МОДУЛЬ 5 · ТЕМА 5.1

Внутренние решения и инструменты за пределами разрешённых сервисов



Результат определяет не только модель, но и всё, что стоит вокруг неё

5.1

День 5 · Виталий Пиков · архитектор информационной безопасности · pikov.expert

Сильные зарубежные модели: снимок на 10.11.2026

Разработчик	Модель	Что умеет по описанию
OpenAI	GPT-6 Astra	Контекст 1,05 млн токенов, инструменты: поиск, файлы, работа с компьютером
Google	Gemini 4 Carbon	Контекст 1 млн токенов; текст, изображения, звук, видео, PDF
Anthropic	Claude Fable 5.2	Контекст 1 млн токенов; текст и изображения; инструменты; сложные рассуждения и долгая работа агентов

По официальным страницам разработчиков, открыты 10.11.2026. Дата выпуска указана только у Anthropic: 09.11.2026

Граница возможностей неровная

Однако на сложной управленческой задаче, специально выбранной за пределами этой границы, участники, работавшие с ИИ, на 19 % реже приходили к правильному решению, чем те, кто работал без ИИ, что указывает на возможные ограничения ИИ в поддержке работников умственного труда.

Dell'Acqua и соавт., 2026, doi.org/10.1287/orsc.2025.21838

19 %

реже приходили к правильному решению

Почему данные организации туда не отправляются

В списках стран, где OpenAI, Google и Anthropic открывают доступ к моделям, **России нет**

Работа вне списка, по правилам OpenAI, может привести к **блокировке учётной записи**

Данные уходят из-под контроля организации, под чужие правила и право

Куда можно отправлять сведения, решают **правила организации**

Что бывает внутри организации

- **Внутренний чат** на локальных моделях: текст не уходит наружу
- **Помощник по документам**: отвечает по базе документов со ссылками
- **Конструктор агентов**: сотрудник собирает помощника без программирования
- **Подключаемые инструменты**: поиск, расчёт, запрос в систему
- **Инфраструктура** для тех, кто программирует

Есть ли это у вас и как называется – уточните в своей организации

Покупать или делать самим

Зрелое покупают: видеоаналитика, контент, анализ отзывов

Сами делают то, где своя экспертиза выше рынка

И то, что нельзя выпускать наружу: рецепты, технологические карты

Для инженера это КД, ТД и результаты испытаний

Н. Галкин, «Кофемания», форум All-over-IP, 11.11.2026

Закрытые и открытые модели

Закрытые модели легко внедрить, но это риски для данных и зависимость от поставщика

Открытые модели на своём оборудовании дают больше настройки, но требуют ресурсов и квалификации

Модели, которые можно развернуть у себя, часто уступают лучшим закрытым

Запустить модель можно и на рабочей станции с одной видеокартой

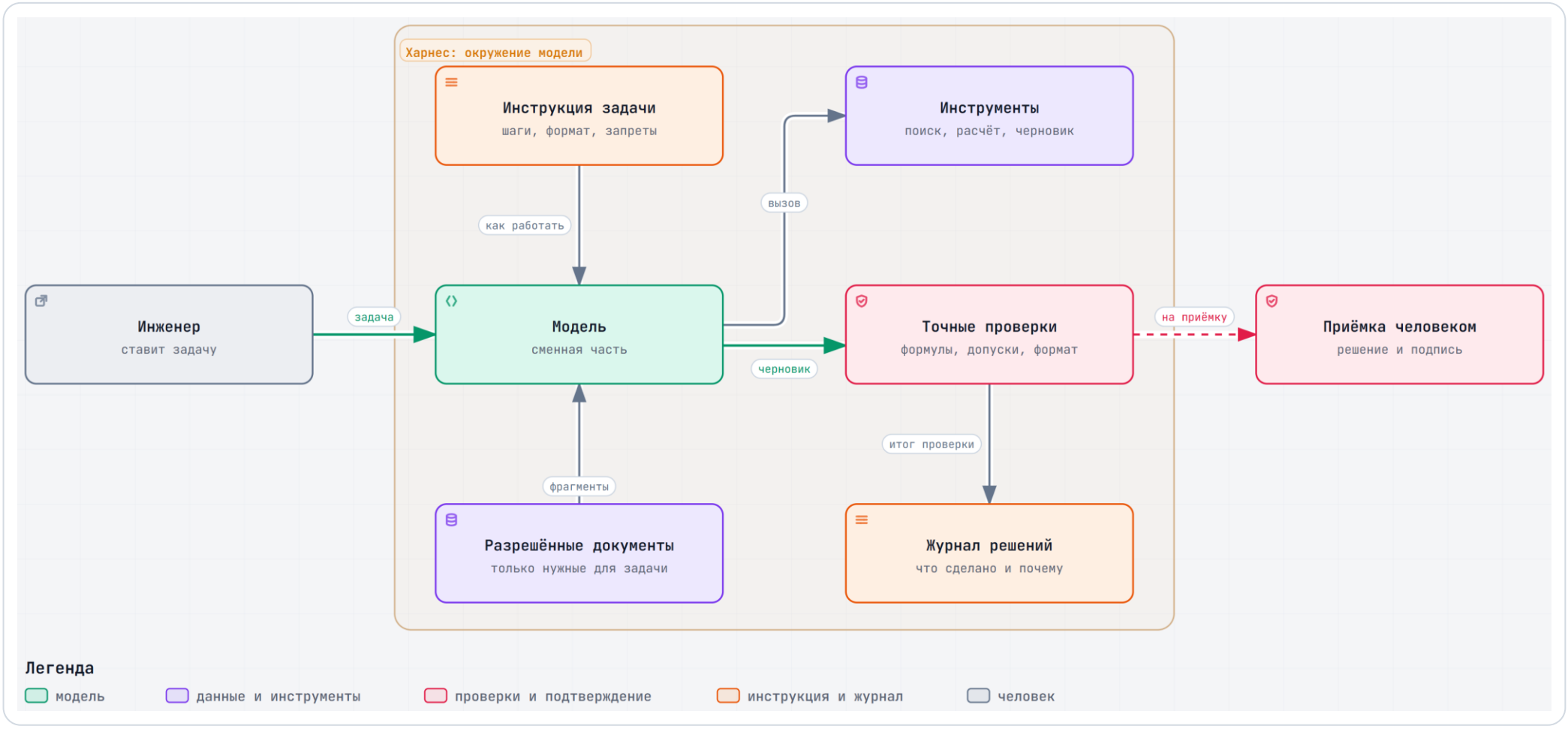
Модель и харнес



- Харнес – окружение модели: инструкция, разрешённые данные, инструменты, проверки, журнал, решение человека
- «Модель это повар. Харнес это кухня» – А. Субботин
- Инструкцию пишут так, чтобы справилась и более простая модель
- Сначала описать работу, потом звать агента

Схема «Модель и харнес»

Модель и харнес: что стоит вокруг модели



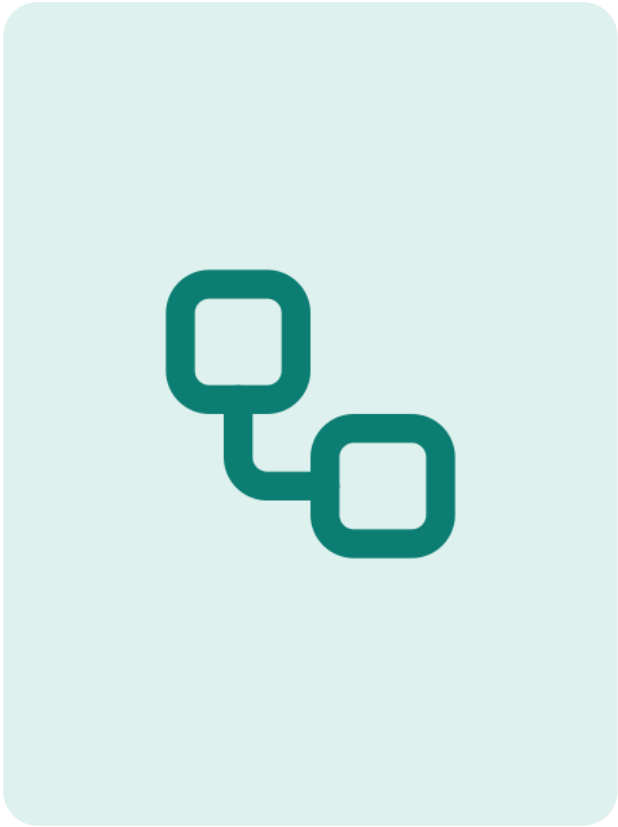
Интерактивная версия – 08-Схемы/d5-model-harness.html

Агент или фиксированная цепочка

Вариант	Когда подходит
Шаблон запросов	Задача укладывается в несколько одинаковых шагов
Цепочка в конструкторе	Шаги заранее заданы схемой, модель пишет текст в отмеченных местах
Агент	Порядок шагов заранее неизвестен, модель сама выбирает следующий шаг

Начинать с самого простого решения и усложнять, только когда это необходимо – руководство Anthropic

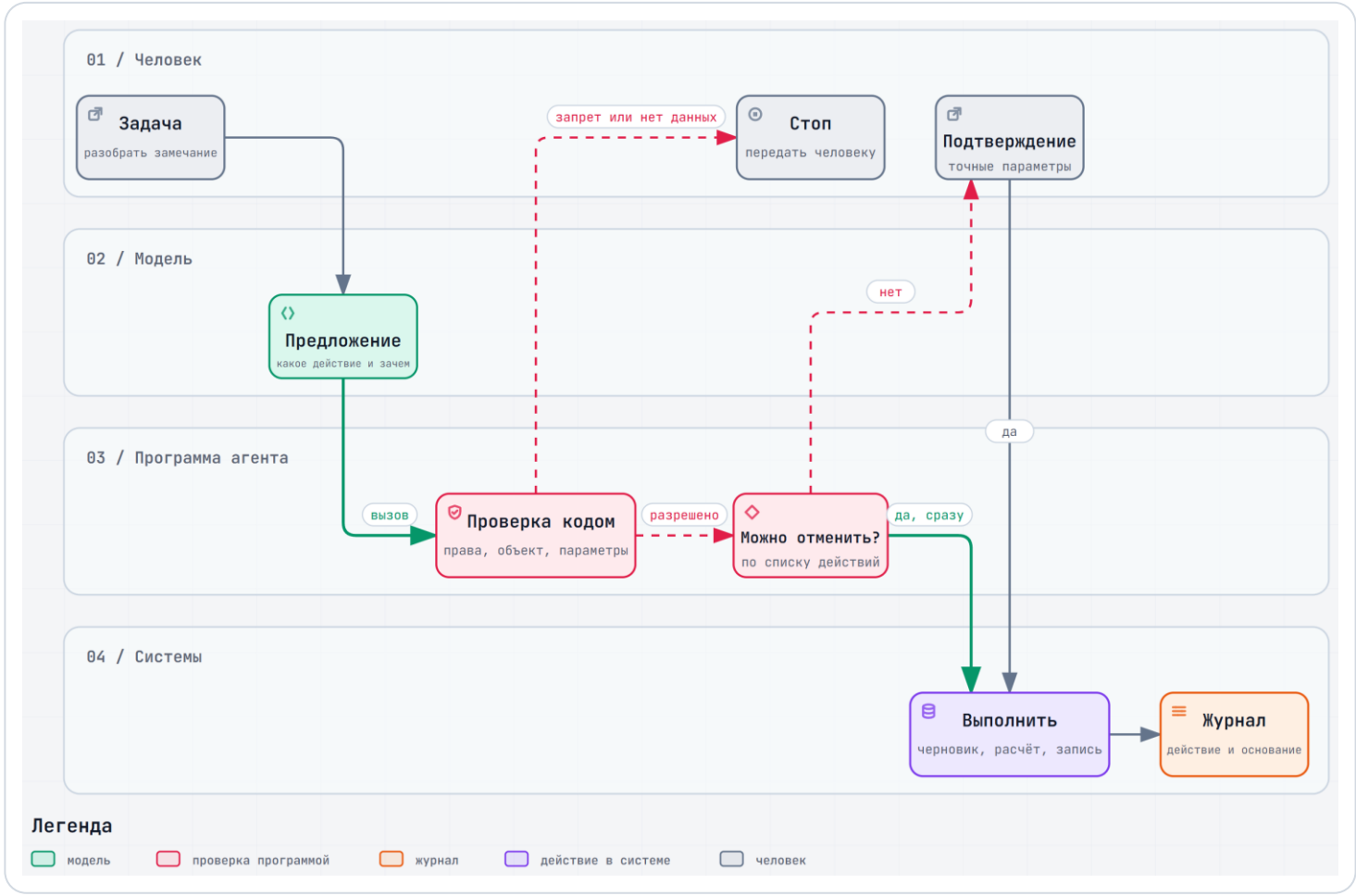
Агент-советчик



- **Модель** разбирает текст и предлагает следующий шаг
- **Программа** проверяет права, объект и параметры действия
- **Человек** подтверждает то, что нельзя отменить
- Обратимое – сразу, необратимое – со спросом

Схема «Агент-советчик»

Агент-советчик: модель предлагает, программа проверяет, человек решает



Интерактивная версия – 08-Схемы/d5-agent-sovetchik.html

Права проверяет код, а не модель

Модель Qwen2.5 без внешних ограничений распознала повышенный риск, но не передала на обязательное согласование человеку ни одного из 43 чувствительных случаев: доля случаев, переданных человеку, оказалась равна нулю.

Pacella и соавт., 2026, doi.org/10.3390/a19080627

ни одного

из 43 чувствительных случаев

Модель-«судья»: дешёвая проверка шага



1. Закрытый вопрос: пропустить, заблокировать или передать человеку
2. Ответ – вариант и уверенность за доли секунды
3. Порог задаёт политика: чем дороже ошибка, тем выше
4. «Гарантирован формат ответа, а не его правота»
5. Правило курса прежнее: права – программа, необратимое – человек

Слайды А. Мирошкина, форум All-over-IP, 11.11.2026; числа на них иллюстративные

Что оставить точным проверкам

Сравнение 85 мА с 40 мА – формула

Расчёт разряда батареи – таблица

Комплектность документации – сверка по перечню

Стоп-условия: нет данных, противоречие, подозрительная строка

Подтверждение тоже нужно уметь делать



- Если подтверждать приходится слишком часто, «да» нажимают не читая
- Люди оценивают, правильно ли сделано, а не безопасно ли
- Подтверждение – там, где необратимо
- Человек видит **точные параметры**, а не пересказ агента

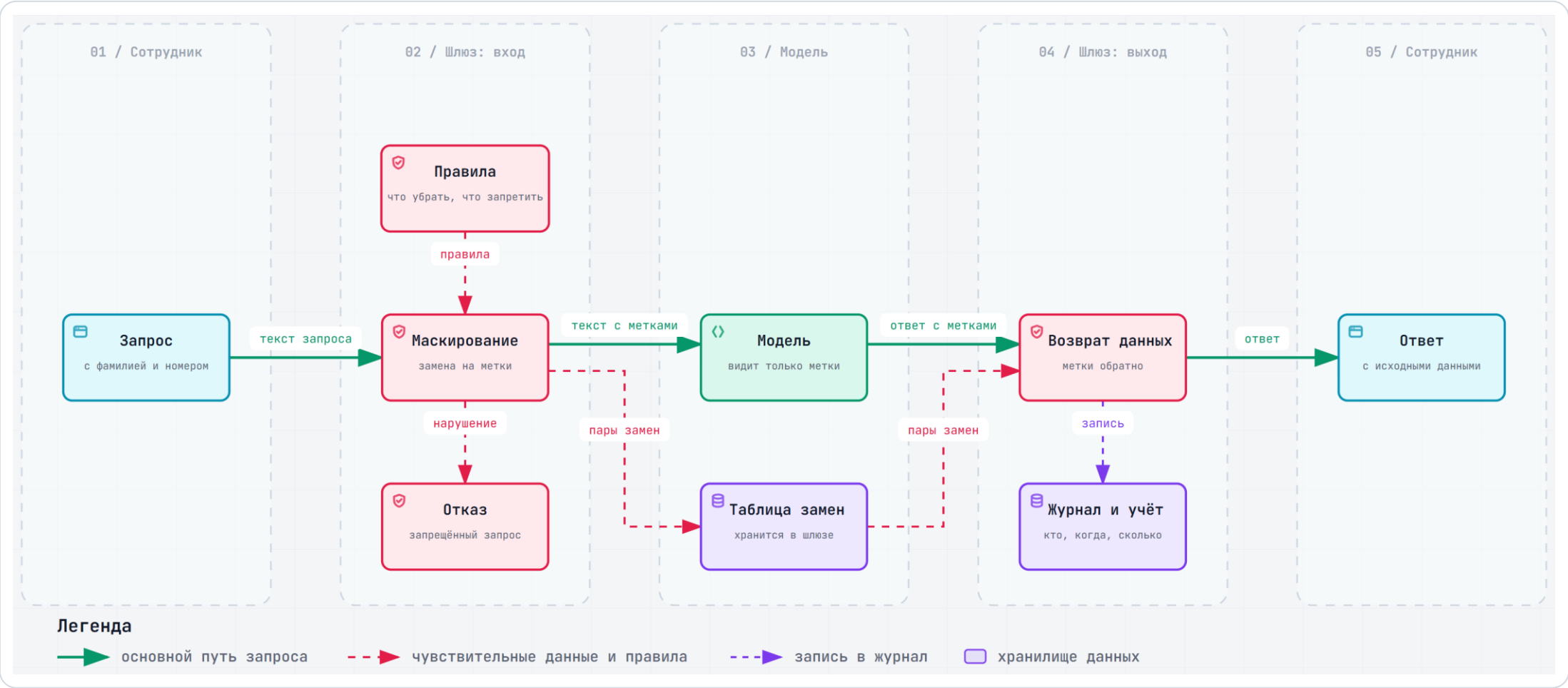
Шлюз безопасности ИИ



1. Находит фамилии, номера, ключи и заменяет метками
2. Модель видит только метки
3. В ответе возвращает исходные значения
4. Ведёт журнал и считает расход

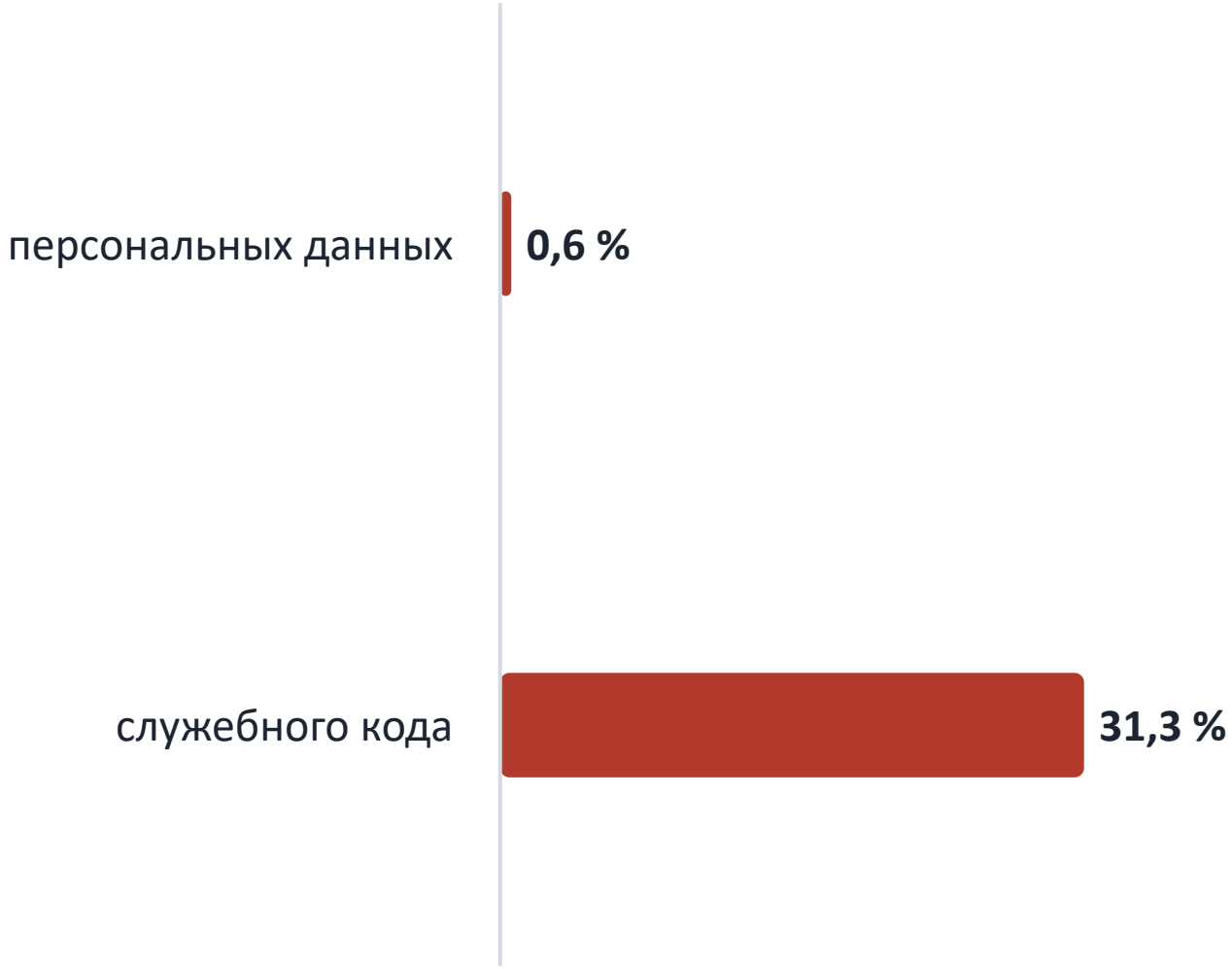
Схема «Шлюз безопасности ИИ»

Шлюз безопасности ИИ: убрать данные до модели и вернуть после



Интерактивная версия – 08-Схемы/d5-shlyuz.html

Чего шлюз не делает



- Не делает чертёж или протокол открытым документом
- В исследовании 2026 года лучшее сочетание мер: 0,6 % утечки персональных данных, 31,3 % – служебного кода
- Восемь защит от скрытых команд обошли подобранными атаками
- Обезличивание не даёт права на передачу

Направления развития – кратко

Агенты выполняют больше шагов подряд без человека

Модели приходят в производство: документация цеха, диагностика, техпроцессы

Открытые модели ненамного отстают от закрытых в отдельных задачах

Процессы придётся перестраивать – позиция К. Овчинникова

Задание 1. Одна задача в разных инструментах



35 мин

1. Задача: перечень проверок по замечанию 3-2 (ТЗ – редакция 1)
2. Один и тот же текст запроса по шаблону – в двух разрешённых сервисах
3. Второй сервис не открылся – отправить полный текст запроса преподавателя и сравнить с его ответом
4. Тот же вопрос без данных – в одном из сервисов
5. Сравнить по таблице критериев, записать сервис, модель, дату

Задание 2. Выбор инструмента для задач



60 мин

1. 2а: учебный перечень из 12 задач, закрыть три последних столбца
2. Пройти схему выбора инструмента по порядку вопросов
3. 2б: свои 5–8 задач, маршрут данных, инструмент, проверка
4. Отметить задачу для итоговой работы

Задание 3. Агент-советчик: кто что делает



30 мин

1. Разложить шаги разбора 3-2: модель, точная проверка, человек
2. Отметить необратимые шаги и кто их подтверждает
3. Записать три стоп-условия

Задание повышенной сложности. Модель-«судья» для шагов агента



30–45 мин

1. Для тех, кто справился с основными заданиями. Можно сделать и после занятия
2. Девять учебных действий агента
3. Пропустить, заблокировать или передать человеку
4. Сверка с правилом курса; убираем факты – меняется ли решение
5. Результат: что поручить судье, программе, инженеру

Что запомнить

- Сильные зарубежные сервисы – не для данных организации
- Харнес важен не меньше модели
- Модель предлагает, программа проверяет, человек решает
- Шлюз помогает, но не делает служебное открытым

Автор курса – Виталий Пиков



Регалии

Архитектор информационной безопасности · РБПО · DevSecOps · AppSec

Лектор и автор образовательных программ по ИБ и РБПО

Телефон

+7 (963) 782-92-78

E-mail

vitaly@pikov.expert

Сайт

https://pikov.expert

Telegram

@UnderLineSecurity