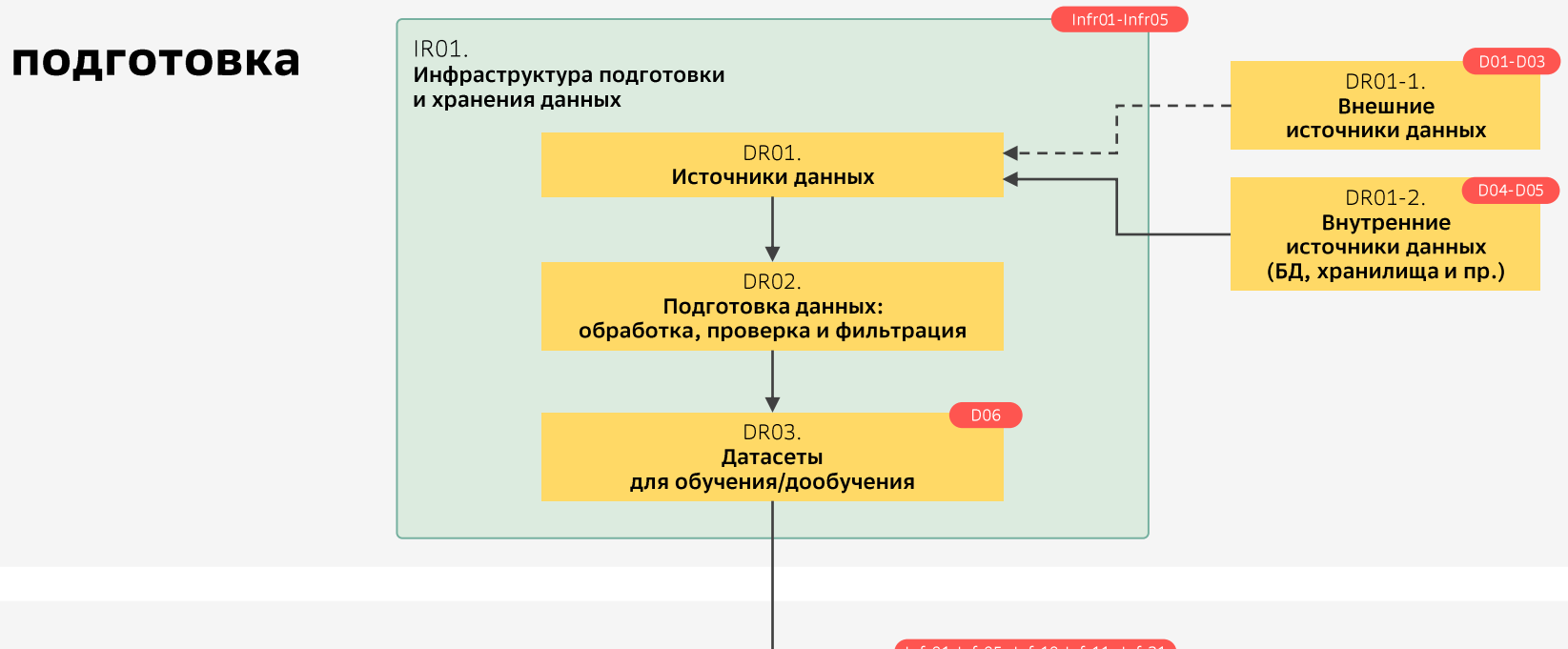
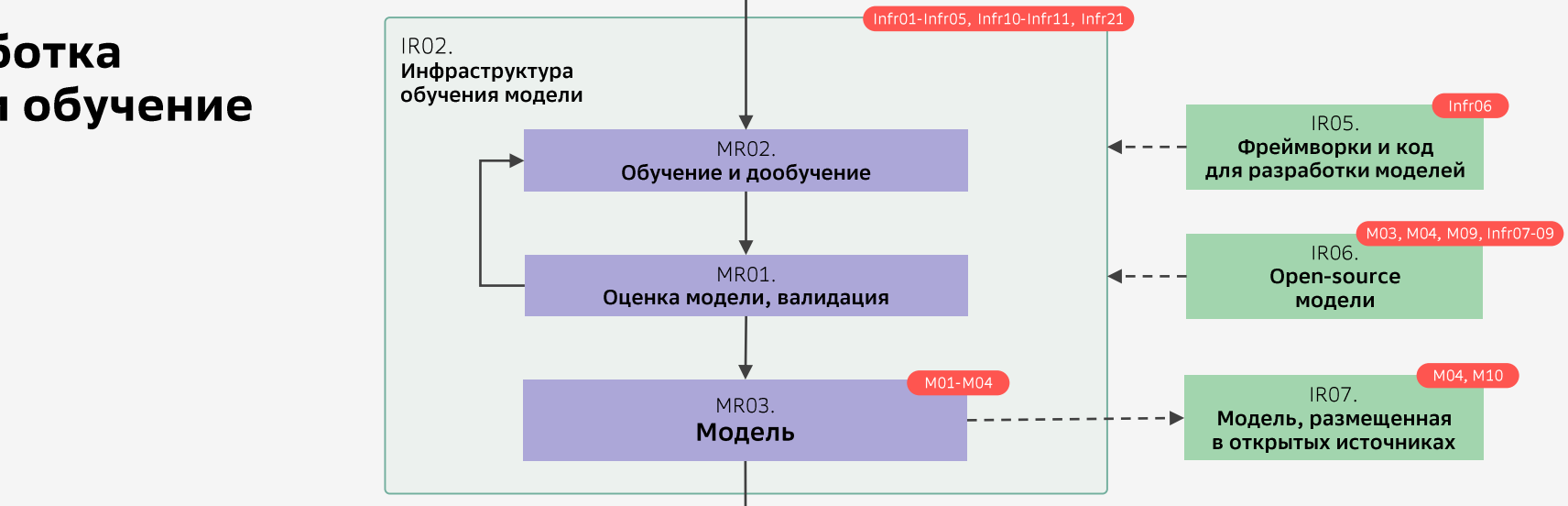
Обобщенная схема объекта защиты и актуальных угроз для КБ AI на этапах сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с приложениями

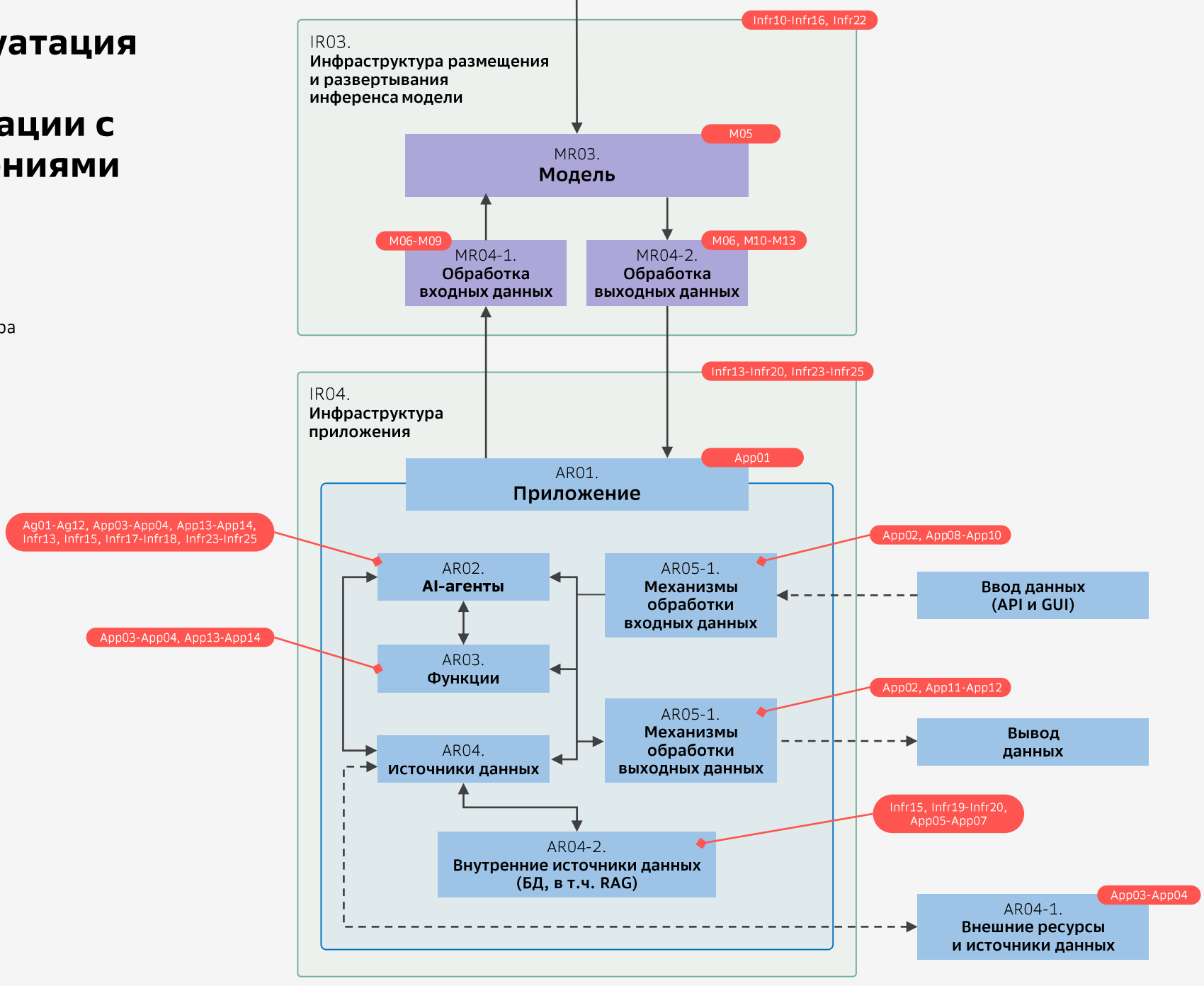
1. Сбор и подготовка данных



2. Разработка модели и обучение



3. Эксплуатация модели и интеграции с приложениями



Легенда

- Данные
- Инфраструктура исполнения
- Модель
- Приложение
- Угрозы

Перечень угроз для КБ AI на этапах сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с приложениями

Угрозы

D01	Использование для обучения/дообучения модели отравленных данных или датасетов, загруженных из внешних источников
D02	Использование для обучения/дообучения модели модифицированных данных или датасетов, загруженных из внешних источников
D03	Воспроизведение в ответах модели персональных данных (ПДн), полученных из внешних источников
D04	Использование для обучения/дообучения модели отравленных данных или датасетов, загруженных из внутренних источников
D05	Неконтролируемая загрузка данных, содержащих конфиденциальную информацию, в датасеты для обучения моделей
D06	Неконтролируемое использование, модификация, удаление данных для обучения или дообучения модели
Infr01	Несанкционированная модификация реестра источников данных, датасетов
Infr02	Несанкционированная модификация обучающих данных
Infr03	Небезопасная передача данных/датасетов между этапами подготовки
Infr04	Кража обучающих данных
Infr05	Утечка конфиденциальной информации из наборов обучающих данных
Infr06	Использование уязвимых версий сторонних библиотек, использование программного кода с закладками
Infr07	Использование open-source моделей, содержащих программные закладки в файлах
Infr08	Использование open-source моделей, содержащих логические закладки, заложенные при её обучении
Infr09	Использование open-source моделей, содержащих закладки в её весах open-source
Infr10	Подмена или модификация модели
Infr11	Кража модели
Infr12	Нарушение доступности модели
Infr13	Утечки конфиденциальной информации из систем логирования, в том числе логирования запросов и вызовов функций
Infr14	Невозможность или несвоевременное выявление, реагирование и расследования событий безопасности и инцидентов из-за отсутствия логирования взаимодействий
Infr15	Перехват или подмена запросов или ответов модели или данных передаваемых при взаимодействии с БД RAG (MiTM)
Infr16	Несанкционированное отключение или модификация механизмов фильтрации или контроля входных и выходных данных
Infr17	Хищение системного промпта
Infr18	Несанкционированная модификация системного промпта
Infr19	Несанкционированная модификация данных во внутренних источниках данных (в тч в БД RAG)
Infr20	Утечки информации из внутренних источников данных
Infr21	Несанкционированная модификация тестовых и валидационных датасетов
Infr22	Несанкционированные подключения к модели
Infr23	Утечки данных AI-агента или информации об особенностях его реализации
Infr24	Несанкционированная модификация AI-агента
Infr25	Утечка информации об архитектуре мультиагентной системы (MAC) через интерфейсы инструментов разработки или взаимодействия пользователя с AI-агентом или MAC
M01	Невозможность реагирование и расследования событий безопасности и инцидентов из-за отсутствия информации о данных, на которых выполнено обучение модели
M02	Использование модели с высокой уязвимостью к состязательным атакам (в том числе промпт-атакам)
M03	Нежелательное поведение, вредоносные генерации, галлюцинации
M04	Подбор атак с использованием знаний уровня white-box об open-source модели
M05	Отсутствие информации об инференсах модели
M06	Обход механизмов обработки входных/выходных данных, реализуемых на уровне модели
M07	Нарушение доступности модели (DoS) из-за отсутствия единого контроля запросов на уровне модели
M08	Исчерпание лимитов интеграции (DoW) из-за отсутствия единого контроля запросов на уровне модели
M09	Обход встроенных защитных механизмов модели в том числе с использованием методов состязательных атак и промпт-атак
M10	Утечка информации о модели
M11	Утечки конфиденциальной информации из дообученной модели или LoRA
M12	Эксфильтрация, инверсия или реверс-инжиниринг модели
M13	Эксфильтрация данных
App01	Использование небезопасных интеграций компонент
App02	Обход механизмов обработки входных/выходных данных, реализуемых на уровне приложения
App03	Загрузка вредоносного программного обеспечения (ВПО) из внешних источников (Интернет)
App04	Загрузка отравленных данных из внешних источников (Интернет)
App05	Внедрение непрямых промпт-инъекций во внутренние источники (в т.ч. БД RAG)
App06	Утечки информации из внутренних источников (в т.ч. БД RAG)
App07	Выполнение вредоносных инструкций, созданных моделью
App08	Реализация прямых промпт-инъекций из-за отсутствия контроля входных данных
App09	Нарушение доступности (DoS/DoW) интеграции
App10	Нарушение логики выполнения задачи из-за отсутствия контроля входных данных
App11	Утечка информации о системном промпте из-за некорректной обработки выходных данных
App12	Токсичная или вредоносная генерация из-за некорректной обработки выходных данных
App13	Вывод информации о среде
App14	Автоматического распространения вредоносной инструкции на другие приложения
Ag01	Ошибки в проектировании AI-агентов и MAC
Ag02	Вредоносные генерации в ответе AI-агента на запрос пользователя
Ag03	Отправка информации из среды исполнения функций AI-агента (действий) на внешние ресурсы
Ag04	Удаление или модификация файлов в среде исполнения функций AI-агента (действий)
Ag05	Размещение в среде исполнения функций AI-агента (действий) файлов с ВПО, полученных с внешних ресурсов
Ag06	Нарушение доступности (DoS/DoW) среды исполнения AI-агента (в т.ч. функций)
Ag07	Утечка информации об архитектуре MAC через интерфейсы пользовательского ввода
Ag08	Передача другому AI-агенту ложной информации в MAC
Ag09	Нарушение цели другого AI-агента при кооперативном взаимодействии в MAC
Ag10	Распространение промпт-атаки по AI-агентам в MAC для усиления ее эффекта
Ag11	Нарушение рабочего процесса АС, реализующей AI-агента
Ag12	Утечка информации о цели, функциях, содержимом памяти или инструкциях механизма планирования AI-агента