

Использования инструментов ИИ в операционной деятельности

Декабрь 2025

Выбор подхода к ИИ - конфиденциальность данных

Внутренний (On-Premise) ИИ

- Модели на Python (Scikit-learn, PyTorch, TensorFlow): Для специализированных задач (прогнозирование, классификация, компьютерное зрение)
- Локальные большие языковые модели (LLM - например, через Ollama): для задач генерации текста, анализа документов, чат-интерфейсов внутри защищенного периметра

Внешний (Cloud/SaaS) ИИ

- Использование облачных API (OpenAI GPT, Gemini, Claude и др.).
- Готовые сервисы для аналитики, распознавания и генерации



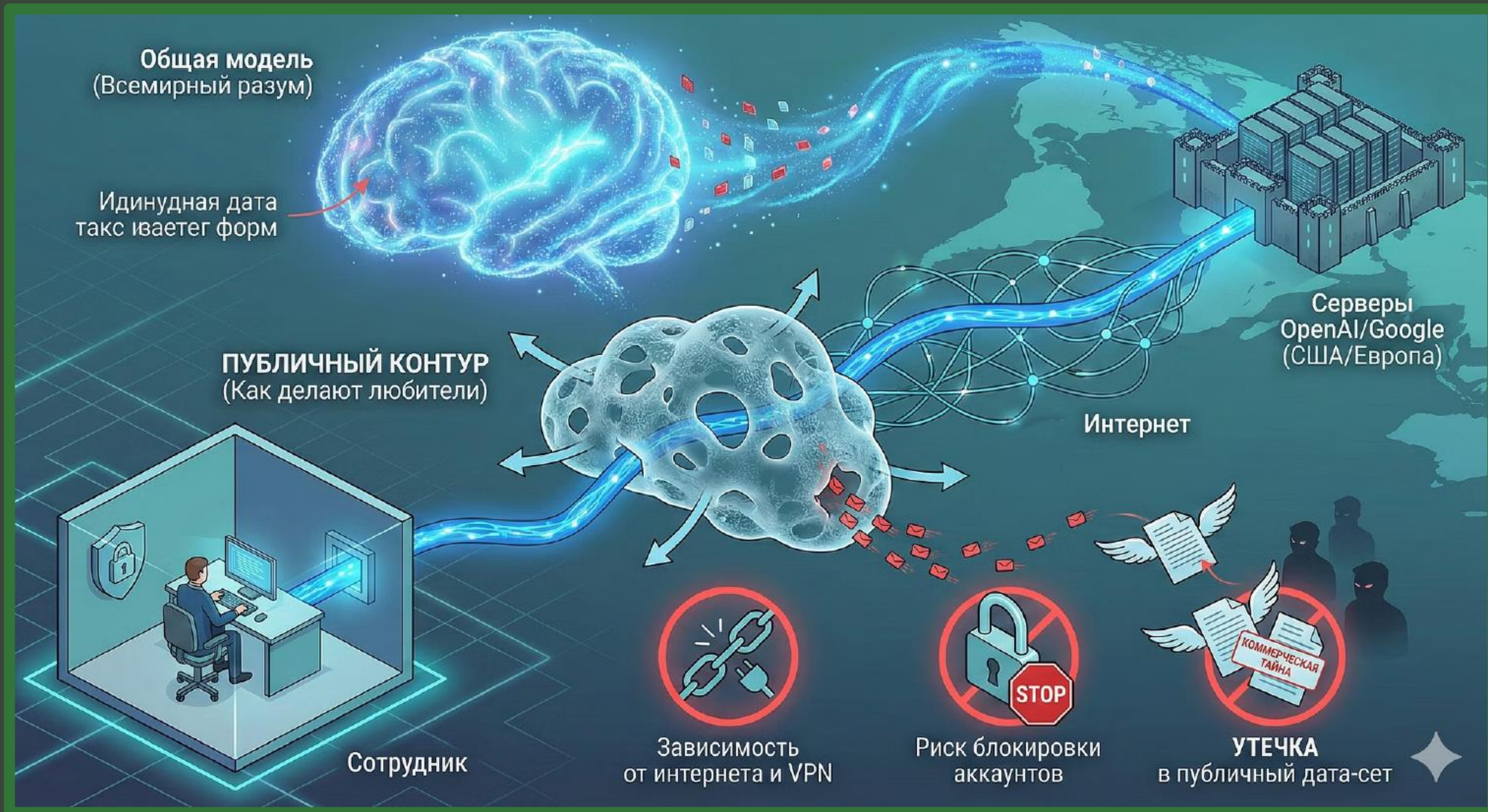
Тип данных	Рекомендуемый подход	Обоснование
Открытые или неконфиденциальные	Внешние ИИ-сервисы	<ul style="list-style-type: none">▪ Максимальная простота, мощность и функциональность без капитальных затрат
Конфиденциальные, персональные, коммерческая тайна	Внутреннее развертывание	<ul style="list-style-type: none">▪ Полный контроль над данными▪ Информация не покидает инфраструктуру компании, что исключает риски утечек.



Принцип принятия решения

- Классификация данных - потом выбор инструмента ИИ
- Высокий приоритет безопасности данных

Выбор подхода к ИИ – внешние модели



Выбор подхода к ИИ – внутренние модели

ЗАКРЫТЫЙ КОНТУР / On-Premise



Результат: Работаем без интернета. Секреты остаются в периметре.

Выбор подхода к ИИ - технические ограничения

Ключевое ограничение - Вычислительные мощности

- Отсутствие доступа к масштабным GPU-кластерам
- Единственный ресурс: сервер с ограниченным объемом видеопамяти
- Мы не можем развернуть масштабные LLM (например, GPT-4, Llama 3 70B) или сложные модели компьютерного зрения



Оптимизация и выбор легковесных решений

- Для кастомной аналитики и ML**
- Использование эффективных Python-моделей (легкие нейросети, ансамбли), которые решают конкретные задачи при небольших ресурсах

- Для работы с языком и документами:**
- Применение урезанных (quantized) версии LLM через Ollama (например, Llama 3.1 8B, Qwen2:7b, Gemma 2)
 - Эти модели работают на одной видеокарте с 8-24 ГБ VRAM и дают качественный результат для многих бизнес-задач



Приоритет — эффективность

- Точечное решение задач, а не создание универсального решения
- Оптимизация и облегчение моделей под наш ресурс

Выбор подхода к ИИ – выбор под задачи

Внешние большие модели - ChatGPT-4, Claude 3, Gemini Pro, DeepSeek, Midjourney

- Написание коммерческих предложений и отчетов
- Генерация маркетинговых текстов и контента
- Анализ рыночных трендов и конкурентов
- Подготовка презентаций и стратегий
- Перевод документов с контекстом
- Обучение и консультации сотрудников
- Креативный дизайн (промпты для изображений)
- Обработка обращений в поддержку (триаж)

Локальные GPU-модели - Llama 3, Mistral, Qwen, Stable Diffusion, Whisper-large

- Глубокий анализ юридических документов
- Полноценный Q&A по техдокументации (RAG)
- Распознавание и анализ медицинских снимков
- Контроль качества продукции по изображениям
- Анализ видео с камер наблюдения
- Генерация внутренней документации
- Моделирование бизнес-процессов (агенты ИИ)
- Персонализированные рекомендации для клиентов

Локальные CPU-модели - scikit-learn, LightGBM, TinyLlama, Sentence-BERT

- Глубокий анализ юридических документов
- Полноценный Q&A по техдокументации (RAG)
- Распознавание и анализ медицинских снимков
- Контроль качества продукции по изображениям
- Анализ видео с камер наблюдения
- Генерация внутренней документации
- Моделирование бизнес-процессов (агенты ИИ)
- Персонализированные рекомендации для клиентов

Рутину на CPU, креатив на API, секретное на GPU

Универсальные генеративные ИИ

Название	Ссылка	Краткое описание
DeepSeek	deepseek.com	Мощный и бесплатный ИИ-ассистент от китайской компании, поддерживает длинный контекст, имеет веб- и мобильную версию, умеет работать с загружаемыми файлами (изображения, PDF, Word и др.)
ChatGPT	chatgpt.com	Самый известный ИИ от OpenAI. Есть бесплатная (GPT-4o mini) и платные версии (GPT-4o) с расширенными возможностями, голосовым режимом и доступом в интернет
Claude	claude.ai	Ассистент от Anthropic, известный длинным контекстом (до 200K токенов), высокой безопасностью и качественной работой с текстом (анализ, написание, суммирование)
Google Gemini	gemini.google.com	Мультимодальный ИИ от Google, интегрирован в поиск и другие сервисы. Умеет работать с текстом, изображениями, аудио, есть продвинутая платная версия
Microsoft Copilot	copilot.microsoft.com	ИИ на базе моделей OpenAI, встроен в Windows, Bing и Office. Бесплатен, с доступом в интернет и разными режимами творчества
Grok	grok.x.ai	Ироничный и саркастичный ИИ от xAI (компания Илона Маска). Имеет доступ к данным платформы X (Twitter) в реальном времени

Российские ИИ

Название	Ссылка	Краткое описание
ЯндексGPT (Yandex Alice с умным ассистентом)	В приложении Яндекс и Алиса	Умный ассистент в экосистеме Яндекс. Доступен в приложениях Яндекс, в голосовой Алисе и в Yandex Browser .
GigaChat	https://giga.chat/	Мультимодальный ИИ от Сбера. Распознаёт и генерирует текст, изображения, аудио. Доступ по заявке, есть платные тарифы.

ИИ для генерации изображений, видео и дизайна

Название	Ссылка	Краткое описание
Midjourney	midjourney.com	Один из лидеров в генерации высококачественных художественных изображений. Работает через Discord-бота
DALL-E 3	Встроен в ChatGPT Plus и Copilot	Модель от OpenAI, известная точностью следования текстовому запросу и высоким качеством деталей
Stable Diffusion	stability.ai	Открытая модель для генерации изображений. Есть много онлайн-сервисов на её основе: Clipdrop, DreamStudio и др.
Gamma	gamma.app	То, что вы спрашивали! ИИ-инструмент для создания презентаций, веб-страниц и документов из текстовой идеи
Tome	tome.app	Ещё один мощный генератор презентаций и сторителлинга с помощью ИИ
Runway	runwayml.com	Продвинутый набор инструментов для генерации и редактирования видео, изображений с помощью ИИ

ИИ для работы с аудио и музыкой

Название	Ссылка	Краткое описание
Suno	suno.ai	Лидер в генерации полноценной музыки (мелодия и текст) по текстовому описанию
Udio	ud.io	Ещё один мощный генератор музыкальных треков с хорошим качеством и удобным интерфейсом
ElevenLabs	elevenlabs.io	Лучший в своем классе ИИ для генерации и клонирования реалистичных голосов
Mubert	mubert.com	Генерация фоновой музыки и треков для контента по текстовому промпту

Специализированные и нишевые ИИ

Название	Ссылка	Краткое описание
Perplexity AI	perplexity.ai	ИИ-поисковик, который отвечает на вопросы, цитируя актуальные источники из сети. Минимум "галлюцинаций".
Notion AI	Встроена в notion.so	ИИ-помощник прямо внутри популярного workspace-редактора Notion для написания, редактирования и суммирования текстов.
Grammarly	grammarly.com	Продвинутый ИИ-помощник для проверки грамматики, стиля и тона письма на английском и других языках.
Otter.ai	otter.ai	Специализированный ИИ для транскрибации аудио (например, лекций, встреч) с высокой точностью.
Jenni AI	jenni.ai	ИИ-ассистент для научного и академического письма, помогает в исследованиях и создании статей.