GreenData

Ищем эффекты от генеративного ИИ в корпоративной разработке

Perm ~~Winter~~ Wesna School '26

Александр Перевалов, PhD
Руководитель группы разработки ИИ, GreenData

О спикере

Curriculum vitae

2025 - н.в.	**Team Lead команды AI** Гриндата
2020 - 2025	**PhD (Dr. rer. nat.)** HTWK Leipzig & University of Paderborn
2020 - 2025	**Research assistant** Hochschule Anhalt → HTWK Leipzig
2019 - 2020	**MSc Data Science (двойной диплом)** Hochschule Anhalt & Пермский Политех
2018 - 2020	**Разработчик** Форсайт

Некоторые успехи в AI

🔬 **Статьи на конференциях A/A*: The Web Conf, WSDM, LREC, ISWC**
h-Index (Индекс Хирша): 10

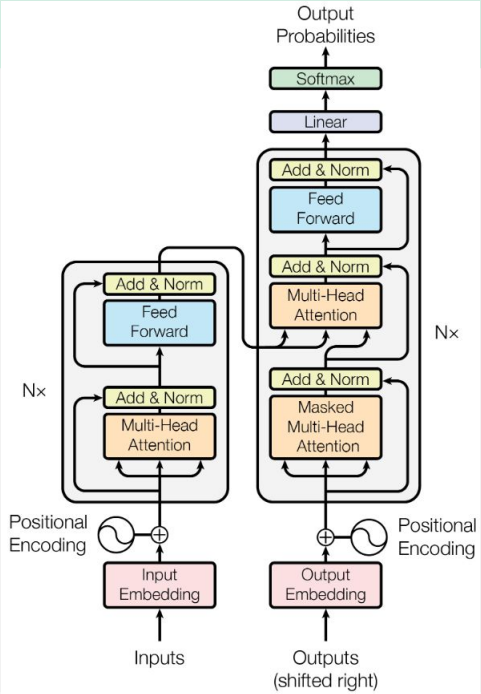
🏆 **Победы на хакатонах**
- TenderHack 2024 – ИИ-ассистент для портала поставщиков
- Hack.Genesis 2024 – ИИ-ассистент для анализа отчётности
- TenderHack 2026 – Система обоснования НМЦК

🥈 **Призовые места на ML/AI соревнованиях**
- Sber AI Journey 2024
- Sber AI Journey 2025

🏆 **Победитель Stepik Awards в номинации «Анализ данных и AI»**
Курс AI DevTools

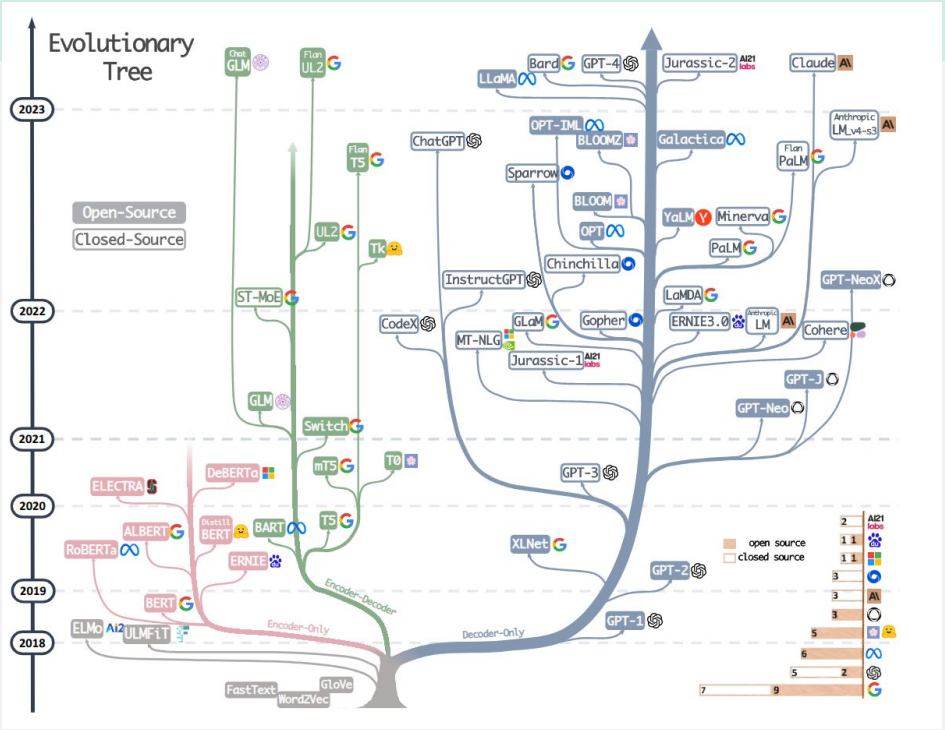
Языковые модели как драйвер GenAI – триггер №1

Attention Is All You Need, Vaswani et al. (2017)



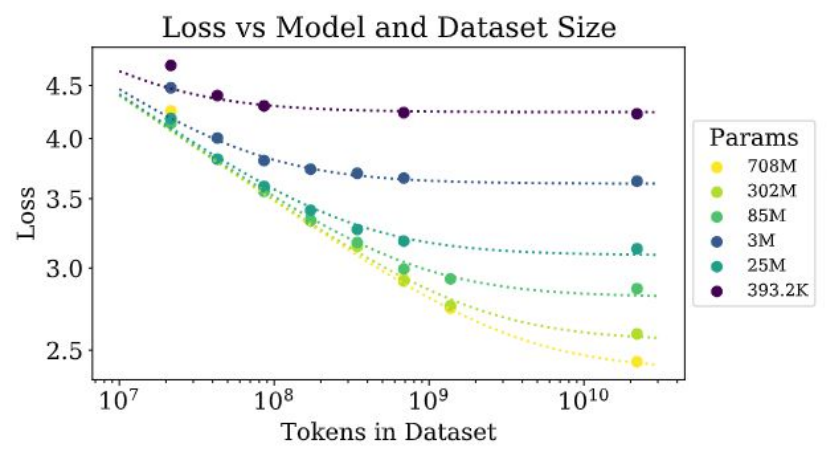
Механизм внимания (attention) породил несколько семейств современных моделей: от BERT до GPT

Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond (2023)

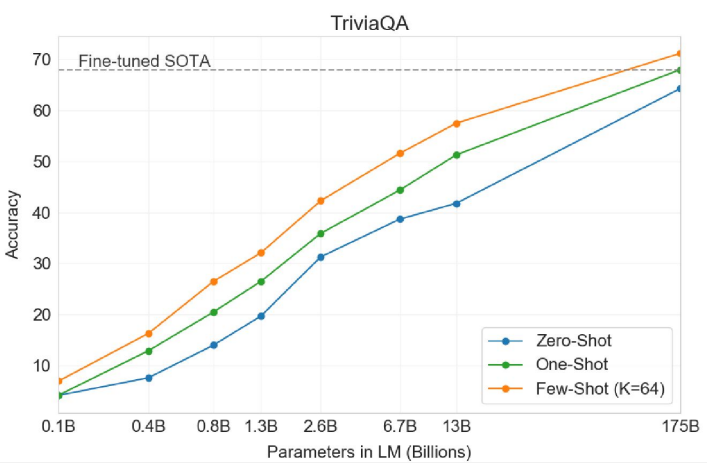


Языковые модели как драйвер GenAI – триггер №2

Scaling Laws for Neural Language Models by OpenAI (2020)



Language Models are Few-Shot Learners by OpenAI (2020)



Законы масштабирования языковых моделей – новый виток в развитии:

- При фиксированном размере датасета более крупные модели достигают меньшей ошибки;
- Предсказуемость “scaling laws” сделала инвестиции в крупные модели просчитываемыми;
- Предобученные LLM начали работать на уровне дообученых “специализированных” LM.

Ключевые различия ML и GenAI

Машинное обучение – ML

- Одна модель – одна задача
- Задачи узкоспециализированы и привязаны к обучающим данным
- Проще добиться высокого качества, для ряда моделей доступна интерпретируемость
- Многие классические ML-модели (линейные, бустинги, небольшие нейросети) могут обучаться и инференситься на CPU

Генеративный искусственный интеллект – GenAI

- Одна модель – несколько задач. Обычно подразумеваются большие языковые модели (LLM)
- Возможности строить рассуждающих агентов и мультиагентные системы
- Низкий порог входа, быстрый “вау-эффект”, значительное время до промышленного использования
- GPU-friendly – требует ощутимо больше ресурсов для старта во внутреннем контуре

GenAI-модель позволяет быстро стартовать на нескольких задачах с минимальной разметкой, но требует ощутимых вычислительных ресурсов и инженерной работы для вывода в промышленную эксплуатацию.

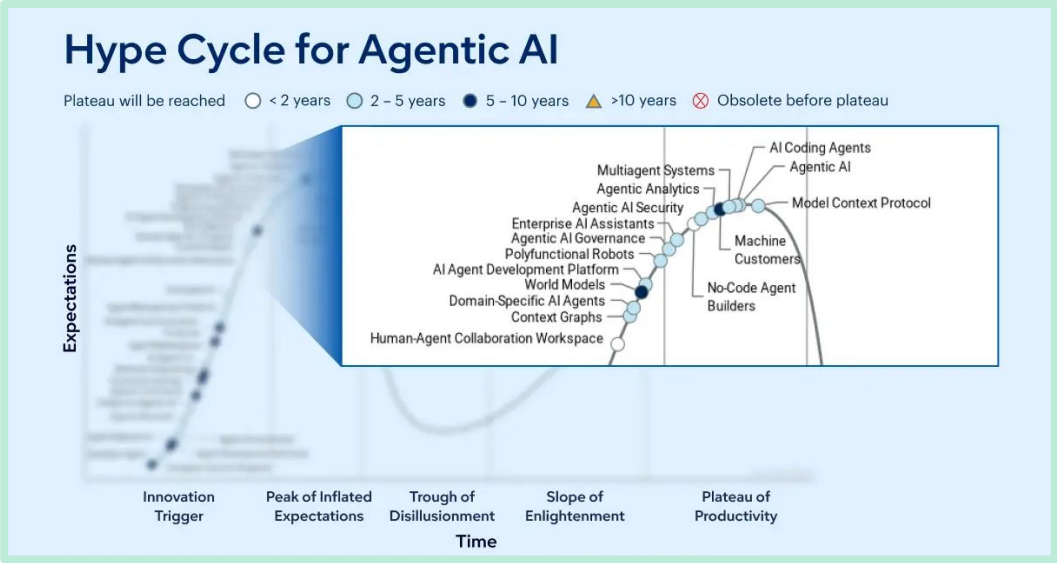
ML-модель требует больше трудозатрат на сбор и разметку данных под каждую задачу, но на узких структурированных задачах даёт более высокое качество при меньших вычислительных затратах.

AI-агенты

AI-агенты это дополненные языковые модели

Источник: Lei Wang, Chen Ma, Xueyang Feng, et al. “A survey on large language model based autonomous agents”. Frontiers of Computer Science. 2024

Где мы находимся – общемировые тренды



0 RPA → “AI agents”

1 AI TRiSM

2 AI-ready data & platforms

3 AI and the workforce

Источник: What the 2026 Hype Cycle for Agentic AI Reveals, Gartner

Большинство технологий генеративного ИИ уже в фазе утраты иллюзий

Где мы находимся – контекст РФ и GreenData

Совокупный объем инвестиций в ИИ в России снизился с 530 до 451 млн. долларов. Доля инвестиций в ИИ-агентов выросла с 10 до 31 млн. долларов.

Специализация инвестиций свидетельствует о закономерном созревании ИИ и стремлении бизнеса получать конкретный экономический эффект.

Источник: МТС, ИИ-агенты: обзор мирового и российского рынков, 2025

Бизнес в России свернул или отложил 9 из 10 проектов по внедрению генеративного ИИ

Источник: Опрос консалтинговой компании «Интеллектуальная аналитика», 2026

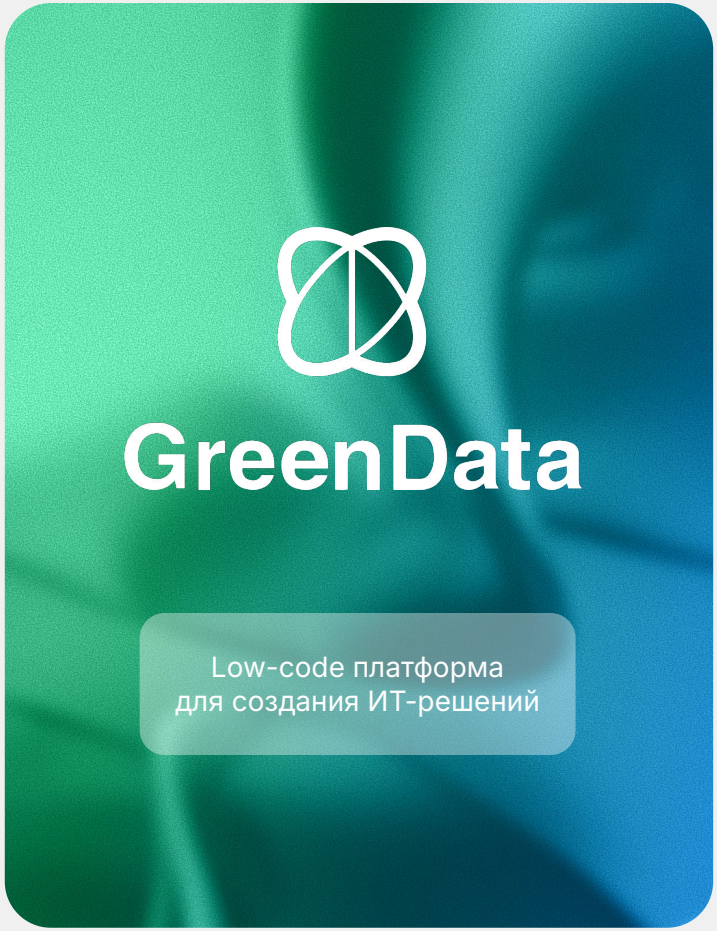
Успешный пилот у крупнейшего клиента (банк) по распознаванию документов

В процессе внедрения обновленный модуль распознавания отчётности (РСБУ)

В основе – технологии OCR и извлечения атрибутов с помощью генеративного ИИ

Десятки демо-показов как новым, так и существующим клиентам

От ИИ-ассистентов и генерации ответов на входящую почту до распознавания десятков типов документов



Стратегический фокус компании в части искусственного интеллекта

AI-nativeness в процессах и продукте

Продукты и проекты с AI под капотом

Собственные данные и модели

Работа с большими языковыми моделями

Экономика обучения больших языковых моделей – высокая стоимость создания и дообучения LLM заставляет бизнес разрабатывать простые и эффективные решения вокруг фундаментальных моделей.

*Для модели 32 млрд. параметров с учетом рыночной стоимости ресурсов

Рецепты для повышения доступности LLM

- Context engineering – построение приложения вокруг LLM и процесс подбора релевантного контекста для неё (RAG + Prompt Tuning)
- Использование моделей, адаптированных под нужный язык или домен
- Методы эффективного обучения и инференса: Квантизация, Дистилляция, P-tuning, Steering vectors, Memory offload, KV Caching

Источник: AI-экосистема Т-Банка на основе SLM, Виктор Тарнавский. AI Journey 2024

Воронка внедрения генеративного ИИ в enterprise

Цифровая готовность процессов

Наличие инфраструктуры

Интеграция с legacy

ИБ / Законодательство

- Учёт в рукописном журнале, а несколько томов базы знаний лежат шкафу
- Хотим on-prem, у нас нет GPU, ждать 10 минут до первого токена не можем
- ИИ-агент должен “ходить” во внешнюю систему, документации к которой никогда не было
- Использовали open-source библиотеки, но словили десяток critical-уязвимостей

Препятствия на каждом этапе снижают экономическую целесообразность

Цифровая готовность процессов – топливо для ИИ

**Парадокс: для того, чтобы внедрить ИИ, знания должны быть оцифрованы.
Но чтобы их оцифровать нужен ИИ.**

В 2024 году компания Anthropic (Claude) поставила задачу получить “все книги в мире”.

Anthropic потратили несколько миллионов долларов на покупку книг, а затем оцифровали их для использования в качестве обучающих данных.

Источник: Материалы окружного суда США по делу Anthropic

Базы данных Springer Materials (материаловедение) были получены путём оцифровки печатных книг высококвалифицированными специалистами.

Специализированная и уникальная информация – конкурентное преимущество в эпоху генеративного ИИ.

OCR модели на базе мультимодальных LLM позволяют извлекать графики, формулы и таблицы, существенно упрощая оцифровку → увеличение эффективности труда.

Optical Character Recognition (OCR) в 2026 году

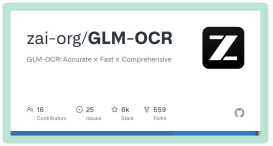
1 Попробовать Open Source

и понять, что под капотом старый-добрый tesseract или EasyOCR

Docling

LiteParse

2 Подключить LLM-OCR



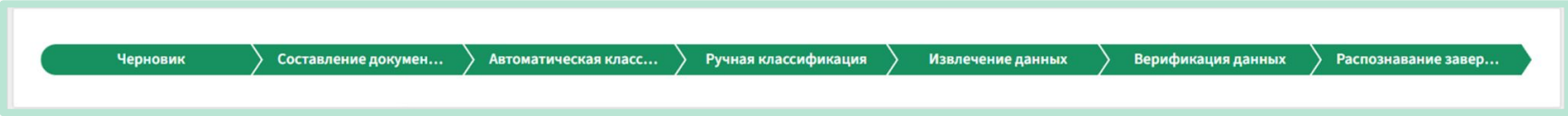
3 Дообучение и эвристики

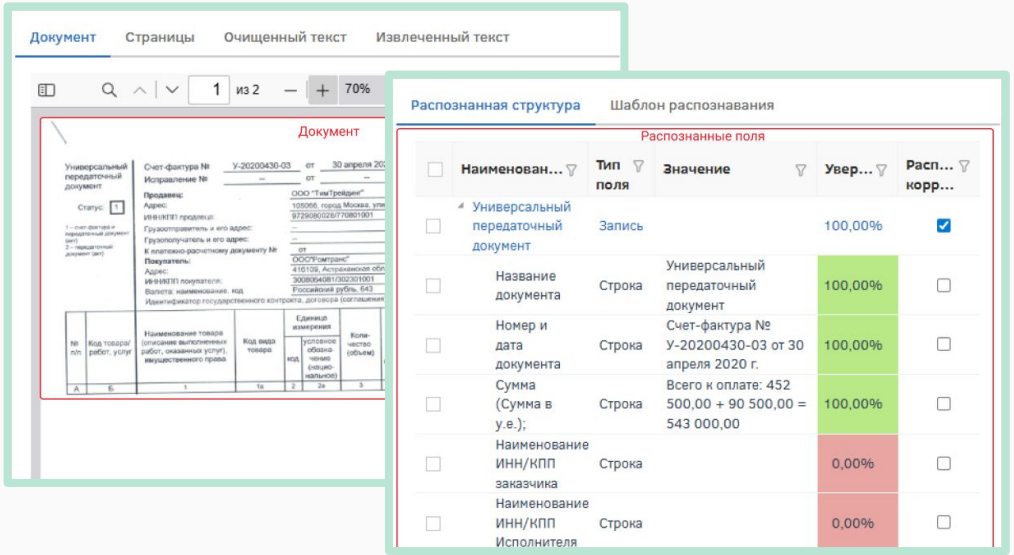
Контекст русского языка и специфики документов не позволяет достигнуть приемлемого качества из коробки.

Open Source позволяет добиться 80% желаемого результата. Остальные 20% лежат в области дообучения моделей и построения эвристик вокруг распознанных данных.

Распознавание документов в корпоративной среде

Конструктор распознавания документов GreenData





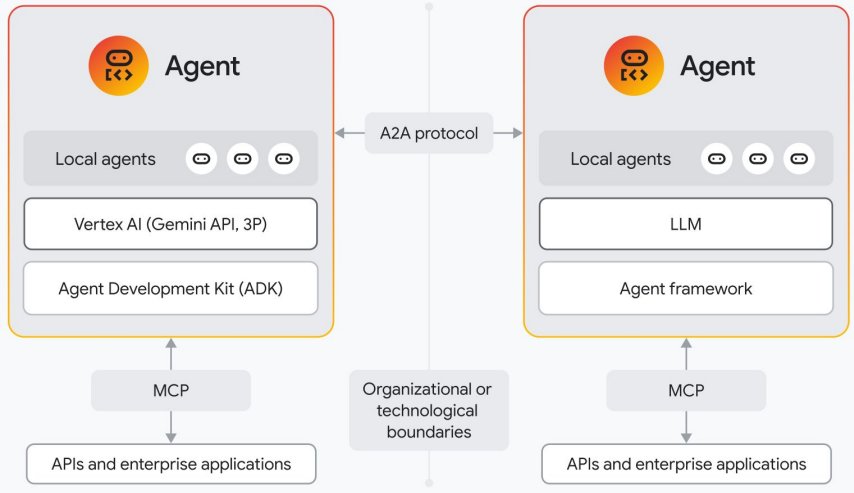
- Разметка и дообучение
- Типовые и нетиповые документы
- Верификация данных человеком
- Нестандартные сканы, множество документов в одном файле

AI и интеграция с legacy

Вызов внешних инструментов (протокол MCP)
конвертирует ответы LLM в реальные действия (напр. вызов API)

Использование навыков (SKILL.md)
навык позволяет дать модели определённый набор инструкций, не перегружая её контекст.

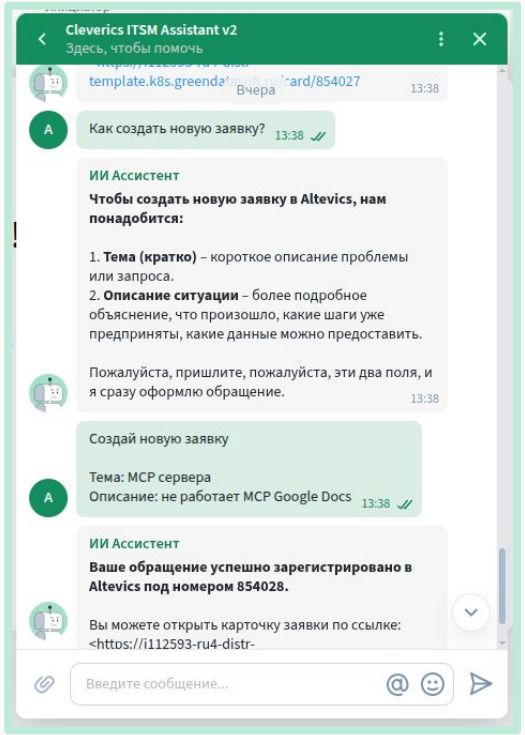
Мультиагентность (протокол A2A)
взаимодействие (оркестрация/хореография) нескольких агентов по универсальному протоколу



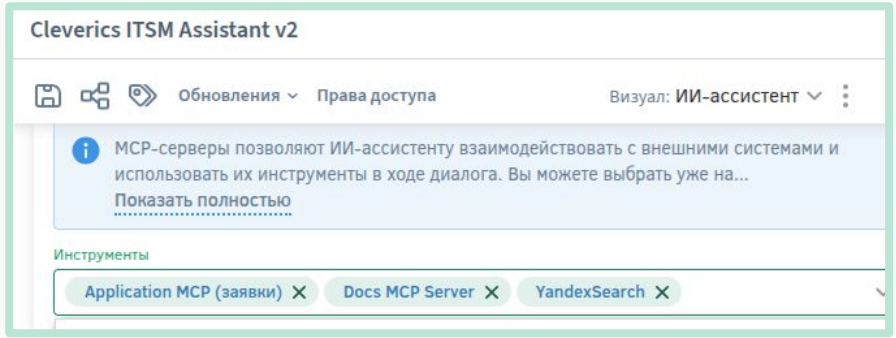
Источник: AI Agent Trends 2026 by Google Cloud

Приложения должны идти по пути AI-native и быть “в доступе” для агентов через MCP. Знания, ранее лежавшие в головах специалистов конвертируются в “скиллы” агентов.

Подводные камни агентских интеграций



- Права доступа конкретных сотрудников на конкретные действия или документы
- “А у нас своё” – продукты с LLM под капотом должны быть независимы от конкретной модели
- MCP – новое API. В интересах каждой платформы поддерживать свой MCP-сервер.



Законодательство и курс на суверенность

Россия: ФЗ об основах государственного регулирования сфер применения технологий ИИ

- Доверенные модели ИИ (реестр моделей)
- Суверенные модели – разработанные полностью в РФ
- Наборы данных также формируются внутри страны

Источник:
https://regulation.gov.ru/projects/166424/

Европа: резолюция о цифровом суверенитете

- Создание собственного “облака” и ИИ моделей
- Предпочтения европейским компаниям при закупке железа
- Госсредства на open source проекты

Источник:
https://www.br.de/nachrichten/netzwelt/eu-parlament-grosse-mehrheit-fuer-digitale-souveraenitaet

Курс на суверенность виден в РФ и Европе. Однако лидерами в ИИ-гонке остаются США и Китай

Подписывайтесь
на наш telegram

Краткие выводы

- Порог входа в генеративный ИИ низкий, но вау-эффект быстро сменяется разочарованием;
- Будущее платформ и приложений, недоступных нативно AI-агентам, под сомнением;
- 80% результата доступно из коробки, остальные 20% не лежат на поверхности;
- Курс на суверенность: обладание уникальными данными → конкурентное преимущество.

Александр Перевалов, PhD

Руководитель группы разработки ИИ, GreenData

perevalov.am@greendata.ru