИСКОВ ИИ
- РЕАГИРОВАНИЕ НА СОБЫТИЯ РИСКА
- ВЕДЕНИЕ БАЗЫ ДАННЫХ РИСК-СОБЫТИЙ

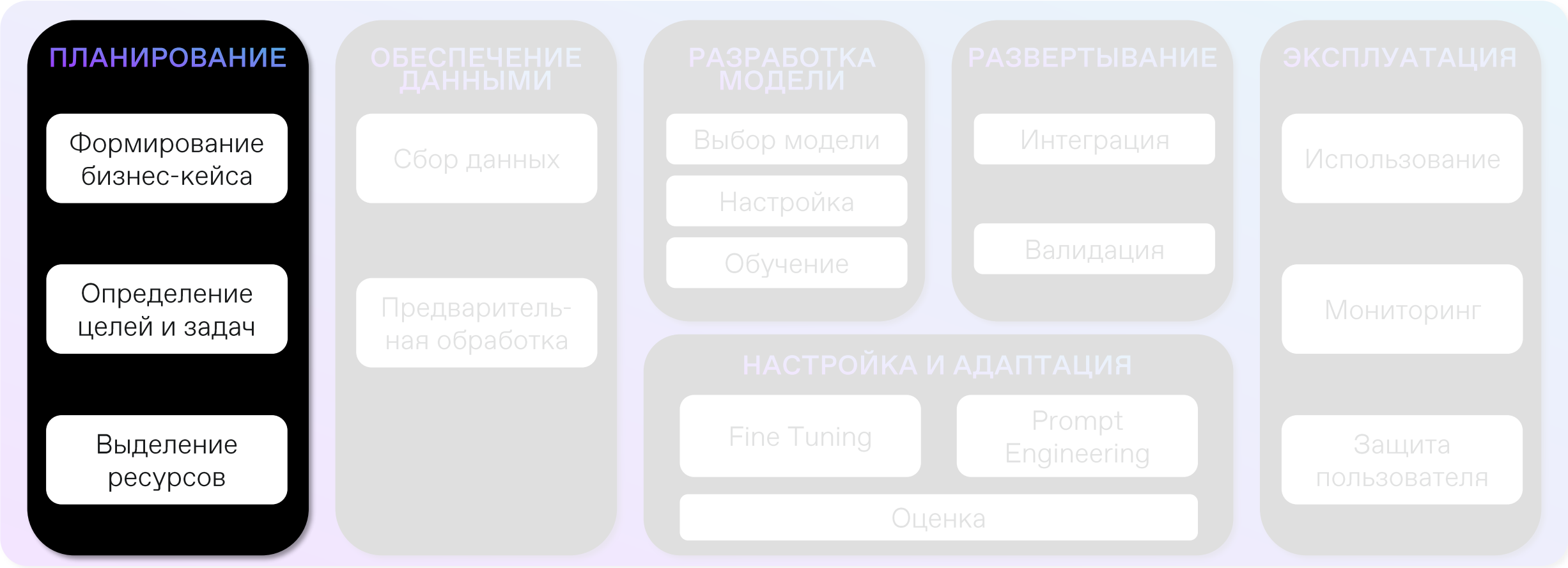
На каждом из этапов ЖЦ осуществляются действия по управлению рисками, конечная цель которых – не допустить реализацию риска ИИ и поддерживать совокупный уровень риска СИСТЕМЫ ИИ на приемлемом уровне.

Каждый идентифицированный риск проходит классическую оценку.
Уровень риска, выявленного на этапе жизненного цикла системы ИИ, определяется в соответствии с ВНД и учитывается при определении совокупного уровня риска системы ИИ.

ЖИЗНЕННЫЙ ЦИКЛ СИСТЕМЫ ИИ

концепция MLSecOps (включая специфику ГенИИ)

MOEX GROUP



MLSecOps — подход, который объединяет операционные аспекты машинного обучения с вопросами безопасности. Направлен на снижение рисков, которые могут принести модели AI/ML/GenAI в организацию.

MOEX GROUP

СОСТАВНЫЕ ЭЛЕМЕНТЫ ЭТАПА ЖИЗНЕННОГО ЦИКЛА

Формирование бизнес-кейса

Определение целей и задач

Выделение ресурсов

Сформированы цели и задачи, которые должна решать система ИИ.

Определены процессы, в которых будет использоваться система ИИ.

Определены потребители.

КЛЮЧЕВЫЕ РИСКИ

СТРАТЕГИЧЕСКИЙ

Риски черного ящика как стратегическая и юридическая уязвимость, проблемы ответственности за принимаемые решения, проблемы автономности ИИ

РИСК ТРЕТЬИХ СТОРОН

Стратегическая зависимость от одного поставщика ИИ-услуг
- Техническая (технологии)
- Зависимость от данных
- Сервисная
- Экономическая и ресурсная
- Зависимость от «личности» модели

ПРАВОВЫЕ И ЭТИЧЕСКИЕ РИСКИ

Проблемы интеллектуальной собственности (включая права на голос, изображение, сгенерированный контент), ограничения использования OSS, проблемы дискриминации и нарушения морально-этических норм и др.

Ключевые последствия:
- *когнитивный раскол на пользователей и оракулов;*
- *атрофия управленческих компетенций;*
- *скрытый саботаж;*
- *судебные разбирательства, связанные с нарушением интеллектуальных прав;*
- *судебные разбирательства, связанные с дискриминацией и нарушением морально-этических норм.*

ИНСТРУМЕНТАРИЙ СУР/MLSecOps

1. Анализ категорий потребителей результатов модели, анализ критичности модели. Определение ограничений модели

↓

Определение значимости модели

Определение политик конфиденциальности, правовых и этических ограничений использования

2. Анализ предполагаемой схемы реализации. Оценка техстека, зависимости от третьих лиц

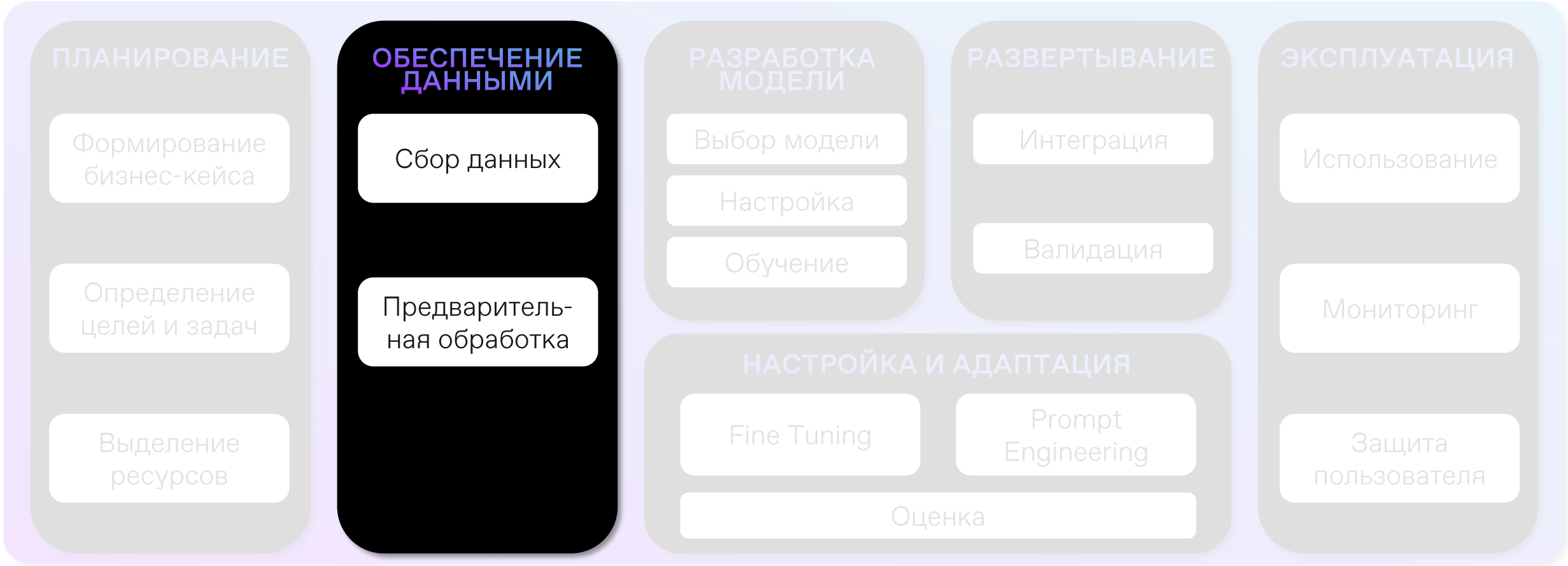
↓

Определение уровня риска зависимости от третьих сторон

ЖИЗНЕННЫЙ ЦИКЛ СИСТЕМЫ ИИ

концепция MLSecOps (включая специфику ГенИИ)

MOEX GROUP



MLSecOps — подход, который объединяет операционные аспекты машинного обучения с вопросами безопасности. Направлен на снижение рисков, которые могут принести модели AI/ML/GenAI в организацию.

MOEX GROUP

СОСТАВНЫЕ ЭЛЕМЕНТЫ ЭТАПА ЖИЗНЕННОГО ЦИКЛА

- Сбор данных
- Разметка данных
- Обработка данных
- Формирование признаков

Сформирован обработанный, высококачественный набор данных, который подчиняется некоторой закономерности.

КЛЮЧЕВЫЕ РИСКИ

УТЕЧКИ ДАННЫХ

Риск утечки конфиденциальных данных, таких как ПДн, бизнес-информация или алгоритмы.

РИСКИ ЦЕПОЧКИ ПОСТАВОК

Уязвимости в цепочке поставок могут привести к нарушениям безопасности и смещению данных.

ОТРАВЛЕНИЕ ДАННЫХ/ФУНКЦИЙ

Манипуляции с данными на этапах обеспечения данными и обучения модели, что влияет на ее надежность и результаты.

ИНСТРУМЕНТАРИЙ CYP/MLSECOPS

Обеспечение организованного контроля доступа к данным

(ролевая модель, аутентификация, DLP)

Классификация и управление данными

(классификация, настройка DQ, DQ для оценочных данных, организация безопасного хранения данных, разные пайплайны разработки для данных разной критичности)

Отслеживание происхождения данных

(определение формата информации о происхождении, формирование White-List источников, внедрение контроля модификации)

Защита от отравления данных

(Внедрение контроля целостности, внедрение алгоритмов обнаружения аномалий и предотвращения атак)

Защита извлеченных признаков от манипулирования

(использование шума во время извлечения, орг. безопасного хранилища)

Организация платформы управления данными

(безопасная цепочка поставок, шифрование при хранении и передаче, доверенная база компонентов, руководства по безопасной работе, автоматизация конвейера по безопасной работе с данными, формализация ЖЦ обеспечения данными)

ЖИЗНЕННЫЙ ЦИКЛ СИСТЕМЫ ИИ

концепция MLSecOps (включая специфику ГенИИ)

MOEX GROUP

ПЛАНИРОВАНИЕ

- Формирование бизнес-кейса
- Определение целей и задач
- Выделение ресурсов

ОБЕСПЕЧЕНИЕ ДАННЫМИ

- Сбор данных
- Предварительная обработка

РАЗРАБОТКА МОДЕЛИ

- Выбор модели
- Настройка
- Обучение

РАЗВЕРТЫВАНИЕ

- Интеграция
- Валидация

ЭКСПЛУАТАЦИЯ

- Использование
- Мониторинг
- Защита пользователя

НАСТРОЙКА И АДАПТАЦИЯ

- Fine Tuning
- Prompt Engineering
- Оценка

MLSecOps — подход, который объединяет операционные аспекты машинного обучения с вопросами безопасности. Направлен на снижение рисков, которые могут принести модели AI/ML/GenAI в организацию.

MOEX GROUP

СОСТАВНЫЕ ЭЛЕМЕНТЫ ЭТАПА ЖИЗНЕННОГО ЦИКЛА

- Проектирование архитектуры
- Настройка модели
- Обучение и тестирование
- Fine Tuning
- Prompt Engineering
- Оценка и валидация

Разработана модель, соответствующая требованиям к точности, надежности, масштабируемости, скорости, устойчивости к шуму и неточным данным.

Модель обучаема и адаптивна, ее результаты объяснимы.

КЛЮЧЕВЫЕ РИСКИ

УЯЗВИМОСТИ ВЕКТОРОВ И ЭМБЕДДИНГОВ

Слабые места в моделях векторных представлений, способные вызвать непредсказуемые ошибки.

НЕАВТОРИЗОВАННЫЕ ДАННЫЕ ДЛЯ ОБУЧЕНИЯ

Модель делает вывод на основе данных, к которым у нее не должно быть доступа. Может привести к утечке конфиденциальной информации, непредсказуемому поведению модели.

ПРОМТ-ИНЪЕКЦИИ

Вмешательство пользователя в запросы, чтобы изменить результаты и функции модели.

ОТРАВЛЕНИЕ МОДЕЛИ/ ПОДДЕЛКА ИСТОЧНИКОВ

Манипуляции с данными на этапах обеспечения данными и обучения модели, что влияет на ее надежность и результаты, может повлечь риск фальсификации исходного кода модели.

НЕКОРРЕКТНАЯ/ИЗБЫТОЧНАЯ ОБРАБОТКА ВЫХОДНЫХ ДАННЫХ

Отсутствие проверок и обработки вывода может привести к уязвимостям в приложениях, сохранению или использованию данных сверх допустимых пределов, недостаточной детализации ответов.

ИНСТРУМЕНТАРИЙ СУР/MLSECOPS

Контроль входных/выходных данных, использование доверенных компонент

(проверка и нормализация вводимых данных, логирование и мониторинг данных, определение формата выходных значений, проверка и фильтрация выходных значений, установка границ доверия между моделью, внешними источниками и расширяемой функциональностью, введение этических ограничений, децентрализация, Federated Learning)

Состязательное обучение

(Adversarial Training - генерация adversarial примеров и включение их в процесс обучения для повышения точности и устойчивости к атакам.

Дифференциально приватный SGD (DP-SGD) - обеспечивает конфиденциальность за счёт равномерного распределения privacy costs по всем этапам обучения.)

Тестирование и валидация

(анализ защищенности модели, внедрение методов интерпретации результатов и оценки точности

Проверки обучаемости – кросс-валидация. Контроль производительности на валидации, мониторинг точности модели на валидационных данных и тд

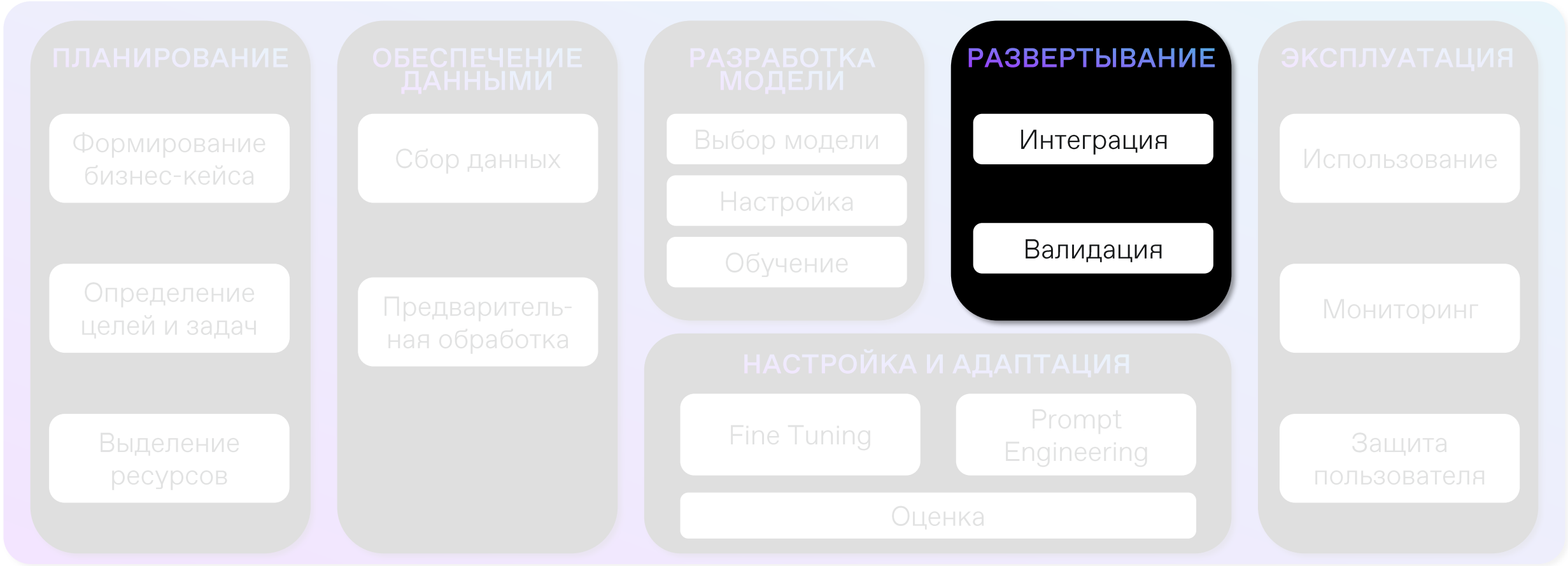
Проверки адаптивности - Онлайн обучение, Transfer Learning, A/B тесты, Concept Drift Detection и тд

Проверки объяснимости - SHAP, LIME, PDP, Feature Importance, библиотеки интерпретации и тд)

Анализ защищённости – тестирование на устойчивость, анализ смещений (Bias Detection)

ЖИЗНЕННЫЙ ЦИКЛ СИСТЕМЫ ИИ

концепция MLSecOps (включая специфику ГенИИ)



MLSecOps — подход, который объединяет операционные аспекты машинного обучения с вопросами безопасности. Направлен на снижение рисков, которые могут принести модели AI/ML/GenAI в организацию.

MOEX GROUP

СОСТАВНЫЕ ЭЛЕМЕНТЫ ЭТАПА ЖИЗНЕННОГО ЦИКЛА

- Разработка
- Сборка
- Тестирование системы
- Развертывание
- Настройка мониторинга

Система ИИ готова к началу эксплуатации

КЛЮЧЕВЫЕ РИСКИ

УЯЗВИМОСТИ ВЕКТОРОВ И ЭМБЕДДИНГОВ

Слабые места в моделях векторных представлений, способные вызвать непредсказуемые ошибки.

УЯЗВИМОСТИ КОДА

Слабые места в коде системы, позволяющие злоумышленнику осуществлять атаки на систему и саму модель.

УТЕЧКИ ДАННЫХ И СИСТЕМНЫХ ИНСТРУКЦИЙ

Утечка внутренних системных подсказок, которые раскрывают конфиденциальные настройки.

РИСКИ ЦЕПОЧКИ ПОСТАВОК

Уязвимости в цепочке поставок могут привести к нарушениям безопасности и смещению данных.

НЕКОРРЕКТНАЯ/ИЗБЫТОЧНАЯ ОБРАБОТКА ВЫХОДНЫХ ДАННЫХ

Отсутствие проверок и обработки вывода может привести к уязвимостям в приложениях, сохранению или использованию данных сверх допустимых пределов.

ИНСТРУМЕНТАРИЙ СУР/MLSECOPS

Контроль и воспроизводимость проводимых экспериментов

(документирование и стандартизация экспериментов, использование контейнеризации)

Защита активов

(шифрование, контроль аномалий, создание центрального хранилища для моделей с контролем доступа, внедрение в активы цифрового водяного знака, Certified Defenses, DLP-фильтрация, контент-фильтрация, SBOM,SPDX)

Организация платформы управления разработкой

(организация хранилища обученных моделей, контроль версионности, разработка политик и руководств для процесса разработки, автоматизация конвейера разработки модели, формализация всего ЖЦ разработки модели, организация безопасной цепочки поставок)

MLOps-конвейер

(построение конвейера, автоматизированные инструментальные проверки, настройка и обновление сигнатурных правил, организация взаимодействия команд)

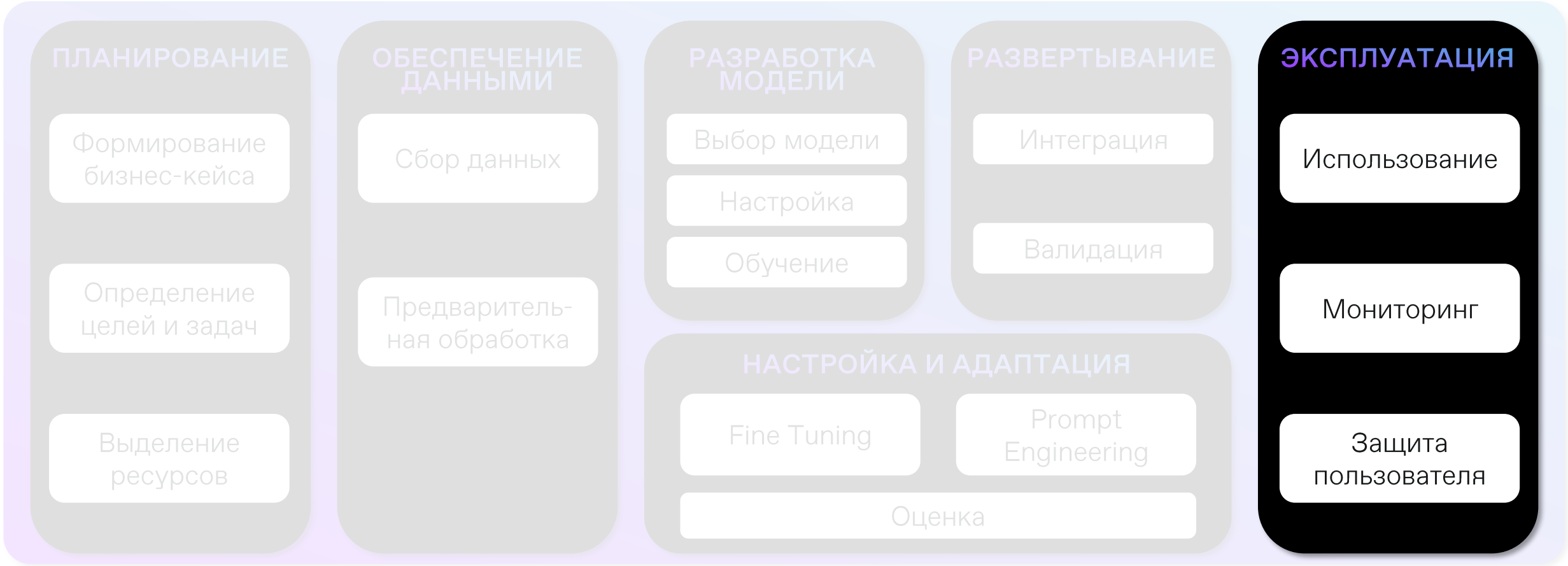
Анализ защищенности системы ИИ

(Robust Training, правила Malware-gen, firewall фильтрация retrieved-chunk, BugBounty, управление инцидентами, аудиты, модели угроз)

+ КЛАССИЧЕСКАЯ ИБ

ЖИЗНЕННЫЙ ЦИКЛ СИСТЕМЫ ИИ

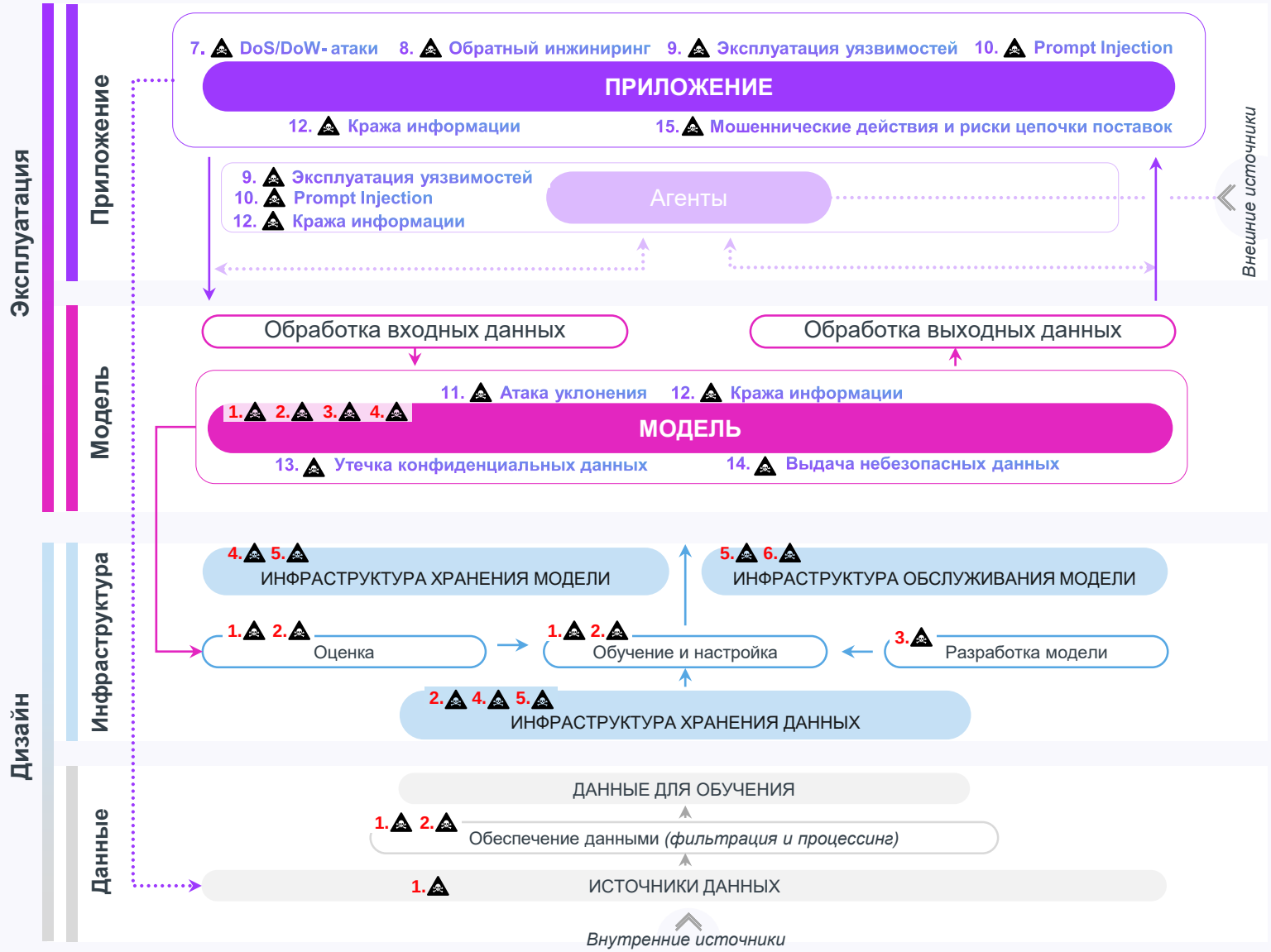
концепция MLSecOps (включая специфику ГенИИ)



MLSecOps — подход, который объединяет операционные аспекты машинного обучения с вопросами безопасности. Направлен на снижение рисков, которые могут принести модели AI/ML/GenAI в организацию.

MOEX GROUP

НЕПРЕРЫВНОЕ СОВЕРШЕНСТВОВАНИЕ СИСТЕМЫ КОНТРОЛЕЙ СУР/MLSECOPS



РИСКИ НА УРОВНЕ ПОЛЬЗОВАТЕЛЕЙ

ЧРЕЗМЕРНАЯ АГЕНТНОСТЬ
Чрезмерное использование LLM для принятия решений, выходящее за пределы безопасных границ.

ВВЕДЕНИЕ В ЗАБЛУЖДЕНИЕ
Генерация ложной или вводящей в заблуждение информации с потенциальными последствиями, галлюцинирование и хаотичное поведение модели.

НЕОГРАНИЧЕННОЕ ПОТРЕБЛЕНИЕ
Неконтролируемое использование вычислительных ресурсов, что вызывает перегрузку системы.

КЛЮЧЕВЫЕ РИСКИ НА УРОВНЕ ДИЗАЙНА

1. ОТРАВЛЕНИЕ ДАННЫХ
2. НЕАВТОРИЗОВАННЫЕ ДАННЫЕ
3. ПОДДЕЛКА ИСТОЧНИКА
4. ИЗБЫТОЧНАЯ ОБРАБОТКА ДАННЫХ
5. ЭКСФИЛЬТРАЦИЯ МОДЕЛИ
6. ФАЛЬСИФИКАЦИЯ МОДЕЛИ ПРИ РАЗВЕРТЫВАНИИ

ПРИЛОЖЕНИЯ

Типология атак на системы ИИ

Специфические векторы атак на различные типы систем ИИ

Пример применения фреймворка для оценки риска системы ИИ

ОСНОВНЫЕ ТИПЫ АТАК НА СИСТЕМЫ ИИ

MOEX GROUP

АТАКИ НА ЭТАПЕ ИНФЕРЕНСА (EVASION ATTACKS)

Цель: обмануть уже обученную модель на этапе прогнозирования.

Быстрые атаки на основе градиента (White-Box)

- FGSM (Fast Gradient Sign Method): Создает возмущение за один шаг, двигаясь в направлении знака градиента функции потерь.
- PGD (Projected Gradient Descent): Многократная итеративная версия FGSM.
- C&W (Carlini & Wagner): Мощная оптимизационная атака, минимизирующая величину возмущения при условии смены класса.

Атаки, ищущие минимальное возмущение

- DeepFool: Итеративно приближает образец к границе решения, находя минимальное необходимое возмущение для смены класса.
- One-Pixel Attack: Атака в крайне ограниченном пространстве - меняет всего один пиксель, демонстрируя хрупкость моделей.

Универсальные и физические атаки

- Universal Adversarial Perturbation: Генерирует одно возмущение, которое, будучи добавленным к любому изображению из датасета, с высокой вероятностью вызывает ошибку.
- Adversarial Patch: Создает небольшой патч (например стикер), который размещается в кадре и заставляет модель ошибаться. Критично для компьютерного зрения в реальном мире.
- Physical Adversarial Attack: Перенос атак в физический мир (например, наклейки на дорожные знаки для обмана автопилота).

АТАКИ НА КОНФИДЕНЦИАЛЬНОСТЬ (PRIVACY ATTACKS)

Цель: извлечь конфиденциальную информацию о данных обучения или о самой модели.

Атаки на восстановление данных (Model Inversion & Extraction)

- Model Inversion Attack: Угроза приватности.
- Model Extraction / Stealing Attack: Угроза интеллектуальной собственности.
- Attribute Inference Attack: Вывод скрытых атрибутов, не являющихся прямой целью предсказания, на основе выхода модели.

Атаки на членство в обучающей выборке (Membership Inference)

- Основная идея: Определить, присутствовал ли конкретный образец данных в обучающем наборе модели.

АТАКИ В УСЛОВИЯХ "ЧЕРНОГО ЯЩИКА" (BLACK-BOX ATTACKS)

Атакующий не имеет доступа к внутренностям модели (архитектуре, весам), только к ее выходу.

Атаки на основе запросов (Query-Based)

- Score-Based Attack: Атакующий видит вероятности классов, выдаваемые моделью.
- Decision-Based Attack: Атакующий видит только итоговый класс. Атака часто начинается с большого шума, который постепенно уменьшается, пока образец остается адверсариальным.
- Boundary Attack: Разновидность decision-based атаки, которая "скользит" вдоль границы решения, минимизируя возмущение.

Атаки, использующие перенос (Transferability Attacks)

- Transfer Attack: Создание адверсариального примера на локальной модели-суррогате (белый ящик) с последующей атакой с его помощью целевой black-box модели.

АТАКИ НА СПЕЦИФИЧЕСКИЕ АРХИТЕКТУРЫ И МОДАЛЬНОСТИ

NLP / LLM Attacks

- Text Adversarial Attack: Замена слов, добавление опечаток/изменений, меняющих решение модели, но не смысл для человека.
- Prompt Injection / Jailbreak Attack

Атаки на системы обучения с подкреплением (RL)

- Reward Hacking: Агент находит неожиданный способ максимизировать функцию вознаграждения, не решая поставленную задачу.
- Environment Poisoning: Изменение среды обучения, чтобы агент выработал неэффективную или опасную стратегию.

Атаки на графические и другие модели

- Graph Adversarial Attack: добавление/удаление связей или узлов.
- Audio Adversarial Attack: Добавление незаметного для человека шума в аудиозапись, чтобы обмануть систему распознавания речи.

АТАКИ НА ЭТАПЕ ОБУЧЕНИЯ (POISONING ATTACKS)

Цель: злоумышленник вмешивается в процесс обучения, чтобы подорвать работу модели.

Отравление данных (Data Poisoning)

- Внесение в обучающий набор вредоносных примеров, которые ухудшают общую производительность модели или смещают ее решения.

Бэкдор-атаки (Backdoor Attacks)

- Механизм: Модель обучается корректно работать на "чистых" данных, но при появлении в входе специального триггера она выдает заданный злоумышленником ошибочный результат/запускает исполнения команд/кода для исполнения задач злоумышленника.

ВЕКТОРЫ АТАК. CNN-сети

(распознавание изображений)

MOEX GROUP

	ВЕКТОР АТАКИ		СПОСОБ ЗАЩИТЫ
Входные данные	**Атака по пороговому шуму** (изменяет входн. Данные так, чтобы изменения влияли на выводы, но не на семантику)	**Adversarial Patch**	Обнаружение и удаление шумовых карт при предв обработке входных данных
		Hidden Trigger Backdoor	Обнаружение и удаление скрытых триггеров-взломов при предв обработке входных данных
		One-Pixel Attack	Обнаружение и удаление шумовых карт при предв обработке входных данных
	Атака методом соседних образцов (создает новые входн. Данные, близкие к исходным, но нарушающие предсказание)	**Decoy Attack**	Использование алгоритмов обнаружения DNS-угроз
		Evasion Attack	очистка adversarial examples и предварительное обучение модели. Преобразование данных (пороговое преобразование, вейвлет-преобразование, уменьшение глубины цвета, обрезка и масштабирование, сжатие JPEG)
		Hop Skip Jump Attack	
		FGSM (Fast Gradient Sign Method)	Антиградиентные методы обучения, Adversarial Training, обучение на «шумных» данных, поиск и предотвращение внедрений/атак, поиск аномалий
Обучение	**Атака на гиперпараметры** (изменяет гиперпараметры обучения)	**Hyperparameter Tuning**	Защита от атак на гиперпараметры. В тч методы обнаружения и удаления взломов в процессе предв обработки входных данных
		Hyperparameter Injection	
	Атака на выборку (отравление данных) (отравление неправильными или поддельными образцами)	**Data Poisoning**	Методы обнаружения и удаления взломов в процессе предв обработки входных данных
		Model Extraction	Защита моделей от извлечения (исп шифрования)
		Data Injection	Методы обнаружения и удаления взломов в процессе предв обработки входных данных
Выводы	**Атака на модель** (восстановление входных данных из вывода с целью получения доступа к конф информации)	**Model Inversion**	Защита от инверсии
		Model Extraction	Защита моделей от извлечения (исп шифрования)
		Model Stealing	Защита моделей от кражи (в тч защита через обучение модели)
	Атака на метаданные (вывод информации о модели, в тч архитектура, параметры обучения)	**Metadata Leakage**	Защита от утечек
		Metadata Injection	Защита от внедрения метаданных
Передача информации	**Атака на передачу информации** (вывод информации о модели, в тч архитектура, параметры обучения)	**Information Leakage**	Защита от утечек
		Side-Channel Attack	Защита от атак по сторонним каналам
		Model Extraction	Защита моделей от извлечения (исп шифрования)
Выборка данных	**Атака на выборку данных** (влияние на выборку для нарушения обучения)	**Membership Inference**	Защита от атак на выборку
		Data Leakage	Защита от утечек
		Data Sampling	Защита от атак на выборку

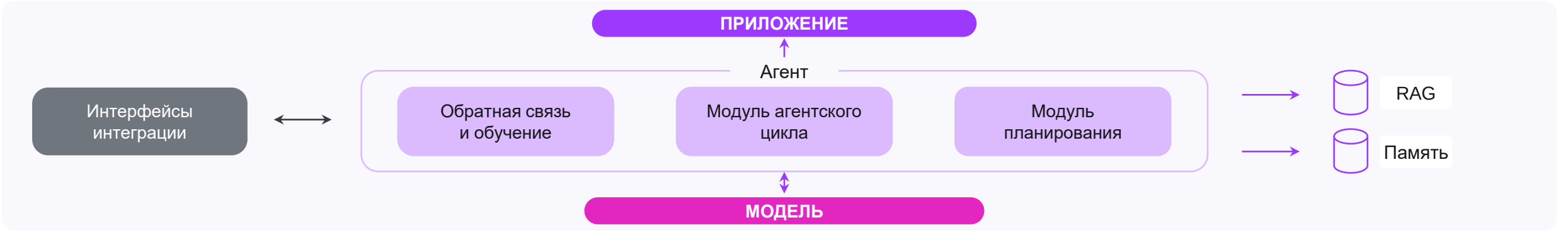
ВЕКТОРЫ АТАК. LLM-модели
(языковые модели)

	ВЕКТОР АТАКИ		СПОСОБ ЗАЩИТЫ
Входные данные	**Атака с чрезмерно длинными последовательностями текста** (цель – переполнение буфера и сбой)	**Text Bombing**	Фильтрация и ограничение длины входных данных. Обеспечение производительности модели для обработки больших данных
	Атака с внедрением вредоносных команд в запросы модели	**Command Injection**	Валидация и фильтрация входных данных. Контроль доступа и ограничения выполнения отд. команд
Обучение	**Атака на обучение с нежелательной информацией**	**Poisoning Attack**	Обнаружение аномалий, проверка данных на адекватность и соответствие ожиданию. Контроль за исходными данными
	Атака на распространение спама	**Spaming Attack**	Обнаружение и фильтрация спама. Ограничения на кол-во запросов от одного источника за период времени
Выводы	**Атака на изменение выходных данных** (изменение выходных данных для нарушения работы модели)	**Output Manipulation Attack**	Контроль за выходными данными. Проверка на адекватность и соответствие. Мониторинг и контроль за искажением
Передача информации	**Атака на передачу конфиденциальной информации**	**Data Leakage Attack**	Контроль доступа и шифрование. Мониторинг утечек и аудит использования
	Атака на вызов модели для нежелательных действий	**Misuse Attack**	Мониторинг и обнаружение неправомерного использования модели. Авторизация и аутентификация для ограничения доступа
Выборка данных	**Атака на выборку данных** с нежелательной информацией	**Data Poisoning Attack**	Обнаружение аномалий, проверка данных на адекватность и соответствие ожиданию. Контроль за исходными данными
Интерфейс	**Атака на использование небезопасных протоколов**	**Insecure Protocol Usage Attack**	Безопасные протоколы, их мониторинг и обновление
	Атака на использование небезопасных методов обмена данными	**Insecure Data Exchange Method Attack**	Безопасные методы обмена данными, сетевая защита. Мониторинг и обновление реестра уязвимостей

ВЕКТОРЫ АТАК. GenAI-модели
(генеративные)

	ВЕКТОР АТАКИ		СПОСОБ ЗАЩИТЫ
Входные данные	**Атака на входные данные** (вредоносные команды, искажения)	**Input Data Attack**	Валидация и фильтрация входных данных. Контроль доступа и ограничения выполнения отд. команд
Обучение	**Атака на генерацию**	**Generation Attack**	Обнаружение и фильтрация спама. Ограничения на кол-во запросов от одного источника за период времени
	Атака на дискриминатор	**Discriminator Attack**	Валидация и фильтрация входных данных. Контроль доступа и ограничения выполнения отд. команд
Выводы	**Атака на изменение выходных данных** (изменение выходных данных для нарушения работы модели)	**Output Manipulation Attack**	Контроль за выходными данными. Проверка на адекватность и соответствие. Мониторинг и контроль за искажением
Передача информ.	**Атака на передачу конфиденциальной информации**	**Data Leakage Attack**	Контроль доступа и шифрование. Мониторинг утечек и аудит использования
Выборка данных	**Атака на выборку данных** с нежелательной информацией	**Data Selection Attack**	Обнаружение аномалий, проверка данных на адекватность и соответствие ожиданию. Контроль за исходными данными

ВЕКТОРЫ АТАК. RAG-агенты



	ВЕКТОР АТАКИ		СПОСОБ ЗАЩИТЫ
Интерфейс	**Атака на поведение модели**	**Prompt Injection**	Контроль доступа и ограничения выполнения отд. Команд. Управление уязвимостями Санитизация и валидация входных данных
	Атака на приложение	**Denial of Service (DoS)**	Rate Limiting, использование WAF, мониторинг потребления ресурсов
	Атака на передачу информации (вывод информации о модели, в тч архитектура, параметры обучения)	**Output Data Attack**	Валидация и фильтрация выходных данных. Контроль доступа и ограничения выполнения отд. команд
Исполнение и инструменты	**Нанесение ущерба легитимными инструментами**	**Tool Misuse**	Контроль доступа и ограничения выполнения отд. Команд. Принцип минимальных полномочий
	Внедрение вредоносных инструкций в описание (метаданные) самого инструмента	**Tool Poisoning**	Аудит и контроль целостности описаний инструментов, разделение данных и инструкций в архитектуре
	Использование избыточных прав	**Privilege Escalation**	Использование изолированных контуров, SAST
	Эксплуатация уязвимостей в механизме проверки прав	**Auth Bypass & Impersonation**	Аутентификация и авторизация, аудит вызовов, использование второго фактора
Инфраструктура и оркестрация	**Атака через цепочку поставок** (внешние источники данных, компоненты системы)	**Supply Chain Attacks**	Использование доверенных репозиториев, SCA- и SAST-сканирование, SBOM, верификация цифровых подписей моделей
	Атака через цепочку агентов	**Cross Agent Communication Poisoning**	Изоляция агентов, валидация и санитизация данных на входе в каждом агенте, мониторинг межагентских коммуникаций
	Неконтролируемое потребление вычислительных ресурсов, токенов, вызовов API	**Ресурсоёмкие атаки**	Установка квот и лимитов на использование ресурсов для каждого агента/пользователя, Circuit Breakers
Выборка данных и знания	**Внедрение вредоносных или ложных данных** в документы, используемые RAG-системой	**Knowledge Base Poisoning**	Контроль доступа к базе знаний, версионирование данных, использование доверенных источников, криптографическая проверка целостности
	Извлечение конфиденциальных данных	**Sensitive Information Disclosure**	Маскирование данных перед передачей в модель, RBAC для RAG, Fine-tuning для «забывания» данных
	Манипуляция поиском в RAG	**Retrieval Manipulation**	Использование гибридного поиска (векторный + ключевой), внедрение модуля Reranker для повторной оценки релевантности
	Восстановление конфиденциального текста из его векторного представления	**Embedding Inversion**	Differential Privacy при создании эмбеддингов, гранулярный контроль доступа к векторной базе, обнаружение аномальных запросов к базе

ПРИМЕР ПРИМЕНЕНИЯ ФРЕЙМВОРКА НА РЕАЛЬНЫХ ПРОДУКТАХ

MOEX GROUP

AI NEWSFEED ИНТЕГРАТОР НОВОСТЕЙ И КОРПОРАТИЫНЫХ СОБЫТИЙ

Описание

Backend сервис для анализа тональности и важности новостей.

Интеграция потока новостей и событий и представление их в читаемом, понятном виде:
- *поток новостей по эмитентам;*
- *корпоративные действия;*
- *финансовая отчетность;*
- *мнения аналитиков.*

Потребители

Использование внутренними подразделениями Мосбиржи (Compliance, НРД, Рынки, ФинУслуги)

Данные /все из открытых источников/

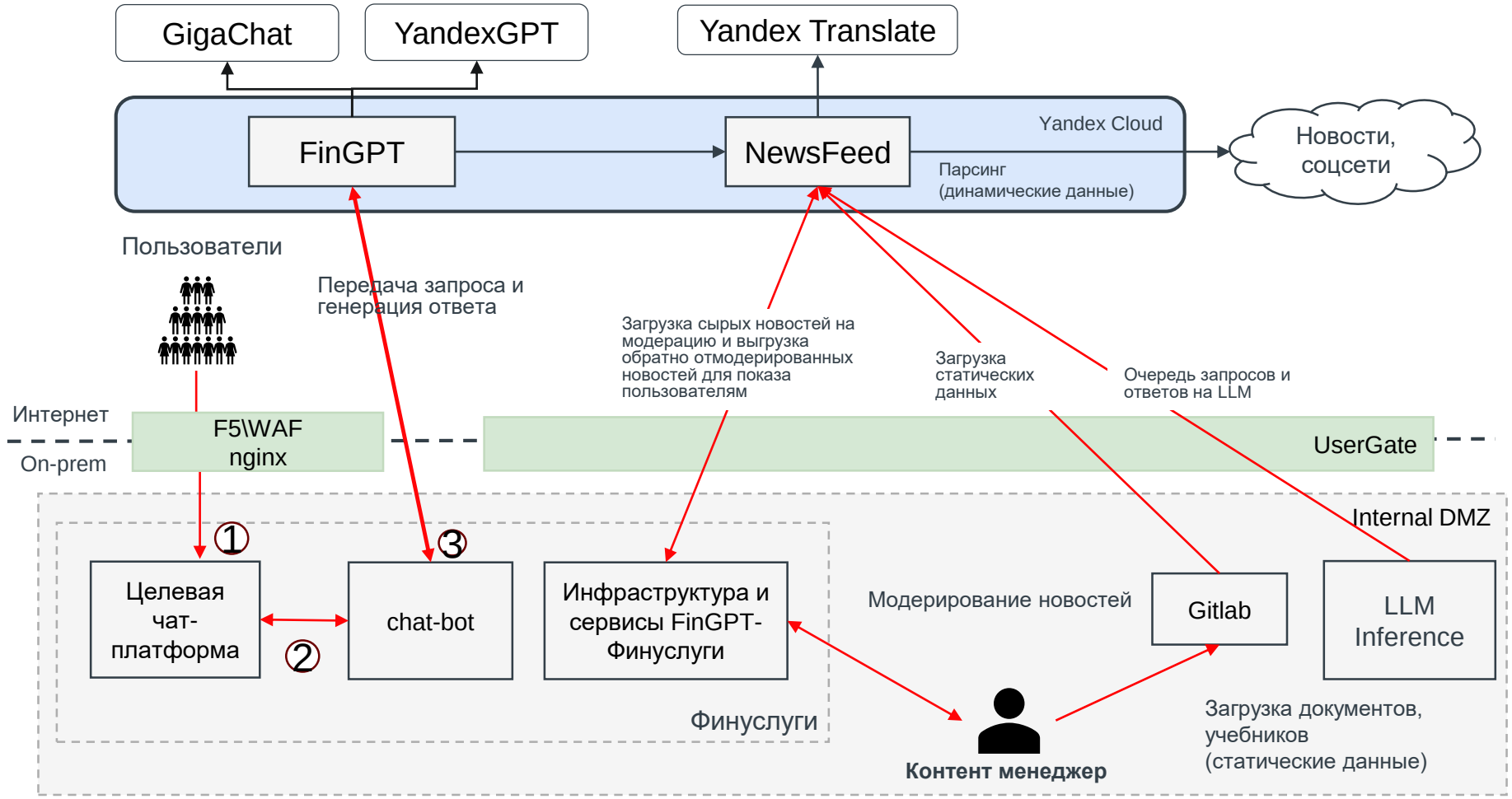
- *новости;*
- *социальные сети;*
- *аналитические обзоры;*
- *открытая корпоративная отчетность.*

РЕЗУЛЬТАТ МОДЕЛИ

Уровень управляемости HITL

Позволяет выгрузить обработанные ИИ новости с указанием:

1. Актуальной ссылки на сайт, с которого подтягивается новость
2. Сантимента (характер новости – негативный, нейтральный, позитивный)
3. Категории новости (экономическая, политическая, прочая)
4. Наименование компании и страны контрагента
5. Даты публикации новости



Модерация информации

- *Статические данные (документы, учебники) для RAG модели модерируются on-Prem и передаются вместе с кодом (git)*
- *Динамические данные(новости) модерируются также on-Prem и передаются в RAG через инфраструктуру Финуслуг*
- *Не прошедшие модерацию данные не участвуют в формировании ответов для пользователя*

ПРИМЕР ПРИМЕНЕНИЯ ФРЕЙМВОРКА НА РЕАЛЬНЫХ ПРОДУКТАХ

MOEX GROUP

AI NEWSFEED ИНТЕГРАТОР НОВОСТЕЙ И КОРПОРАТИВНЫХ СОБЫТИЙ

УРОВЕНЬ РИСКА СИСТЕМЫ ИИ

КРИТИЧНОСТЬ СИСТЕМЫ

Критичность данных - низкая

Функциональная критичность - выше среднего

Потенциальный ущерб - минимальный

Объем воздействия - минимальный

ОБЪЕМ ПОКРЫТИЯ СИСТЕМЫ

Объяснимость решений – ниже среднего

Чистота данных - средняя

Точность модели - средняя

События модельного риска - нет

ЗАЩИЩЕННОСТЬ СИСТЕМЫ

Зависимость от внешних факторов - высокая

Уровень контроля – ниже среднего

Уровень компетенций - средний

События инфоугроз - нет

Ключевые риски системы ИИ AI NEWSFEED

Описание риска	Присущий уровень риска	Мероприятия по минимизации рисков	Остаточный уровень риска	Плановый уровень риска
РИСК ТРЕТЬИХ СТОРОН Стратегическая зависимость от одного поставщика ИИ-услуг • Техническая (GigaChat, YandexGPT) • Зависимость от данных • Сервисная (Yandex облако) • Зависимость от «личности» модели (LLM)		1. Пилотирование продукта на некритичных процессах. 2. Мониторинг и управление рисками критичных поставщиков. 3. Установка SLA для сторонних сервисов.		
РИСКИ ЦЕПОЧКИ ПОСТАВОК Эксплуатация уязвимостей в облачной инфраструктуре/ сторонних агентах, используемых системой, может привести к утечке данных, нарушению бизнес-процессов		1. CI/CD процессы в разработке (sast/dast/sca, код ревью), 2. Использование ПО из внутренних репозитариев 3. Формирование и мониторинг White List источников. 4. Проведение анализа защищенности 5. Использование средств сетевой защиты при выгрузке данных и отправке запросов 6. Мониторинг ИБ (SOC, SIEM) 7. аутентификация внутренняя через idp		
НЕКОРРЕКТНАЯ ОБРАБОТКА ВЫХОДНЫХ ДАННЫХ Отсутствие проверок и обработки вывода может привести к уязвимостям в приложениях. Контент, который создает модель, не проверяют и не очищают, прежде чем передавать в другие системы.		1. Формирование и мониторинг White List источников. 2. Модерация информации 3. Использование средств сетевой защиты при выгрузке данных и отправке запросов		
ОТРАВЛЕНИЕ ДАННЫХ «отравленных» данных, когда злоумышленник загрязняет общедоступные датасеты, чтобы создать скрытую уязвимость при обучении.				
ВВЕДЕНИЕ В ЗАБЛУЖДЕНИЕ Генерация ложной или вводящей в заблуждение информации с потенциальными последствиями.		1. Ограничение функциональности – использование результатов работы модели в некритичных процессах с последующей перепроверкой выходных данных (выборочная перепроверка источников новостей)		
ЧРЕЗМЕРНАЯ АГЕНТНОСТЬ Чрезмерное использование LLM для принятия решений, выходящее за пределы безопасных границ.				