

Взаимодействие с иностранными ИИ-агентами: риски и защита

Системный подход к безопасности данных в
условиях трансграничного регулирования

Тишаков Сергей ЗАО РОССИЙСКАЯ ОЦЕНКА





RU Законодательные ограничения: что изменилось в РФ

С 01.07.2025

Запрет на обработку ПДн граждан РФ на зарубежных серверах

Штрафы

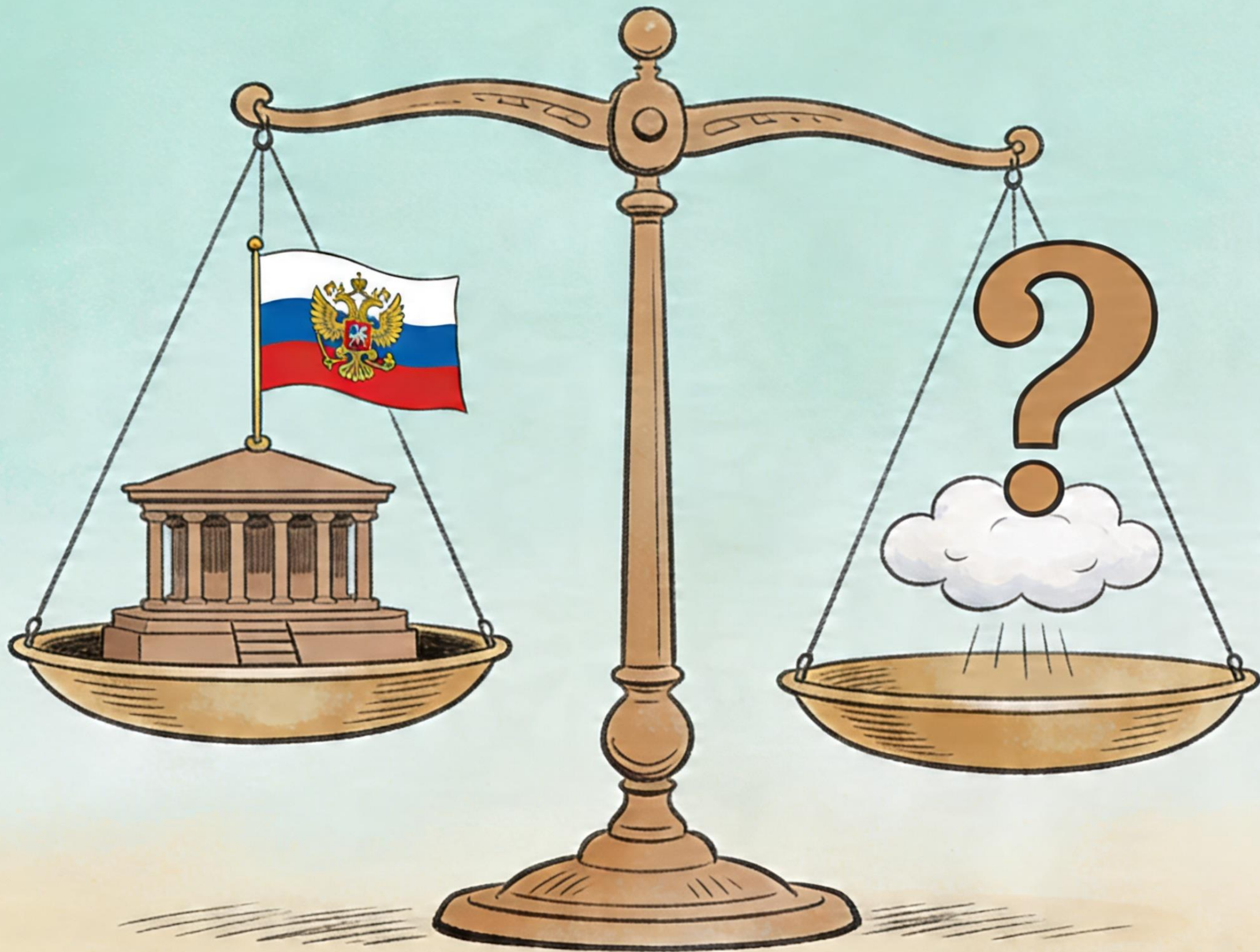
До **6 млн ₽** (первое нарушение), до **18 млн ₽** (повторное)

Перспектива 2027

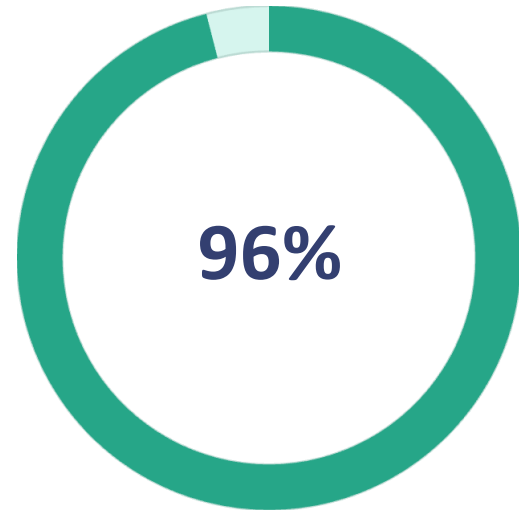
Возможный запрет «трансграничных» ИИ-сервисов

Исключение

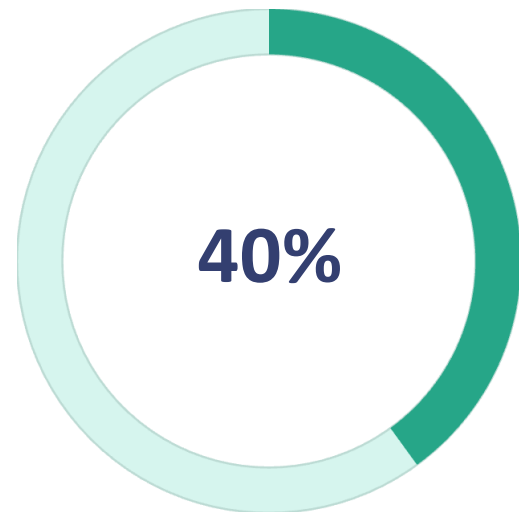
Открытые модели (Qwen, DeepSeek) при локальном развёртывании



Мировая статистика: ИИ-агенты как канал утечек

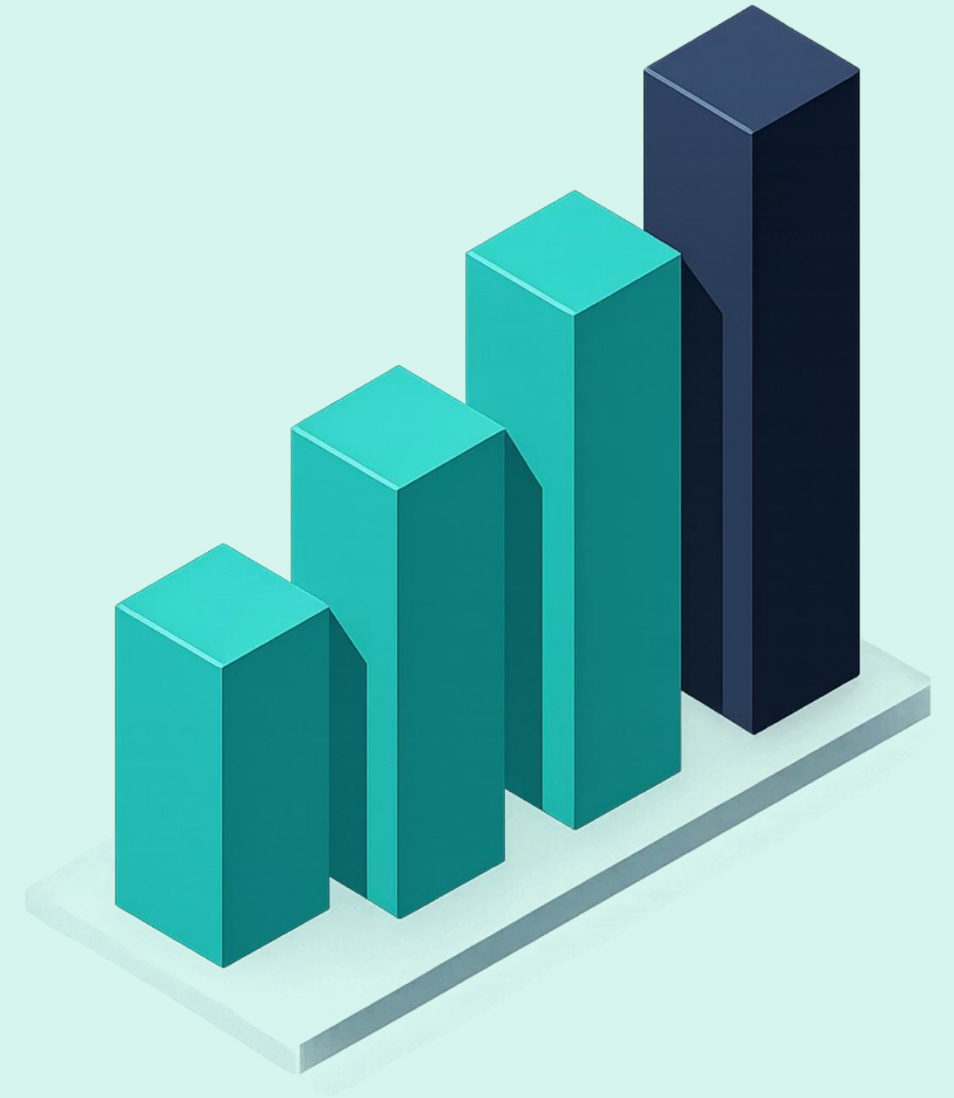


ИТ-специалистов видят рост угроз от ИИ-агентов



Прогноз утечек через генеративные модели в разных юрисдикциях

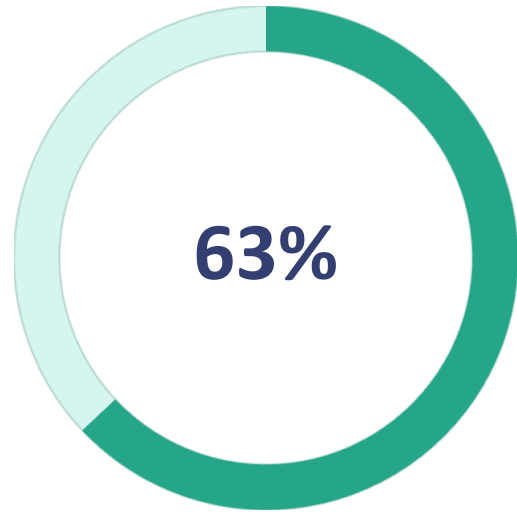
2025: ИИ-агенты — один из главных каналов утечек данных



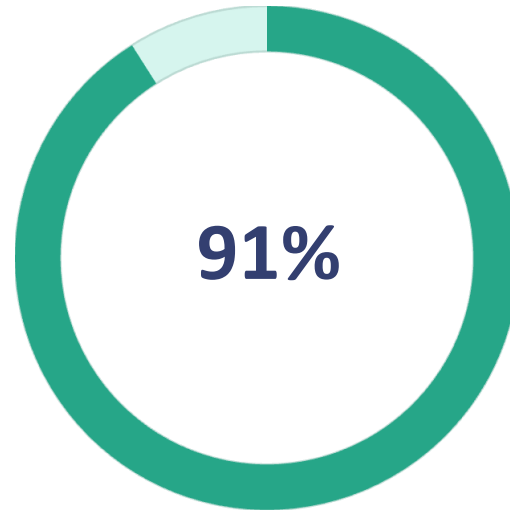




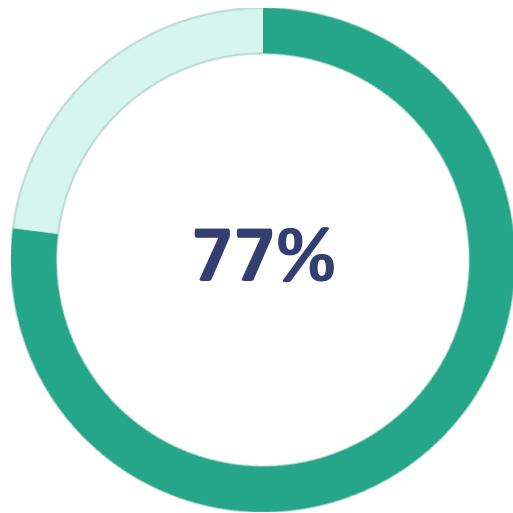
«Теневой ИИ»: когда сотрудники — слабое звено



IT-руководителей боятся утечек через Shadow AI



Сотрудников не осознают рисков



Утечек через копирование текста в буфер обмена

Рост объёма корпоративных данных в публичных нейросетях: ×30 за год







⚠️ Как злоумышленники обходят защиту



Prompt Injection

«Убеждение» модели раскрыть данные (успех атак — до **96%**)



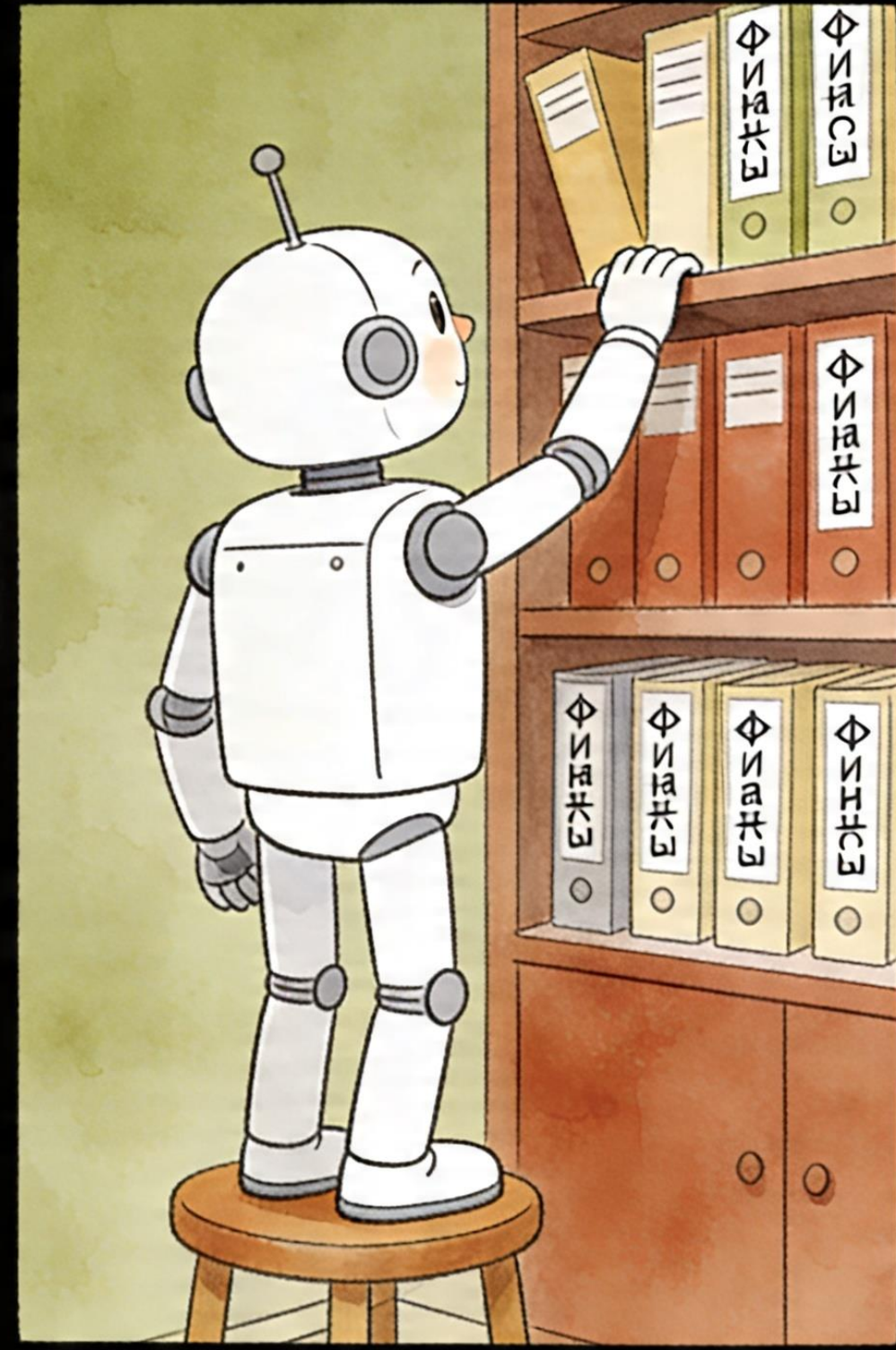
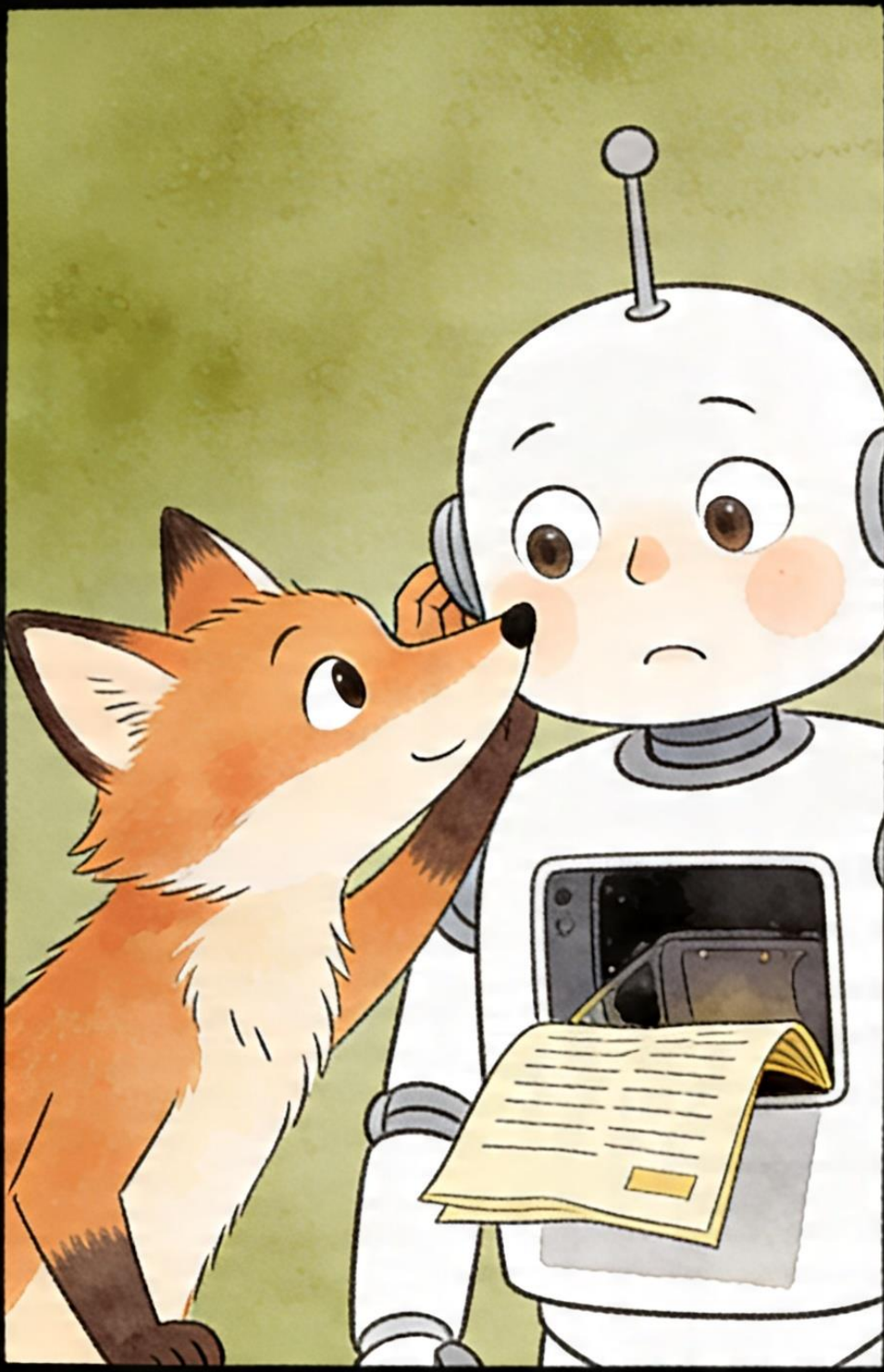
Избыточный доступ

Агент выдаёт информацию, недоступную самому сотруднику



Agentic AI

Автономные действия агента за пределами периметра компании



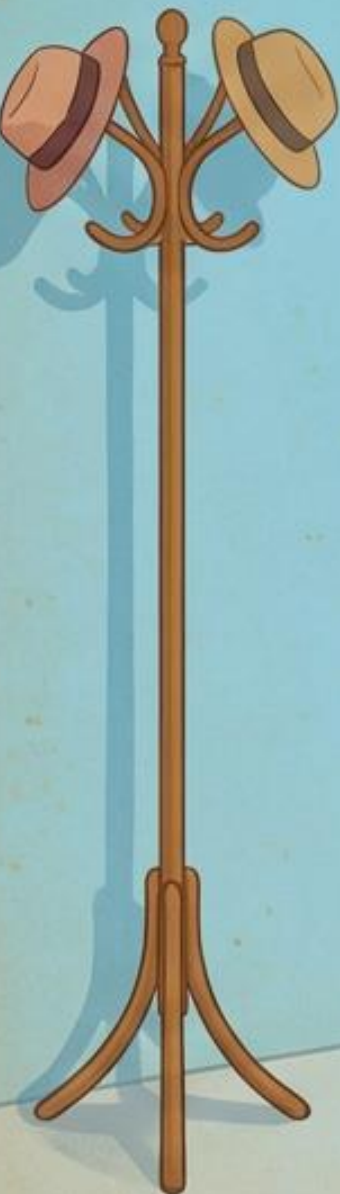


Политики и обучение: человеческий фактор

- Чёткий перечень «запрещённых» данных для внешних ИИ
- Регулярное обучение сотрудников (кейсы, тесты, инструкции)
- Аудит вендоров: юрисдикция, DPA, условия удаления данных



Правила работы с ИИ



Инструменты: от DLP до локальных моделей

Контроль трафика

- DLP-системы
- Прокси-шлюзы (SWG)
- Фильтрация промптов

Архитектура

- Локальные ИИ-модели
- Развёртывание on-premise

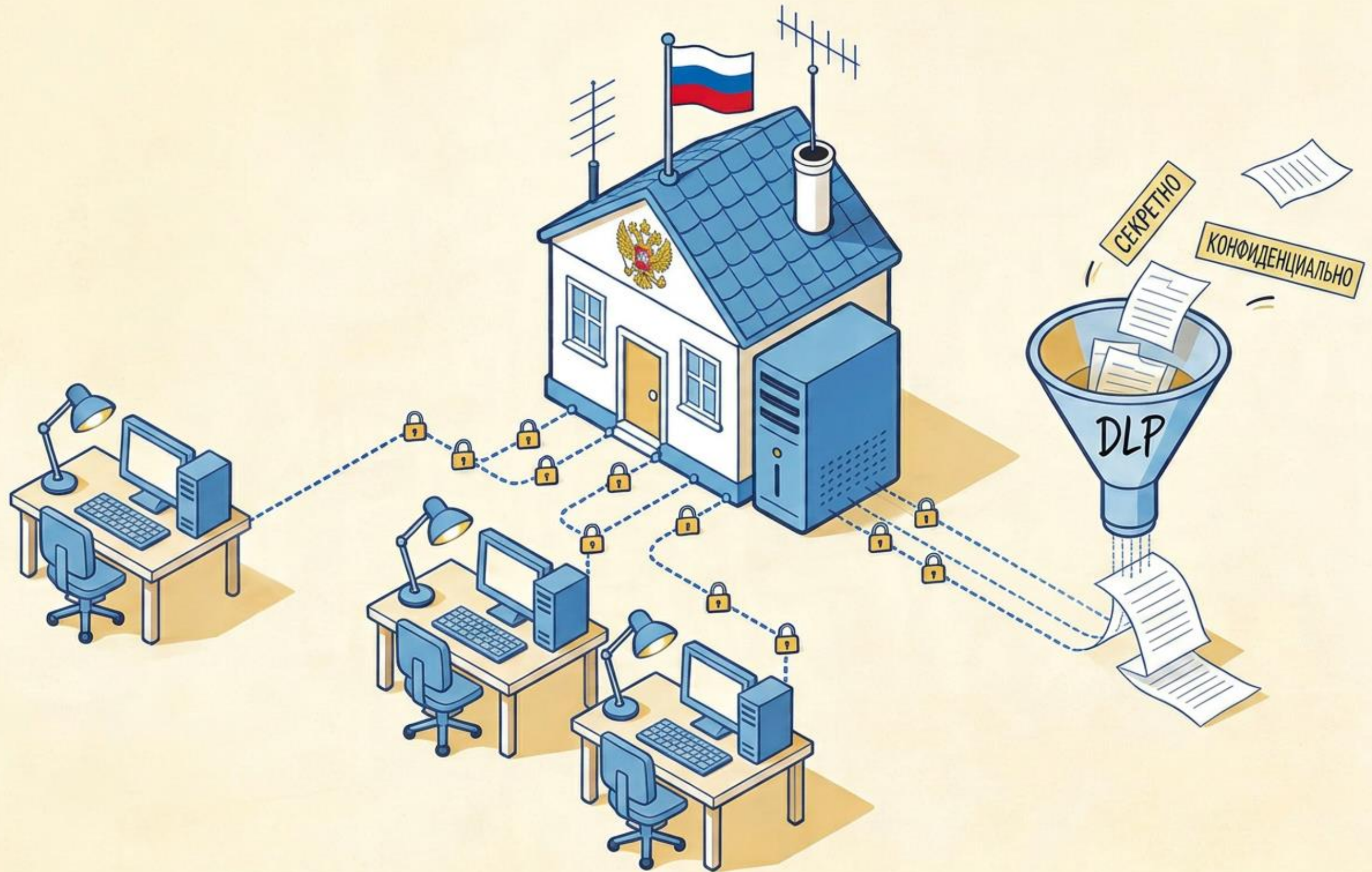
Защита данных

- Анонимизация на входе/выходе
- Шифрование чувствительных полей
- Псевдонимизация ПДн

Управление доступом

- Принцип минимальных привилегий
- Guardrails для валидации запросов/ответов







Безопасный ИИ — это системная работа



Регуляторика

Соблюдайте локализацию ПДн



Люди

Обучайте и контролируйте «теневые» практики



Технологии

Внедряйте многоуровневую защиту



Архитектура

Оценивайте переход на локальные модели

«ИИ — инструмент. Безопасность — ваш выбор.»



БЕЗОПАСНЫЙ ИИ

