# ЛЕКЦИЯ 14

# AI-АГЕНТЫ И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

Основные понятия • определения • схемы • таблицы • графики • примеры

**Преподаватель: проф.Ахмедиярова А.Т.**

# Структура и основные разделы


01 Понятие AI-агента

02 От языковой модели к агенту

03 Основные компоненты AI-агента

04 Цикл работы агента

05 Agentic AI

06 Многошаговые задачи

07 Планирование

08 Инструменты агента

09 Память и состояние

10 Обратная связь

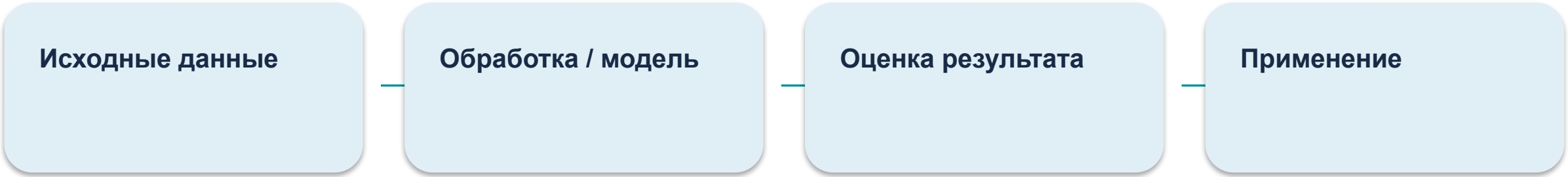
11 Условия завершения

12 Human-in-the-loop

# Понятие AI-агента

AI-агент — программная интеллектуальная система, которая получает информацию о задаче или среде, принимает решения и выполняет действия для достижения определённой цели. В отличие от модели, которая только формирует ответ, агент связывает интеллектуальную обработку с последовательностью разрешённых действий.

# Схема: Понятие AI-агента




Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

# От языковой модели к агенту

Обычная LLM получает контекст и генерирует ответ. Агентная система дополнительно может хранить состояние задачи, выбирать следующий шаг, обращаться к разрешённым инструментам, анализировать результат действия и продолжать работу до достижения заданного условия завершения.

# Сравнительная таблица: От языковой модели к агенту

| От языковой модели к агенту | |
|---|---|
| | Обычная LLM получает контекст и генерирует ответ. Агентная система дополнительно может хранить состояние задачи, вы |

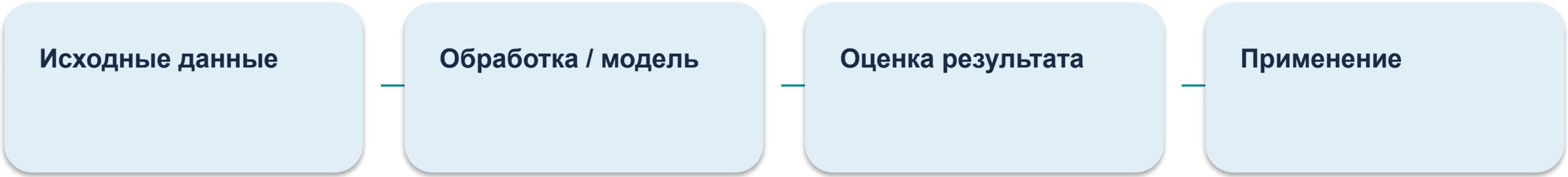
# Основные компоненты AI-агента

Функционально агент включает цель, наблюдения, модель принятия решений, память или состояние, набор инструментов, механизм выполнения действий и контроль результата.

# Цикл работы агента

Типичный цикл можно представить как последовательность: цель → наблюдение → план → действие → результат → проверка → следующий шаг. Цикл прекращается после достижения цели, возникновения ошибки, превышения лимита или необходимости решения человека.

# Схема: Цикл работы агента




Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

# Agentic AI

Agentic AI — подход к системам с большей автономностью при планировании и выполнении многошаговых задач. В загруженных материалах пример представлен как процесс: получить данные → обработать → проанализировать → сформировать таблицу → подготовить документ.

# Многошаговые задачи

Многошаговая задача состоит из взаимосвязанных операций, где результат предыдущего этапа становится входом следующего. Например, подготовка отчёта может включать получение данных, очистку, расчёт метрик, интерпретацию, проверку и оформление.

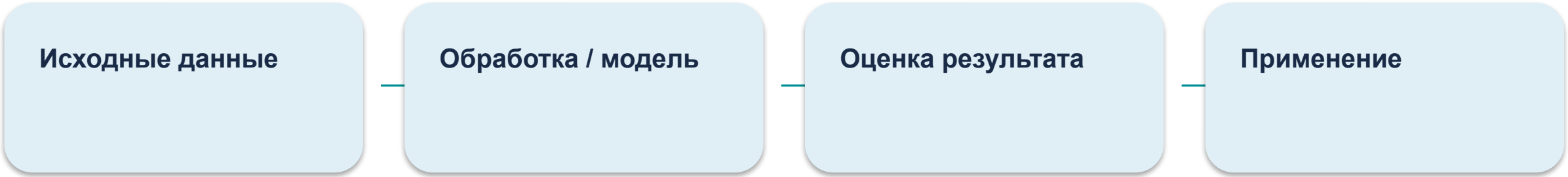
# Сравнительная таблица: Многошаговые задачи

| | |
|---|---|
| **Многошаговые задачи** | Многошаговая задача состоит из взаимосвязанных операций, где результат предыдущего этапа становится входом следующе |

# Планирование

Планирование преобразует общую цель в проверяемые действия. Чем выше риск задачи, тем важнее заранее определить допустимые шаги, ограничения и условия остановки.

# Схема: Планирование




Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

# Инструменты агента

Инструмент — разрешённая функция или сервис, который агент вызывает для выполнения операции: вычисления, чтения файла, поиска по документам, обращения к информационной системе или формирования результата. Модель не должна иметь неограниченный доступ к инструментам.

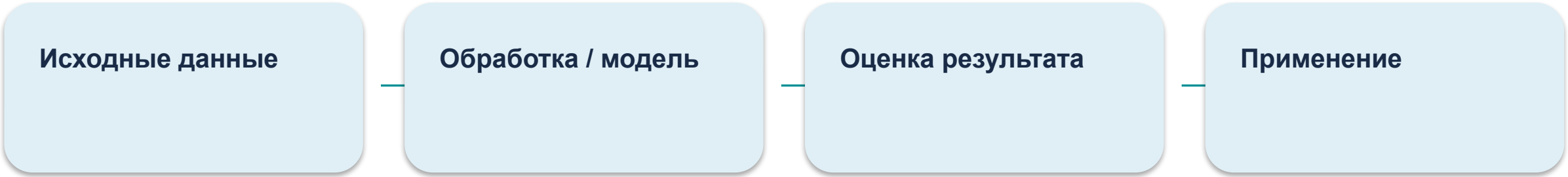
# Память и состояние

Состояние отражает уже выполненные шаги, промежуточные результаты, ошибки и ограничения. Память позволяет сохранять информацию между этапами, но требует правил доступа, хранения и удаления данных.

# Обратная связь

После каждого действия агент должен оценивать результат. Если данные отсутствуют, инструмент вернул ошибку или результат противоречит ограничениям, план необходимо скорректировать либо передать задачу человеку.

# Схема: Обратная связь




Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

# Сравнительная таблица: Обратная связь

| | |
|---|---|
| **Обратная связь** | После каждого действия агент должен оценивать результат. Если данные отсутствуют, инструмент вернул ошибку или резу |

# Условия завершения

Без явных условий остановки агент может выполнять лишние действия. Условиями завершения могут быть достижение цели, исчерпание лимита шагов, отсутствие данных, ошибка безопасности или необходимость подтверждения человеком.

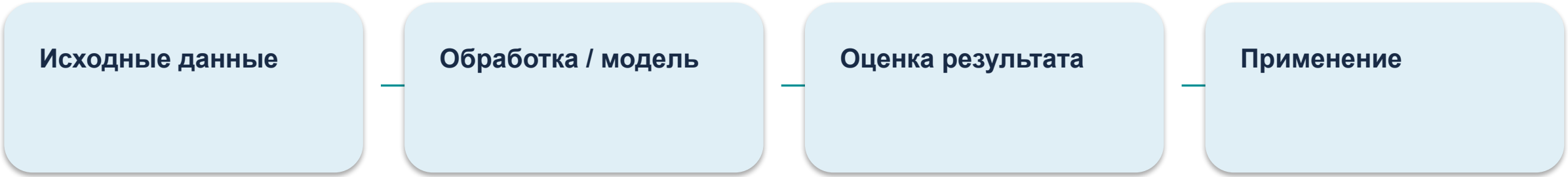
# Human-in-the-loop

Human-in-the-loop означает включение человека в критические точки процесса. Особенно важен контроль действий с финансовыми, правовыми, административными или информационно-безопасностными последствиями.

# Многоагентные системы

Если несколько агентов взаимодействуют для общей задачи, формируется многоагентная система. В загруженных материалах приведён пример: один агент ищет и структурирует информацию, второй анализирует данные, третий проверяет результат.

# Схема: Многоагентные системы




Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

# Координация агентов

При централизованной архитектуре координатор распределяет задачи между специализированными агентами. При распределённой схеме агенты взаимодействуют напрямую. Централизованный вариант обычно проще контролировать и журналировать.

# Сравнительная таблица: Координация агентов

| Координация агентов | |
| --- | --- |
| | При централизованной архитектуре координатор распределяет задачи между специализированными агентами. При распределё |

# Коммуникация между агентами

Обмен должен иметь структурированный формат: задача, входные данные, ограничения, результат и статус. Это снижает риск потери контекста и противоречий.

# Автоматизация процессов

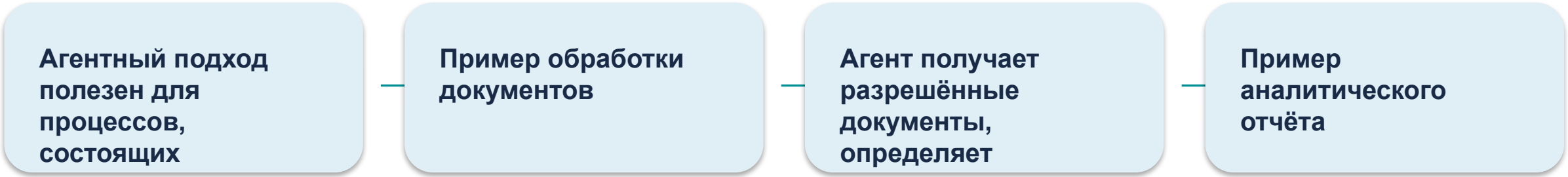
Агентный подход полезен для процессов, состоящих из повторяющихся интеллектуальных операций: получение данных, проверка, классификация, анализ, подготовка черновика и передача результата. Пример обработки документов Агент получает разрешённые документы, определяет тип, извлекает обязательные поля, проверяет полноту, формирует сводную таблицу и передаёт спорные случаи человеку.

Ключевая идея: Агентный подход полезен для процессов, состоящих из повторяющихся интеллектуальных операций: получение данных, проверка, клас

# Автоматизация процессов — продолжение

Пример аналитического отчёта Система получает таблицу, проверяет пропуски и типы данных, рассчитывает метрики, формирует сравнение, готовит черновой вывод и затем сверяет числа с исходными данными.

# Схема: Автоматизация процессов




Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

# AI-агенты в информационной безопасности

В загруженных материалах ИИ связывается с обнаружением аномалий, анализом сетевого трафика, вредоносного ПО, фишинга, журналов событий и поддержкой SOC. Агент может автоматизировать структурирование и подготовку сводок, но итоговые решения должны проверяться специалистом.

# Риски автономных действий

Ошибочный ответ обычной модели остаётся текстом, а ошибочный вывод агента может привести к действию. Поэтому с ростом автономности должны усиливаться ограничения прав, валидация, журналирование и подтверждение критических операций.

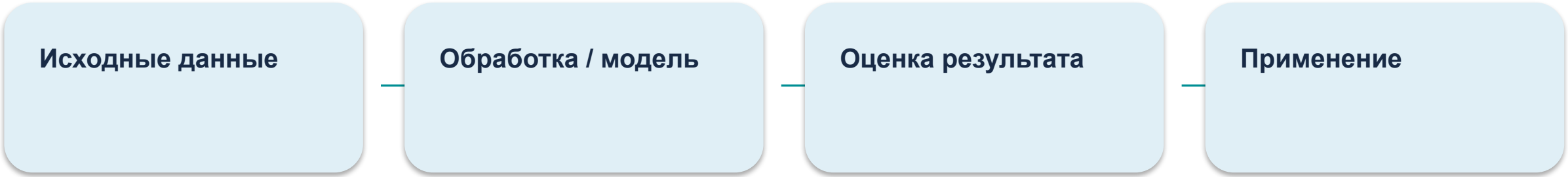
# Сравнительная таблица: Риски автономных действий

| Риски автономных действий | |
|---|---|
| | Ошибочный ответ обычной модели остаётся текстом, а ошибочный вывод агента может привести к действию. Поэтому с рост |

# Принцип минимальных привилегий

Агент получает только права, необходимые для конкретной задачи. Если требуется чтение файла, не нужно предоставлять право удаления; если нужен черновик сообщения, отправка должна подтверждаться отдельно.

# Схема: Принцип минимальных привилегий




Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

# Prompt Injection

Внешний документ или веб-контент может содержать инструкции, пытающиеся изменить поведение агента. Недоверенные данные следует отделять от доверенных системных инструкций, а доступ к инструментам ограничивать.

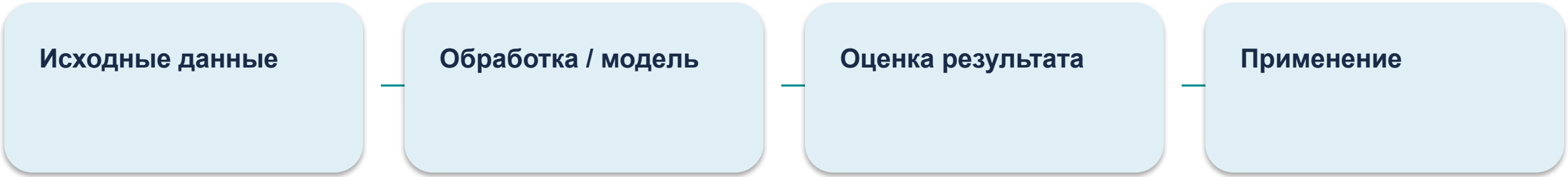
# Ошибки инструментов

Инструменты могут возвращать неполные данные, ошибки или неожиданный формат. Агент должен обнаруживать такие случаи и переходить к безопасному сценарию: повторить разрешённую операцию, запросить уточнение или остановиться.

# Наблюдаемость и аудит

Профессиональная агентная система должна журналировать цель, вызванные инструменты, использованные данные, результаты и подтверждения человека. Это необходимо для диагностики, безопасности и ответственности.

# Схема: Наблюдаемость и аудит




Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

# Сравнительная таблица: Наблюдаемость и аудит

| | |
|---|---|
| **Наблюдаемость и аудит** | Профессиональная агентная система должна журналировать цель, вызванные инструменты, использованные данные, результа |

# Оценка качества

Агент оценивается не только по качеству текста, но и по успешности завершения задачи, числу шагов, времени, стоимости, ошибкам инструментов, доле вмешательств человека и безопасным отказам.

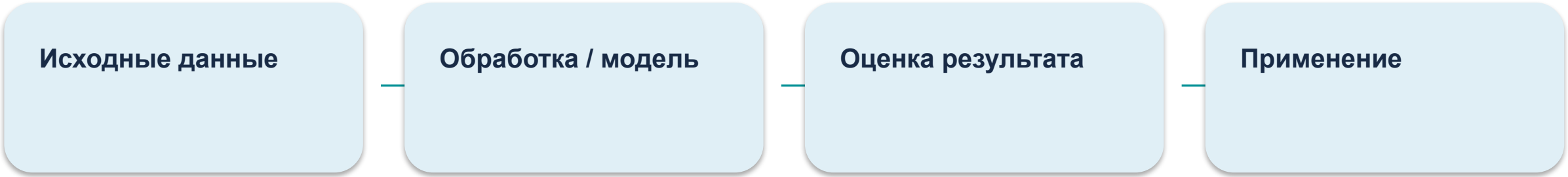
# Проектирование безопасного агента

Проектирование начинается с описания процесса: цель, разрешённые данные, набор действий, критические точки, условия остановки, права доступа и критерии качества. Только после этого выбираются модель и инструменты.

# Ограничения Agentic AI

Агентная система не гарантирует достижение цели. Ошибки модели, неполные данные, сбои инструментов и конфликтующие инструкции требуют безопасного механизма остановки и передачи задачи человеку.

# Схема: Ограничения Agentic AI




Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

# Основные выводы

AI-агент получает информацию, принимает решения и выполняет действия для достижения цели. Agentic AI ориентирован на автономное выполнение многошаговых задач в заданных ограничениях. Многоагентные системы распределяют функции между специализированными агентами. Безопасность требует минимальных привилегий, журналирования, условий остановки и Human-in-the-loop.

Ключевая идея: AI-агент получает информацию, принимает решения и выполняет действия для достижения цели.

# Основные выводы — продолжение

В информационной безопасности агенты могут автоматизировать рутинные аналитические процессы, но не заменяют контроль специалиста.

# Контрольные вопросы

1. Что такое AI-агент и чем он отличается от LLM? 2. Какие компоненты входят в цикл работы AI-агента? 3. Что понимается под Agentic AI? 4. Какую роль играют инструменты, память и состояние? 5. Зачем нужны условия завершения и Human-in-the-loop? 6. Что такое многоагентная система? 7. Чем централизованная координация отличается от распределённой? 8.

Ключевая идея: 1. Что такое AI-агент и чем он отличается от LLM?

# Контрольные вопросы — продолжение

Какие задачи могут автоматизировать AI-агенты в информационной безопасности? 9. Какие риски возникают при доступе агента к внешним инструментам? 10. Какие метрики и меры безопасности используются при оценке агентной системы?

# Схема: Контрольные вопросы

1. Что такое AI-агент и чем он отличается от LLM
2. Какие компоненты входят в цикл работы AI-аген
3. Что понимается под Agentic AI?
4. Какую роль играют инструменты, память и состо

Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

# Список источников

1. Бессмертный, И. А. Системы искусственного интеллекта: методическое пособие / И. А. Бессмертный, А. Е. Авдюшина, А. В. Кугаевских, Ю. А. Королева, Н. Г. Рущенко. – Санкт-Петербург: Университет ИТМО, 2024. – 281 с. 2. Соснило, А. И. Атлас искусственного интеллекта для бизнеса и власти: учебное пособие / А. И. Соснило. – Санкт-Петербург: Университет ИТМО, 2022. – 98 с. 3. Сидоркина, И. Г.

Ключевая идея: 1. Бессмертный, И. А. Системы искусственного интеллекта: методическое пособие / И. А. Бессмертный, А. Е. Авдюшина, А. В. Куга

# Список источников — продолжение

Системы искусственного интеллекта: учебное пособие / И. Г. Сидоркина. – Москва: КНОРУС, 2018. – 248 с. 4. Баланов, А. Н. Искусственный интеллект. Понимание, применение и перспективы: учебник для вузов / А. Н. Баланов. – Санкт-Петербург: Лань, 2024. – 312 с. 5. Калинин, М. О. Основы искусственного интеллекта. Безопасность искусственного интеллекта : учебное пособие / М. О. Калинин, В. М.

# Список источников — продолжение

Крундышев. – Санкт-Петербург, 2024. – 220 с. 6. Андреянов, Н. В. Методы искусственного интеллекта: учебное пособие / Н. В. Андреянов, Т. С. Евдокимова, А. Д. Павлов [и др.]. – Казань: КНИТУ-КАИ, 2024. – 392 с. 7. Основы искусственного интеллекта: учебное пособие. – Гл. 14: AI-агенты и интеллектуальные системы.

# График: иллюстрация зависимости




Иллюстративные значения (не результаты эксперимента).

# Итоги и вопросы для обсуждения

1. Объясните понятие «Понятие AI-агента» и приведите пример.
2. Объясните понятие «От языковой модели к агенту» и приведите пример.
3. Объясните понятие «Основные компоненты AI-агента» и приведите пример.
4. Объясните понятие «Цикл работы агента» и приведите пример.
5. Объясните понятие «Agentic AI» и приведите пример.
6. Объясните понятие «Многошаговые задачи» и приведите пример.
7. Объясните понятие «Планирование» и приведите пример.