red_mad_robot

Рынок GenAI в 2025

Что нужно знать бизнесу

Тренд-репорт
от Аналитического центра
red_mad_robot

Февраль, 2025

Создаём AI-решения, используя многолетний опыт запуска комплексных цифровых продуктов

16
лет мы драйвим цифровой рынок

3
актива в группе компаний с AI-экспертизой

40+
проектов в сфере AI реализовано за 5 лет

650+
специалистов в группе компаний red_mad_robot

GIGA CHAT
стратегический партнер Сбера по LLM Gigachat

Forbes
серебро в рейтинге лучших работодателей России

- Запуск цифровых продуктов под ключ
- AI, ML & CV
- Discovery & Analytics
- Mobile, Web, Backend Dev
- GenAI & LLM
- Internet of Things
- UX/UI-дизайн
- Product & Tech Consulting

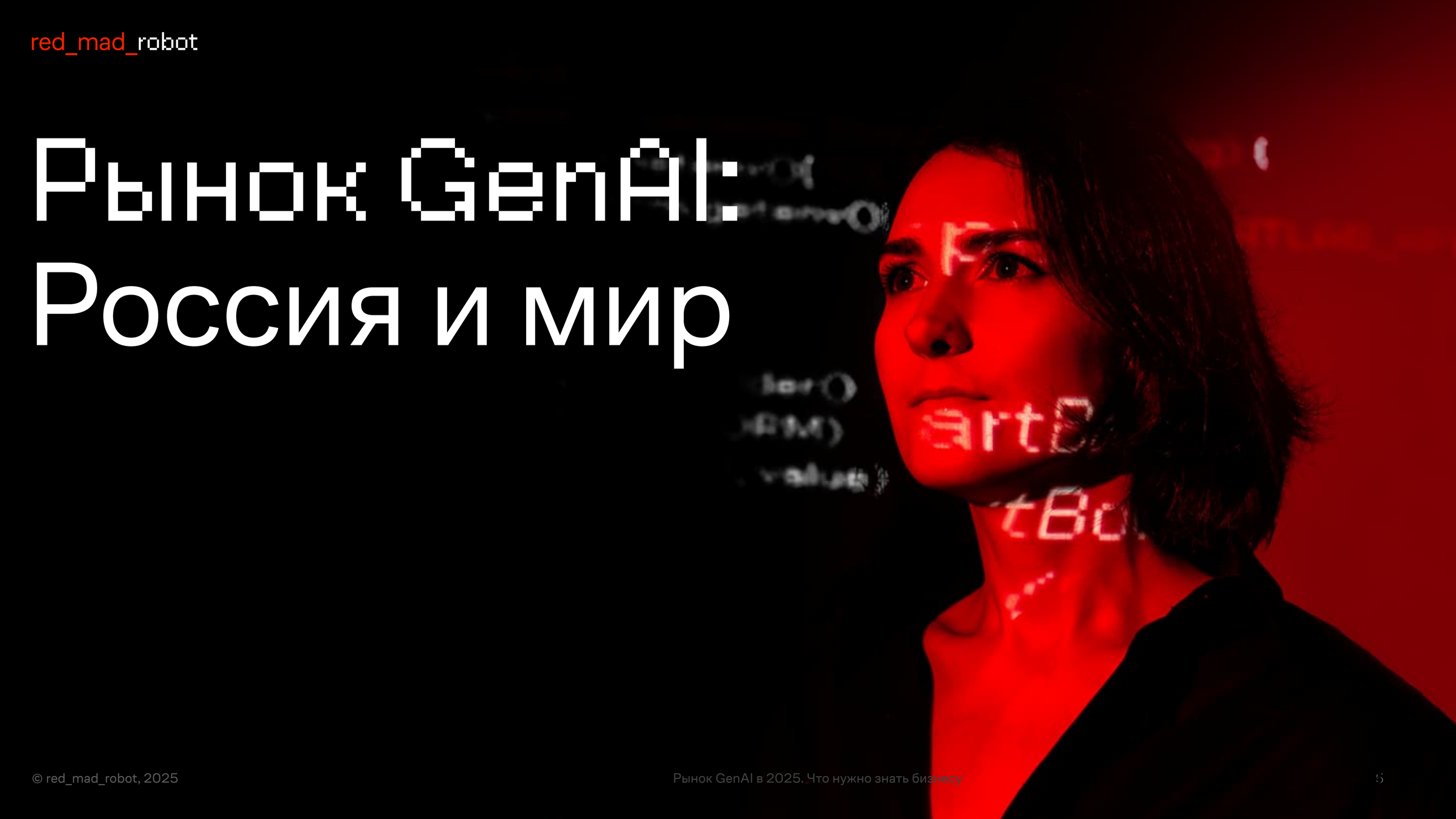
Строим цифровое производство с нуля, ищем и проверяем гипотезы, запускаем продукты и масштабируем их на рынок.

Развиваем R&D-центр в сфере искусственного интеллекта и помогаем компаниям осваивать GenAI-технологии.

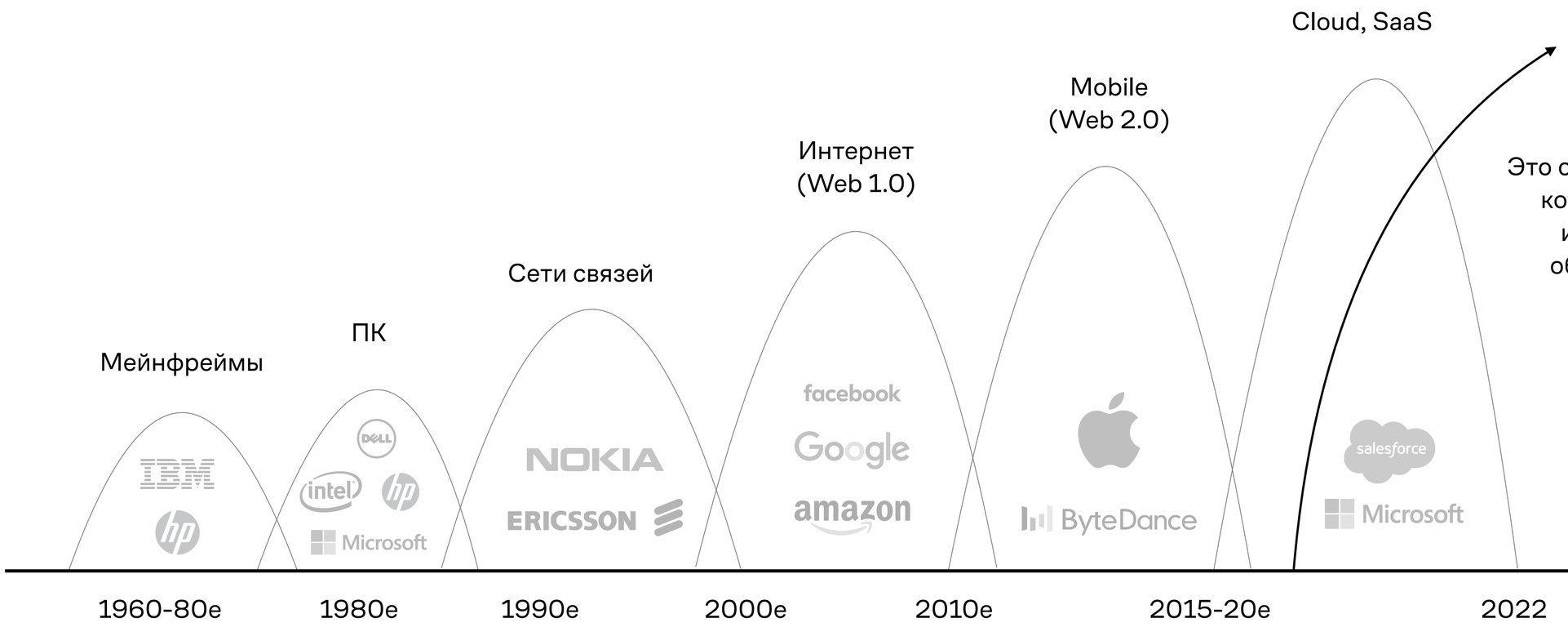
Запустили 100+ продуктов с клиентами по всему миру

Содержание

Рынок GenAI: Россия и мир



Развитие AI предвосхищает очередной цикл технологической трансформации



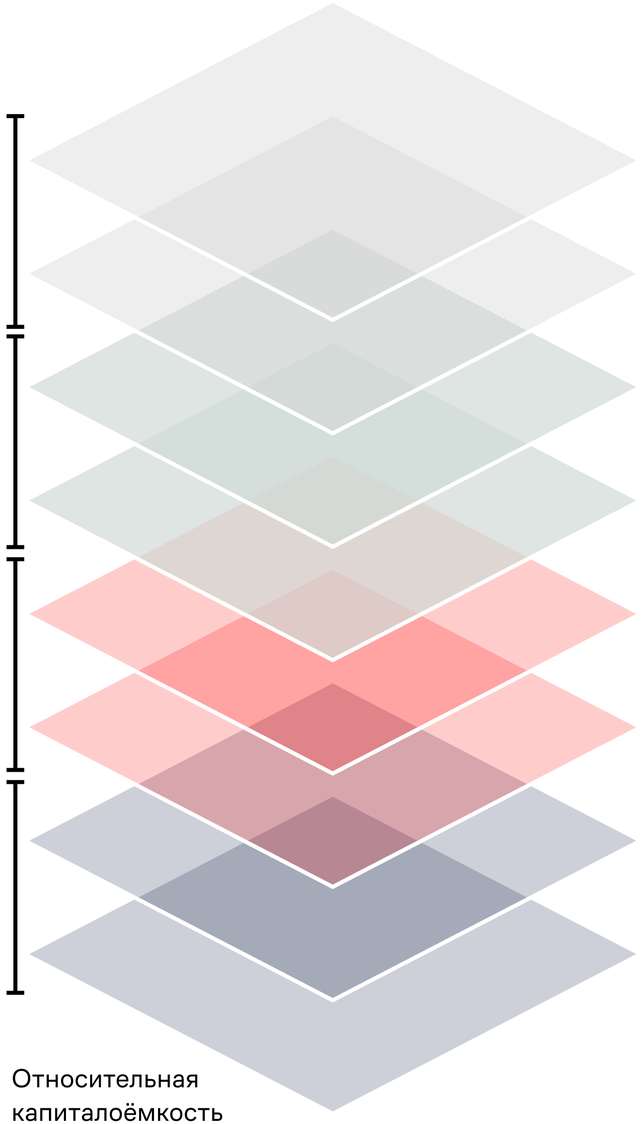
NVIDIA
Meta
OpenAI

AI

не очередной хайп.

Это следующий технологический цикл, который приведёт к масштабным изменениям и долгосрочному обновлению бизнес-процессов.

Архитектура рынка GenAI



Сервисы для конечных пользователей

Инструменты для профессиональных пользователей

Платформенный уровень

Технологический уровень

Относительная капиталоёмкость развития технологии

Цифровые каналы и сервисы

AI-агенты и копилоты

AI-инструменты и студии разработки

Наборы данных и промпты

Мультимодальные AI-платформы

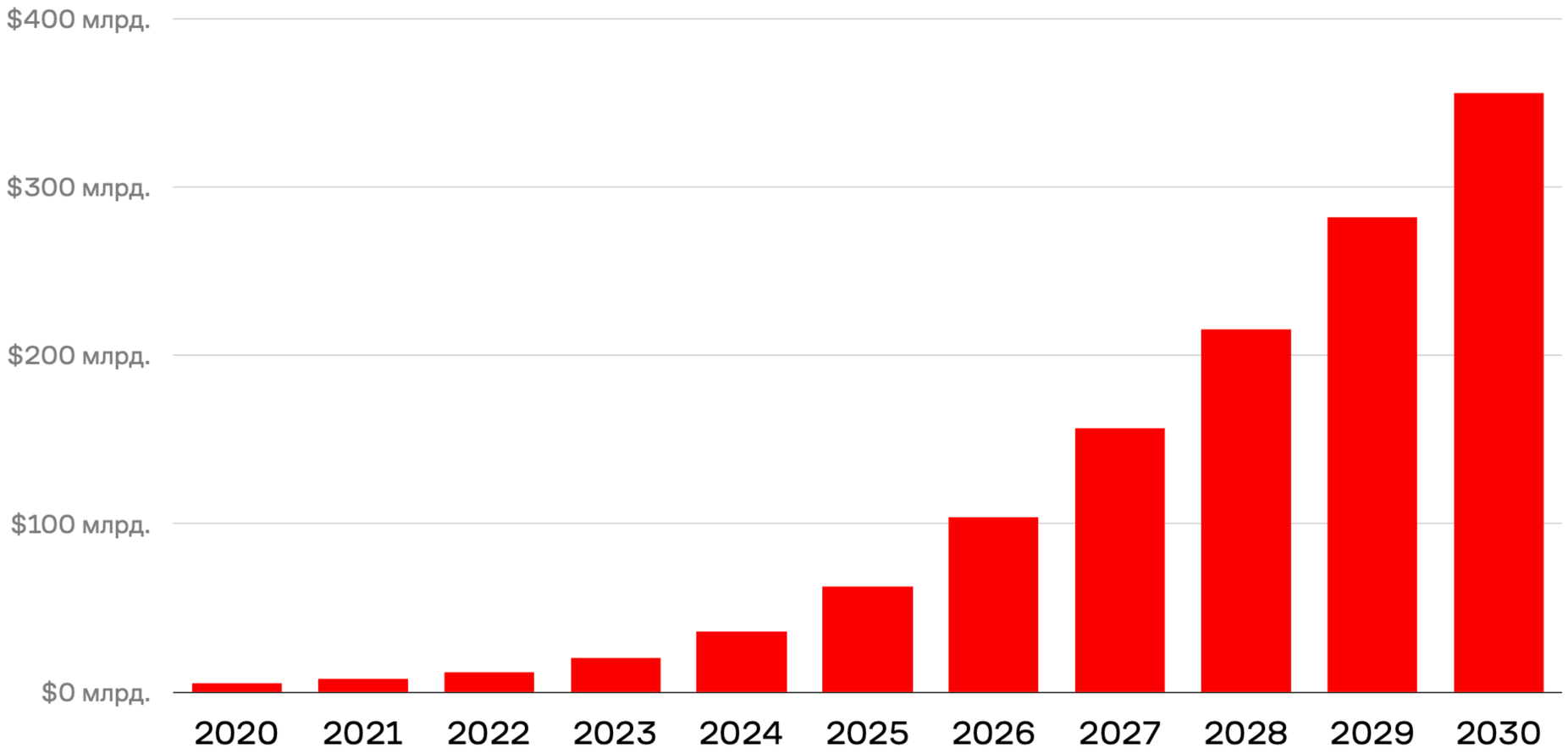
LLM

Векторные базы данных

Технологическая инфраструктура

Ожидаемый среднегодовой темп роста рынка GenAI в 2024–2030 — 46,48%.

К 2030 году его объём увеличится в 10 раз, достигнув $356,1 млрд.

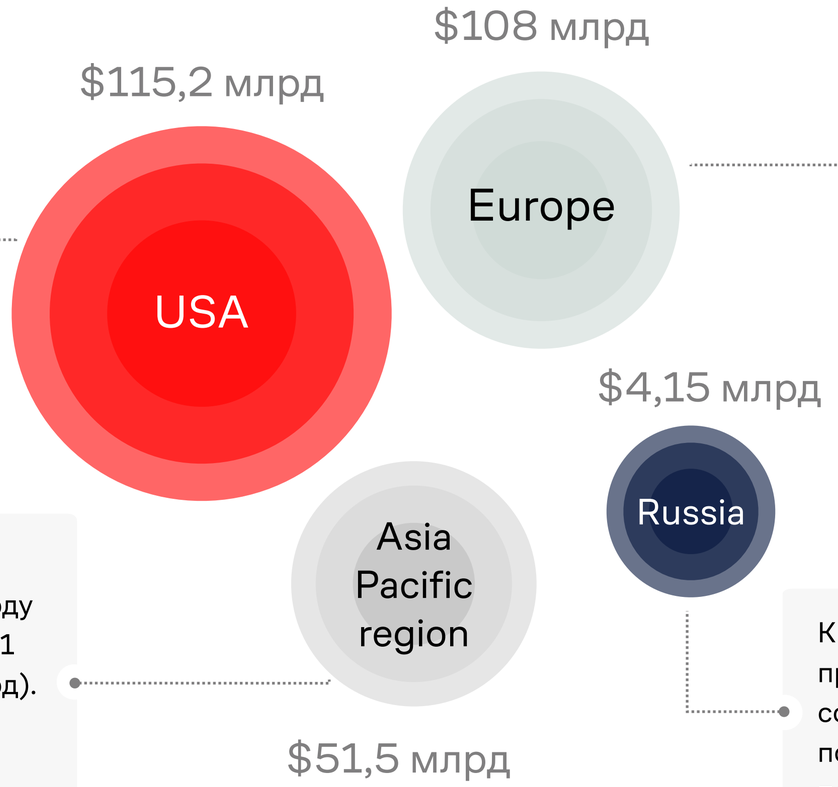


Источник: statista

Мировой рынок GenAI к 2030 году достигнет $356 млрд

США, Европа и Азия будут играть ключевую роль в развитии рынка, однако у России и MENA достаточно потенциала для появления технологического единорога.

США: к 2030 году объём рынка GenAI прогнозируется на уровне $115,2 млрд. Основные сегменты — реклама, медиа и игры, при этом ожидается, что рекламный сектор станет ключевым драйвером роста. США сохранят лидерство благодаря технологическим инновациям и масштабным венчурным инвестициям.



Европа: к 2030 году рынок GenAI прогнозируется на уровне $108 млрд. Ключевые сегменты — здравоохранение, автомобилестроение и производство, при этом наибольший рост ожидается в сфере медицины благодаря развитию AI-решений в диагностике и лечении.

Азиатско-Тихоокеанский регион демонстрирует стремительный рост внедрения GenAI — к 2030 году рынок ожидаемо достигнет $52 млрд (Китай — $31 млрд, Япония — $13 млрд, Южная Корея — $7 млрд). Развитие обусловлено мощной технологической инфраструктурой и активными инвестициями в исследования.

К 2030 году объем рынка GenAI в России прогнозируется на уровне $4,15 млрд, со среднегодовым ростом 25%. Технология набирает популярность как среди бизнеса, так и среди частных пользователей.

Источники: 1, 2, 3, 4, 5, 6.

Два ключевых игрока на рынке AI — США и Китай. Россия активно развивает технологии, чтобы сократить разрыв. Европа и Великобритания только вступают в гонку.

«Западный мир»

Великобритания

с 2025 года объём внебюджетных инвестиций в AI составит $17,36 млрд. Ключевые игроки — Vantage Data Centres, Nscale и Kyndryl.

Европа

С 2025 года совокупное финансирование AI со стороны ЕС и государств-членов составит $2,04 млрд. Также 11 февраля было объявлено о возможных инвестициях в $206 млрд.

США

IT-бюджет правительства США на 2025 год — $75,13 млрд. Внешние инвестиции от таких игроков, как SoftBank, OpenAI, Oracle и MGX, могут достичь до $500 млрд с 2025 года.

«Русский мир»

Россия

К 2030 году правительство планирует выделить $0,3 млрд на развитие AI. Вложения из внебюджетных источников, таких как Сбер и РФПИ, могут составить $1,16 млрд.

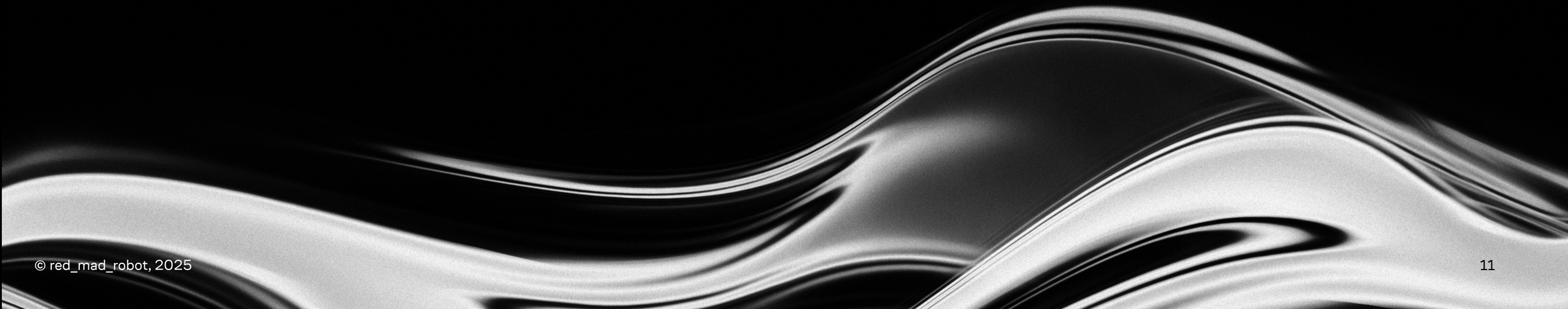
«Великий китайский файрвол»

Китай

Планируется инвестировать $138 млрд, при этом инициатива полностью финансируется государством.

Источники: 1, 2, 3, 4, 5.

Тренды GenAI 2025



Топ-3 события, которые произошли в начале 2025 года

StarGate

- США запускают StarGate — масштабный проект стоимостью $500 млрд, направленный на технологическое доминирование в AI. Уже выделено $100 млрд на первую фазу.
- Проект объединяет усилия ведущих IT-компаний – SoftBank, OpenAI, Oracle, Microsoft, NVIDIA и других. Основные направления работы: развитие AGI, персонализированная медицина (диагностика рака, секвенирование генома, вакцины нового поколения) и прорывы в фундаментальных моделях.
- В рамках проекта строится 20 дата-центров (10 уже возводятся в Техасе) мощностью 5 ГВт – эквивалент двух-трёх атомных электростанций.

AI без ограничений

- Новый президент США Дональд Трамп отменил указ экс-президента Байдена о безопасности AI, что снимает часть ограничений на разработку и внедрение технологий. Это даёт компаниям больше свободы, но одновременно повышает риски: усиление AI-монополий, отсутствие чётких стандартов этики и безопасности, рост числа «серых» разработок.
- В результате рынок ожидает новую волну конкурентной гонки между странами и корпорациями, где регулирование становится вопросом политики, а не технологии.

DeepSeek

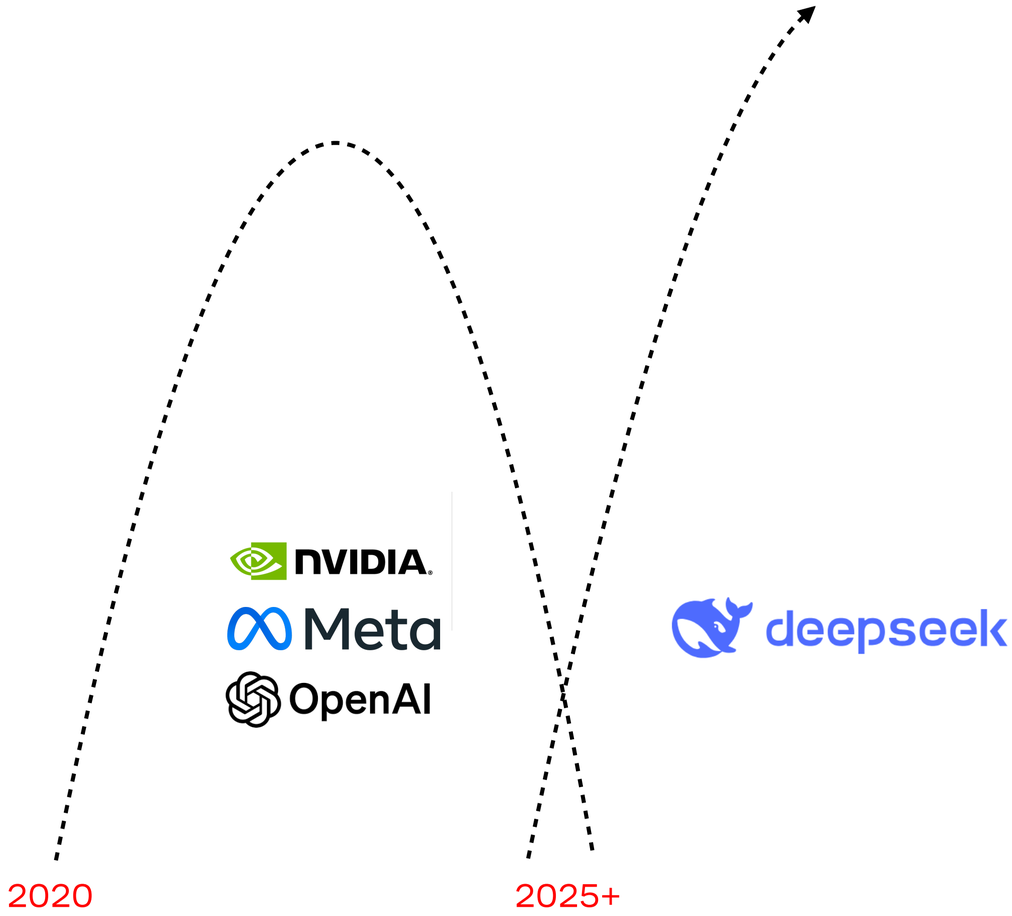
- Китайская модель DeepSeek R1 демонстрирует выдающуюся эффективность: она достигает уровня ChatGPT-4o, но потребляет в 20-30 раз меньше вычислительных мощностей на этапе обучения. Кейс ставит под угрозу доминирование американских AI-компаний и уже повлиял на фондовый рынок США, обвалив акции крупных игроков — например, NVIDIA.
- DeepSeek умеет работать с текстом, решать математические задачи, писать код и доступна бесплатно — в отличие от коммерческих аналогов. Однако модель подвергается цензуре на политические темы. Доступны две версии: 7B (4,7 ГБ) и 671B (более 400 ГБ), а также API для разработчиков.

Источник: tbank, reuters, rg.ru

Прорыв DeepSeek — это начало новой фазы мировой гонки к AGI или очередные завышенные ожидания?

Прорыв DeepSeek — это не достижение одной компании или страны. С выходом новой LLM в открытый доступ улучшения в моделях (алгоритмах и методах обучения) становятся прорывом для всей мировой AI-индустрии.

Теперь эти улучшения доступны каждому, и их невозможно «откатить назад». Никому уже не придётся изобретать это заново. Инновации быстро распространятся и станут вторым скачком прогресса в развитии AI.



Ключевые тренды GenAI 2025

(февраль 2025)

1. Копилоты / Workflow-агенты → автономные AI-агенты → Multi-Agent Systems
2. GenAI с применением RAG становится базовой архитектурой
3. LLM: гонка продолжается — теперь не только в больших, но и в малых языковых моделях (SLM)
4. Модели самообучения могут удешевить процесс создания нейронных сетей
5. Данные как продукт: рост маркетплейсов данных и основанных на них агентов

Ключевые тренды GenAI 2025

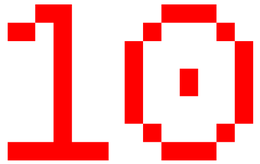
(февраль 2025)

AI-Driven UX —
переосмысление
пользовательского опыта

Массовое внедрение
AI-агентов и копилотов
в физические устройства

Развитие
AI Governance Platforms

Железо: переход к гибридным
и энергоэффективным вычислениям,
трансформация архитектуры



2025: год, когда синтетические
данные станут мейнстримом

Рынок AI движется в сторону агентного подхода: автономные AI-агенты, мультиагентные системы (MAS) и агенты пользовательского интерфейса

Тренд 1: Multi-Agent Systems

Сейчас мы находимся на этом этапе:
2025 год станет переходом к агентскому AI

Копилоты / Workflow-агенты

Human + AI CoPliots

Высокоспециализированные AI-помощники — работают в строго определённой среде, действуют согласно чётко установленным правилам и ограничениям, выполняют конкретные функции и не выходят за пределы своей специализации.

Автономные AI-агенты

Системы AI-агентов

Системы AI-агентов — выполняют многоступенчатые задачи, принимают решения и предпринимают действия для достижения целей при ограниченном или полном отсутствии прямого вмешательства человека.

Также они могут объединяться в гибкие рабочие процессы, где каждый агент обладает своей специализацией, способствующей достижению общих целей.

MAS и UI-агенты

Мультиагентные системы

Самоорганизующиеся системы — состоят из множества AI-агентов, которые сотрудничают, конкурируют или ведут переговоры для достижения общих целей. MAS могут создавать собственные рабочие процессы без заранее заданных сценариев.

UI-агенты

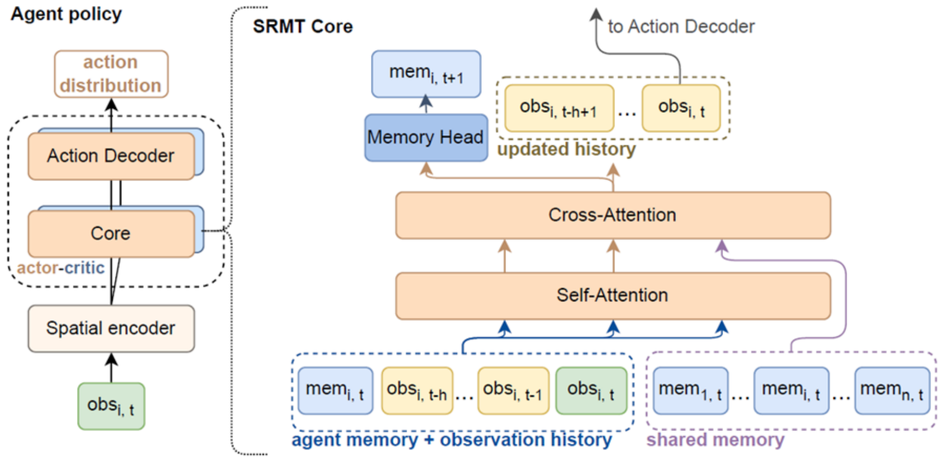
Автономные агенты — способны взаимодействовать с любыми пользовательскими интерфейсами, имитируя реальные действия человека при работе с устройствами.

Что дальше?

Возможно, это будет суперинтеллект или физический AI

Тренд 1: Multi-Agent Systems

Обмен памятью AI-агентов — ещё один шаг на пути к MAS



Архитектура трансформатора общей рекуррентной памяти

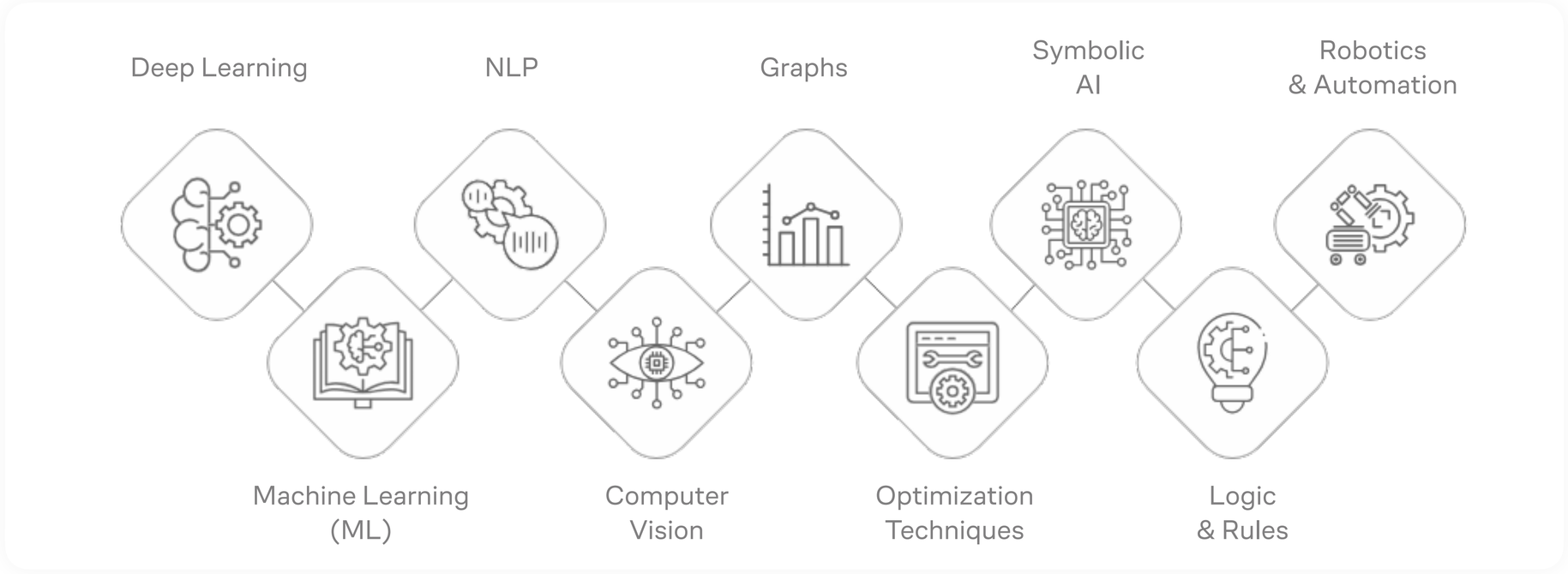
Shared Recurrent Memory Transformer (SRMT) черпает вдохновение из работы человеческого мозга. Подобно тому, как разные части мозга обмениваются информацией через «глобальное рабочее пространство», SRMT позволяет агентам обмениваться данными через общее пространство памяти.

- У каждого агента есть своя «личная» память, которую он обновляет на основе собственных наблюдений.
- Затем эти воспоминания объединяются в общее пространство.
- Все агенты получают доступ к общей памяти, что позволяет им учитывать глобальный контекст при принятии решений.

Источники: community.datascience

Тренд 1: Multi-Agent Systems

Composite AI — новая унифицированная архитектура



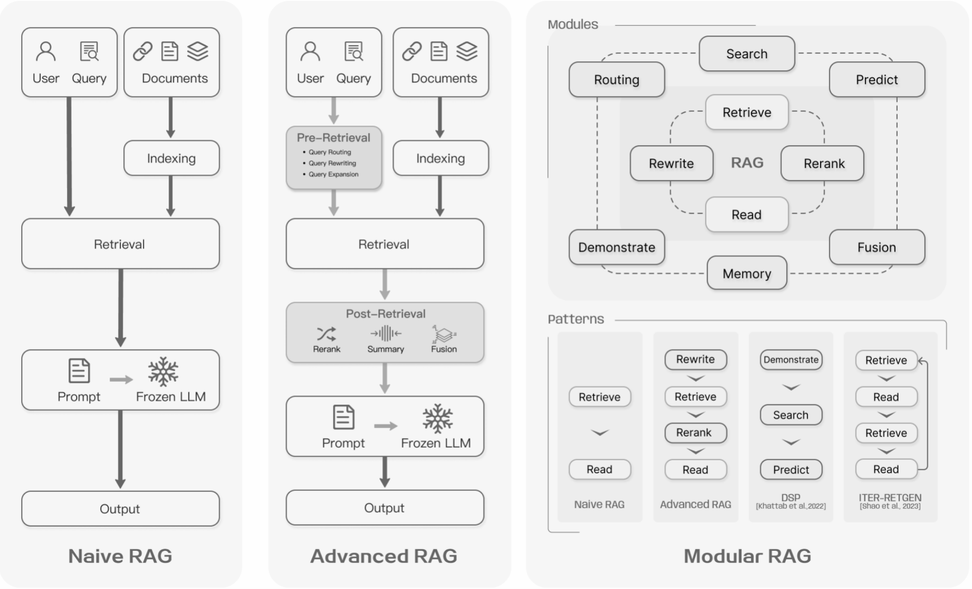
Внедрение AI в бизнес показывает, что подходы с использованием одного метода (одной LLM) часто не справляются с решением сложных и многогранных задач. Композитный AI устраняет эти ограничения, интегрируя такие технологии, как ML, NLP и анализ данных, в единую структуру для создания более эффективных и интегрированных решений.

В 2024 году производитель электроники Fujitsu опубликовал white paper о композитном AI. Документ описывает систему, которая анализирует бизнес-проблемы, автоматически выбирает и объединяет соответствующие модели и данные, а затем предлагает конкретные решения. В компании утверждают, что система может прогнозировать сбои в ПО, планировать маршруты водителей и оптимизировать размещение контейнеров в порту.

Источники: leewayhertz, medium, dzen, Fujitsu Composite AI

Тренд 2: RAG становится базовой архитектурой

RAG становится базово применимой концепцией для LLM и продолжает эволюционировать



В 2024 году RAG (Retrieval-Augmented Generation) подтвердил свою эффективность, а в 2025-м останется основным методом для поиска и генерации информации. Разнообразие архитектур RAG растёт, предлагая новые оптимизации и улучшения.

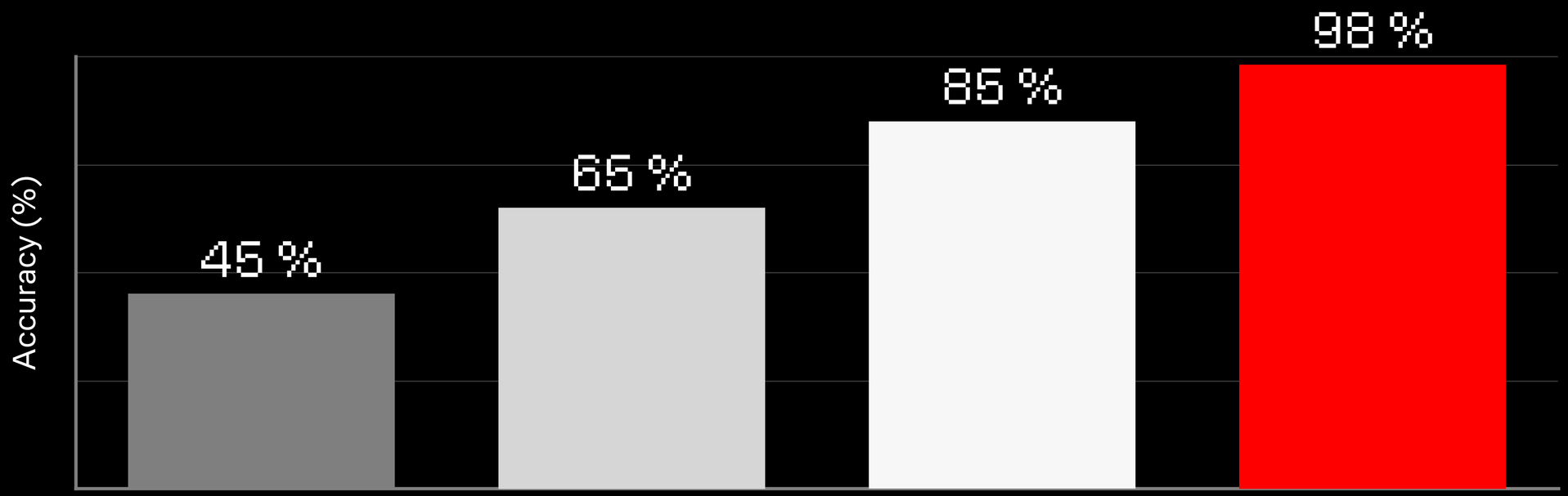
В январе 2025 года появилось исследование, расширяющее концепцию RAG на видеоконтент, что открывает новые возможности для мультимодальных моделей.

Однако распространение Vision-Language Models (VLM) — продвинутых AI-моделей, работающих одновременно с текстом и изображениями — может изменить баланс технологий в этой сфере.

Источник: synoptek, arxiv.org

Тренд 2: RAG становится базовой архитектурой

Увеличивается многообразие инструментов для повышения точности работы RAG-систем



- Retrieval with cosine similarity

- HyDE retrieval
- FT Embeddings
- Chunk/embedding experiments

- Reranking
- Classification step

- Prompt engineering
- Tool use
- Query expansion

Источник: synoptek, arxiv.org

Тренд 3: Развитие языковых моделей

Развитие малых специализированных языковых моделей (SLM)

Аспект	Большие языковые модели (LLMS)	Малые языковые модели (SLMS)
Данные для обучения	Обширные и разнообразные наборы данных	Меньшие наборы данных для конкретной предметной области
Вычислительные требования	Высокий уровень (требуются графические процессоры / TPU, распределенные системы)	Ниже (может работать на менее мощном оборудовании)
Контекстуальное понимание	Отлично, сохраняет контекст на протяжении длинных переходов	Ограниченные возможности, проблемы с долгосрочными контекстными задачами
Производительность	Превосходно справляется со сложными и разнообразными задачами	Подходит для конкретных, сфокусированных задач
Скорость логического вывода	Медленнее из-за большого размера модели	Быстрее за счет меньшего размера модели
Развертывание	Требуется значительная инфраструктура	Более приемлемо для периферийных устройств и мобильных приложений
Универсальность	Высокая универсальность для различных приложений	Более специализированные или сфокусированные приложения
Стоимость обучения	Высокий уровень из-за вычислительных требований и больших наборов данных	Ниже из-за меньшего размера и более простого обучения
Качество генерации контента	Высококачественная, согласованная и детализированная	Подходит для более простых задач или конкретных контекстов
Решение предвзятых и этических проблем	Значительный риск создания предвзятого или неподходящего контента	Меньший риск, но по-прежнему важно управлять
Типичные приложения	Диалоговые агенты, создание контента, научные исследования	Техническая поддержка, чат-боты для конкретного домена, мобильные приложения

SLM (Small Language Models) — это компактные нейросети, оптимизированные для узкоспециализированных задач. Они содержат меньше параметров, но обеспечивают высокую эффективность за счет адаптации к конкретным доменам.

Тренд на самообучающиеся модели ускорит развитие и внедрение отраслевых SLM, снижая затраты на их создание.

По прогнозу Gartner, к 2027 году более 50% моделей GenAI, используемых в бизнесе, будут адаптированы под конкретные отрасли или бизнес-функции.

Источник: Gartner

Тренд 4: Модели самообучения

Модели самообучения могут удешевить процесс создания нейронных сетей

Самообучающиеся модели — ключевой тренд 2025 года. В январе была представлена DeepSeek-R1, созданная с использованием обучения с подкреплением (Reinforcement Learning, RL). Подход позволил снизить затраты на разработку более чем в 40 раз — всего $12 млн против $500 млн у OpenAI.

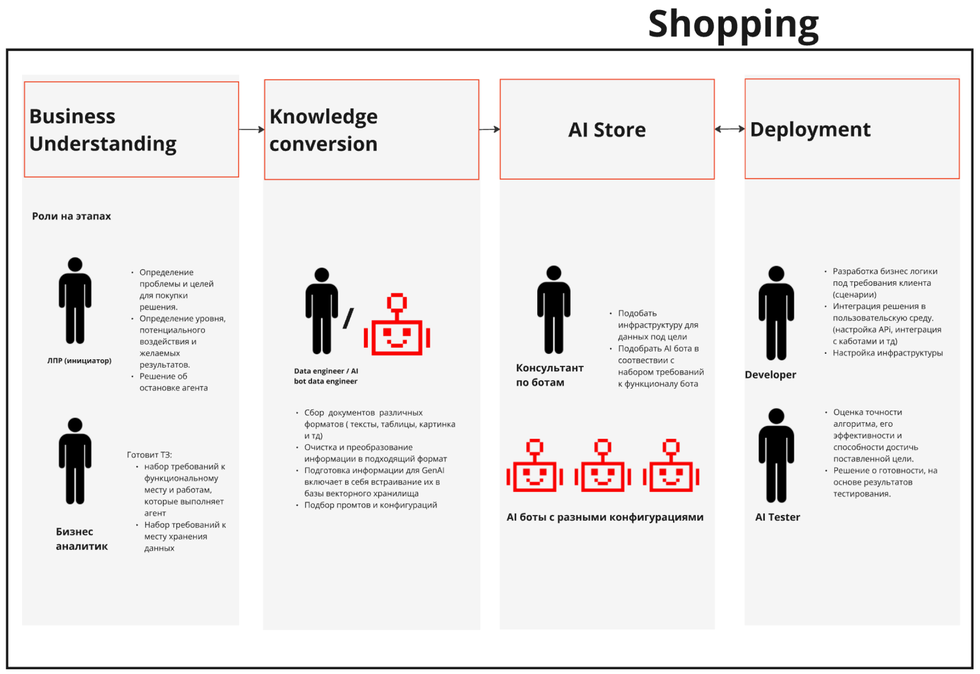
DeepSeek-R1 — одна из первых open-source моделей, в обучении которой применяется RL. Разработчики использовали несколько ключевых методик: длинные цепочки рассуждений, промежуточные высококачественные модели и самосовершенствование через RL, аналогично эксперименту DeepSeek-R1-Zero.

Самообучающиеся модели трансформируют рынок, снижая стоимость AI-услуг. Новые методы, такие как контекстное самообучение (ContextSSL), позволяют нейросетям адаптировать представления к разным задачам. Развитие RL ведёт к появлению AI-систем, способных к самосовершенствованию, что становится важным шагом на пути к AGI.

Источники: synoptek, techxplore, digitalbusiness.kz

Тренд 5: Данные как продукт

Данные как продукт: рост маркетплейсов данных и основанных на них агентов



Бизнес уже осознал ценность данных для оптимизации процессов, но теперь они становятся полноценным продуктом с выделенными командами и стратегиями монетизации. Директора по данным (CDO) будут не только управлять информацией, но и превращать её в источник дохода, интегрируя в бизнес-стратегию и создавая новые AI-решения.

Компании стремятся разрабатывать доменно-специфичных AI-агентов на основе уникальных данных, а рынок движется к формированию AI-маркетплейсов, где такие агенты будут интегрироваться с платформами разных поставщиков. Уже появляются хабы AI-агентов — например, Slack Agent Hub, объединяющий AI-решения от Salesforce, Adobe, Anthropic, Cohere и Perplexity.

AI-экономика развивается в сторону кооперации и обмена данными, что открывает новые модели дистрибуции, включая подписку на AI-экспертов и AI-сотрудников.

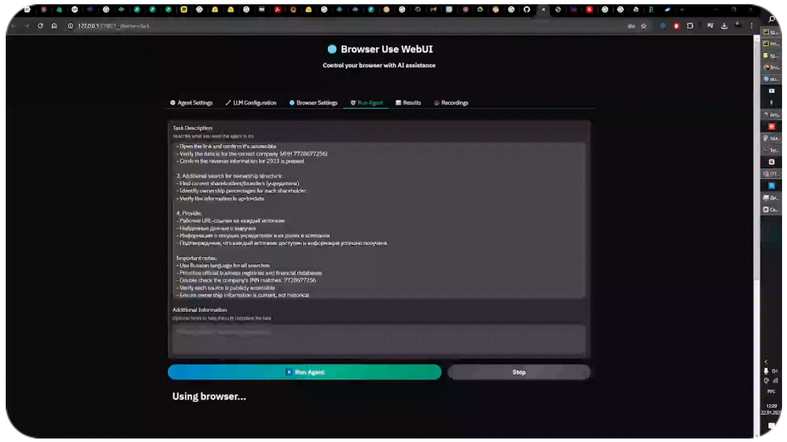
Источник: McKinsey, slack

AI-driven UX — переосмысление пользовательского опыта

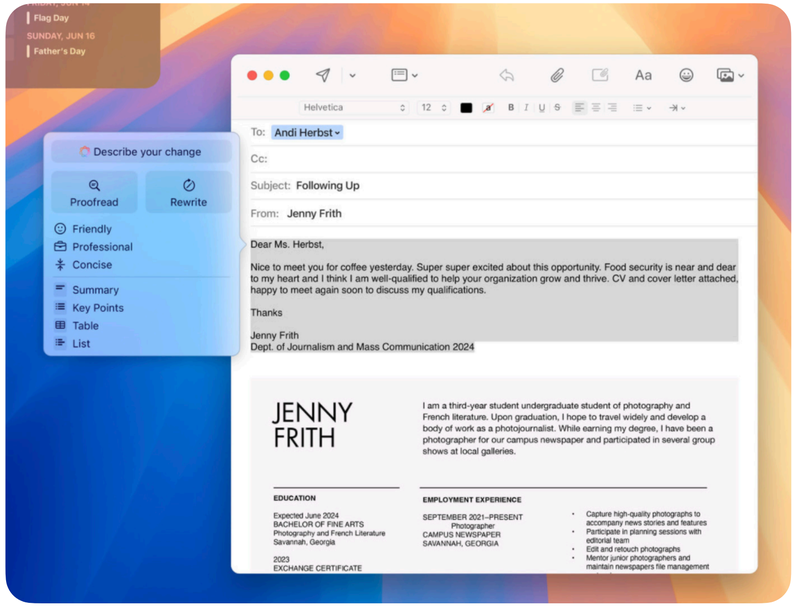
Тренд 6: AI-Driven UX

С развитием AI традиционные подходы к дизайну цифровых продуктов трансформируются — появляется всё больше новых форм взаимодействия с интерфейсом.

↑ Memories создаёт видеоролик из фотографий на основании текстового описания и ключевых слов



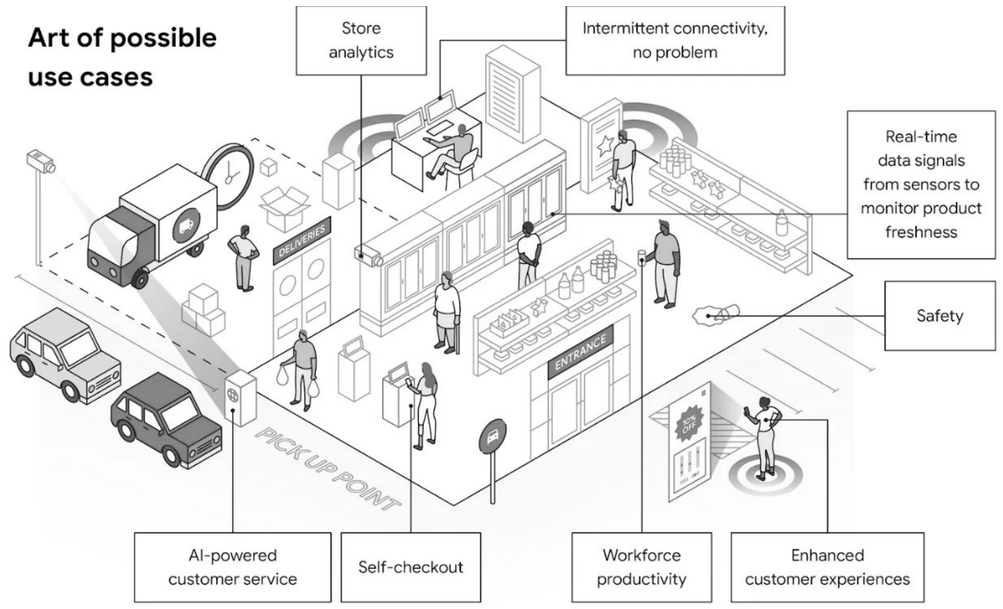
← OpenAI запустила AI-агента Operator, который самостоятельно выполняет задачи в браузере вместо пользователя



← Пользователи могут переписывать текст, проверять данные, находить источники в любом месте — в редакторе, браузере

Тренд 7: AI-агенты и копилоты в физических устройствах

AI-агенты и копилоты массово внедряются в умные устройства



По прогнозам Deloitte, в 2025 году доля смартфонов с поддержкой GenAI превысит 30% — AI-смартфоны станут следующим технологическим трендом, который индустрия предложит массовому потребителю. Китайский бренд Nubia уже делает шаг в этом направлении, интегрируя чат-бота DeepSeek в свои устройства.

Рынок ПК активно развивает связку копилот + ПК. По данным Gartner, в 2025 году глобальные поставки таких устройств достигнут 114 млн единиц, а к 2026 году AI-ноутбуки станут стандартом для крупных предприятий.

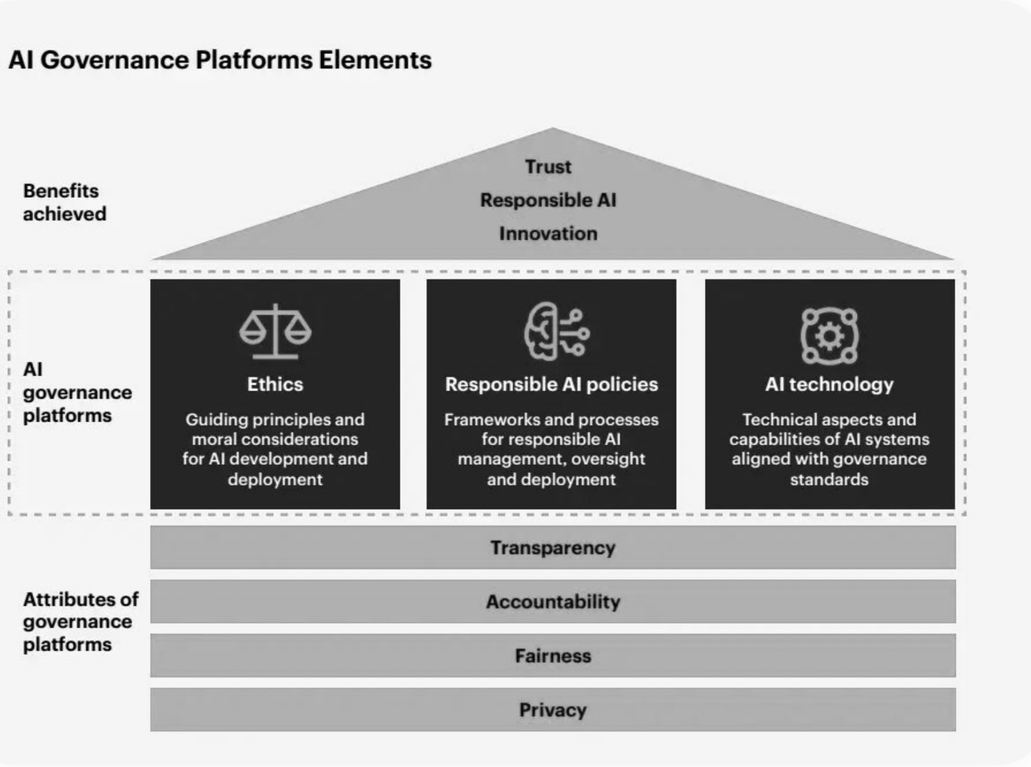
Есть вероятность, что индустрия представит массовое AI-First устройство, исправляя ошибки, допущенные в 2024 году с AI Pin и Rabbit R1.

Google Cloud и Deloitte объединяют усилия, чтобы помочь ритейлерам применять технологии периферийных вычислений через Google Distributed Cloud Edge. Так они смогут запускать AI-приложения прямо в магазинах, анализировать данные в реальном времени, автоматизировать управление запасами и ещё больше персонализировать клиентский опыт.

Источник: Gartner, Medium, Google , Deloitte

Развитие AI Governance Platforms

Тренд 8: Развитие AI Governance Platforms



С развитием AI вопросы управления и этики становятся всё более актуальными. AI Governance Platforms помогают компаниям контролировать юридические, этические и операционные аспекты работы AI-систем.

По данным Gartner, компании, внедрившие такие платформы, могут сократить количество этических инцидентов (судебных исков, PR-кризисов, внутренних конфликтов) на 40% по сравнению с теми, кто этого не сделал.

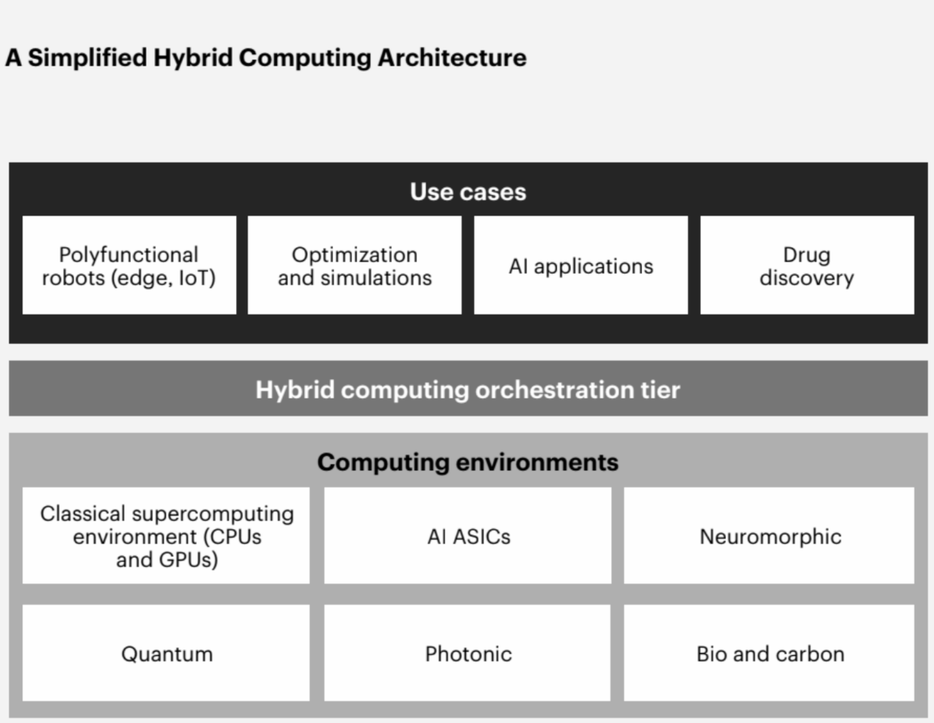
Рынок AI Governance активно развивается, и стартапы в этой области привлекают серьезные инвестиции. Например, Credo AI летом 2024 года привлекла $21 млн в раунде B, а Enzai AI осенью 2023 года получила $4 млн.

Этот тренд широко обсуждается в США и Европе, но в России пока остаётся на периферии внимания.

Источники: synoptek, Gartner

Тренд 9: Гибридные архитектуры

Гибридные и энергоэффективные вычисления



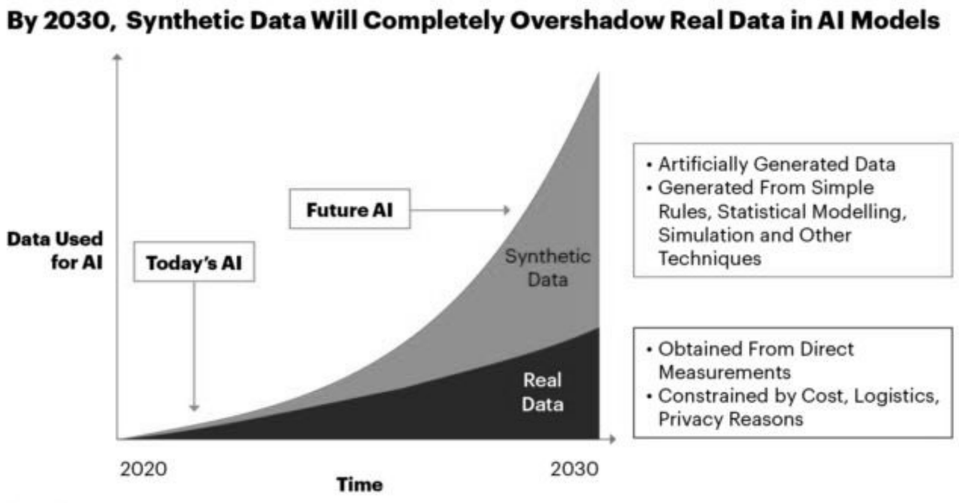
Современные вычисления переходят к гибридным архитектурам, которые объединяют ЦПУ, ГПУ, устройства периферийных вычислений, а также нейроморфные, квантовые и фотонные системы. Такой подход позволяет использовать сильные стороны каждой технологии, повышая производительность и снижая энергопотребление.

Оптимизация AI-моделей становится ключевой задачей: новые вычислительные решения позволяют значительно сократить энергозатраты на обучение и инференс, что снижает углеродный след. По прогнозам Gartner, к 2028 году 30% реализаций GenAI будут использовать энергоэффективные вычисления в рамках стратегии устойчивого развития.

Источники: Gartner

Тренд 10: Синтетические данные

Синтетические данные станут мейнстримом



Синтетические данные необходимы в случаях, когда реальные данные слишком дороги, труднодоступны или ограничены из-за требований конфиденциальности. Они позволяют компаниям безопасно разрабатывать и тестировать AI-модели, не нарушая регуляторные нормы.

По прогнозу Gartner, к 2026 году 75% компаний будут использовать GenAI для генерации синтетических клиентских данных. В России для повышения доступности, безопасности и качества данных Ассоциация больших данных (АБД) совместно со Сбером и другими участниками разработала проект национального стандарта синтеза данных.

Документ содержит математические доказательства, подтверждающие, что при соблюдении стандартов возможно генерировать данные без риска нарушения конфиденциальности.

Источник: Gartner, gretel, rubda

Основные тезисы

Оценка реальных результатов GenAI

2025 год станет годом подведения итогов — компании публикуют первые реальные кейсы внедрения генеративного AI в бизнес-процессы.

Данные как стратегический актив

Компании переходят к data-driven культуре: нанимают директоров по данным (CDO), пересматривают операционные процессы и разрабатывают новые стратегии монетизации данных.

RAG становится стандартом, но требует доработки

Retrieval-Augmented Generation (RAG) остается основным инструментом для работы с неструктурированными данными, но его ограничения стимулируют поиски более эффективных решений.

AI-маркетплейсы и агентные системы

Формируются экосистемы мультимодальных и композитных AI-агентов, способных выполнять задачи, отслеживать их выполнение и корректироваться — по аналогии с junior-специалистами.

Тонкая настройка AI пока недооценена

Модели, оптимизированные под конкретные задачи, значительно повышают производительность, но пока лишь 9% предприятий используют этот инструмент. В 2025 году ожидается рост числа узкоспециализированных SLM.

Оптимизация алгоритмов vs рост мощностей

Будущее AI — за эффективными алгоритмами, а не безграничным увеличением вычислительных мощностей. Это создаёт новые конкурентные возможности и ставит под сомнение долгосрочное доминирование NVIDIA (пример — новая LLM DeepSeek).

Влияние на рынок труда

На смену Digital Natives приходит поколение AI Ready

Digital Natives

Поколение людей, родившихся в эпоху интернета, использующих смартфоны с юных лет.

AI Ready / Generation AI

Дети и молодые люди, выросшие в эпоху развития интеллектуальных устройств, голосовых помощников, алгоритмов и других технологий, основанных на AI.

Текущее молодое поколение будет активнее работать с AI

До 2030 года количество AI-инструментов будет увеличиваться на 40% в год. AI станет технологией, к которой необходимо адаптироваться.

AI-сервисы — это новый Word.

Дети и подростки используют GenAI чаще взрослых

Почти 70% пользователей генеративного AI — это зумеры и миллениалы. При этом 52% зумеров доверяют технологиям, которые помогают им принимать обоснованные решения.

Молодое поколение видит пользу в AI-инструментах

68% студентов считают, что GenAI помогает им лучше усваивать новую информацию, а его регулярное использование позволяет экономить в среднем 5,3 часа в неделю.

Источники: Qustodio, Salesforce, Deloitte, Microsoft, Arthur D.Little

Вместе с поколением AI Ready трансформируется и рынок труда

Поднимается планка для сотрудников

12% рекрутеров в мире уже создают должности, которые будут требовать AI-навыков.

«Руководитель AI-подразделения» становится обязательной позицией в компании — за 5 лет число должностей увеличилось на 28% к 2023 году.

66% руководителей хотят нанимать нетехнических специалистов с AI-навыками.

Число объявлений о вакансиях, связанных с AI растёт в 3,5 раза быстрее, чем в других направлениях.

В LinkedIn по всему миру наблюдается 142-кратное увеличение числа пользователей, которые добавили в профиль наличие AI-навыков.

Места, требующие AI-навыков, предлагают до 25% выше зарплаты, что подчёркивает ценность таких специалистов для компаний.

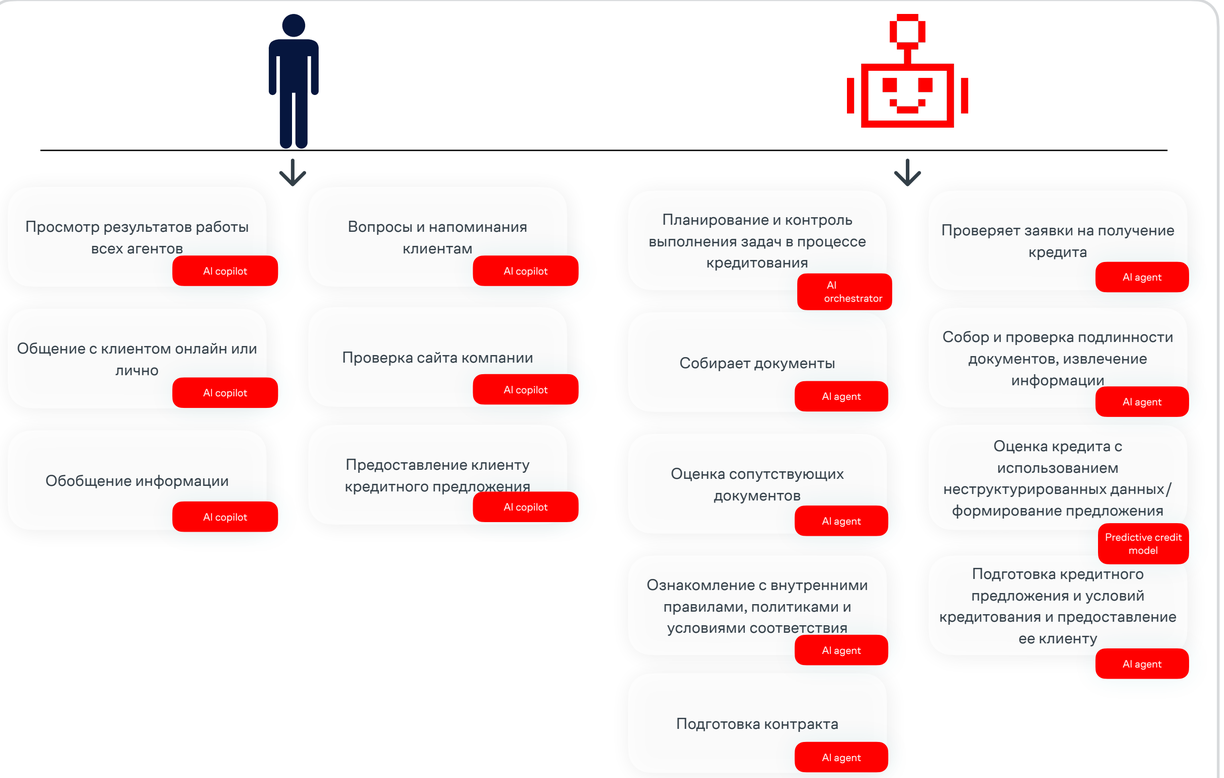
Источники: Microsoft, PwC

AI-агент не заменит человека, а дополнит. Технология требует обновления навыков

Ручная оркестровка и исполнение

Кредитный менеджер обрабатывает задачи вручную, в производство необходимо вовлечь больше специалистов.

Совместно с AI агентом



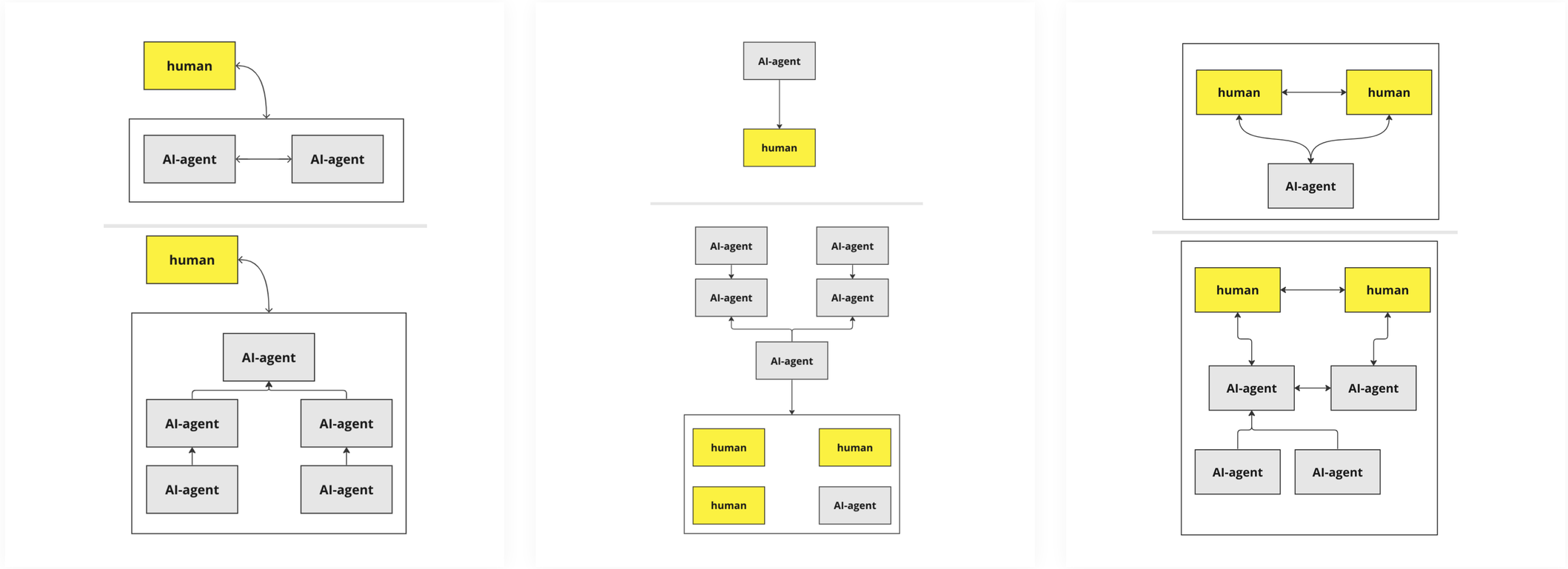
AI справляется со многими задачами самостоятельно, а там, где нет возможности автоматизировать исполнение функции агентом, он выполняет роль копилота, забирая на себя часть работы кредитного менеджера.

Конкуренция возникает не между людьми и машинами, а между теми, кто эффективно использует AI, и теми, кто этого не делает. Часть задач передается ботам, что меняет структуру труда и перераспределяет работу.

Появляются новые требования к сотрудникам, но одновременно и новые возможности: AI-навыки становятся ценным преимуществом. Несмотря на автоматизацию, безработица не должна вырасти, так как появляются новые профессии. Однако большинство специальностей потребуется адаптировать, а сотрудников — обучать работе в среде с AI-агентами.

Источник: Mckinsey

Мы сможем признать AI-агентов полноценными сотрудниками

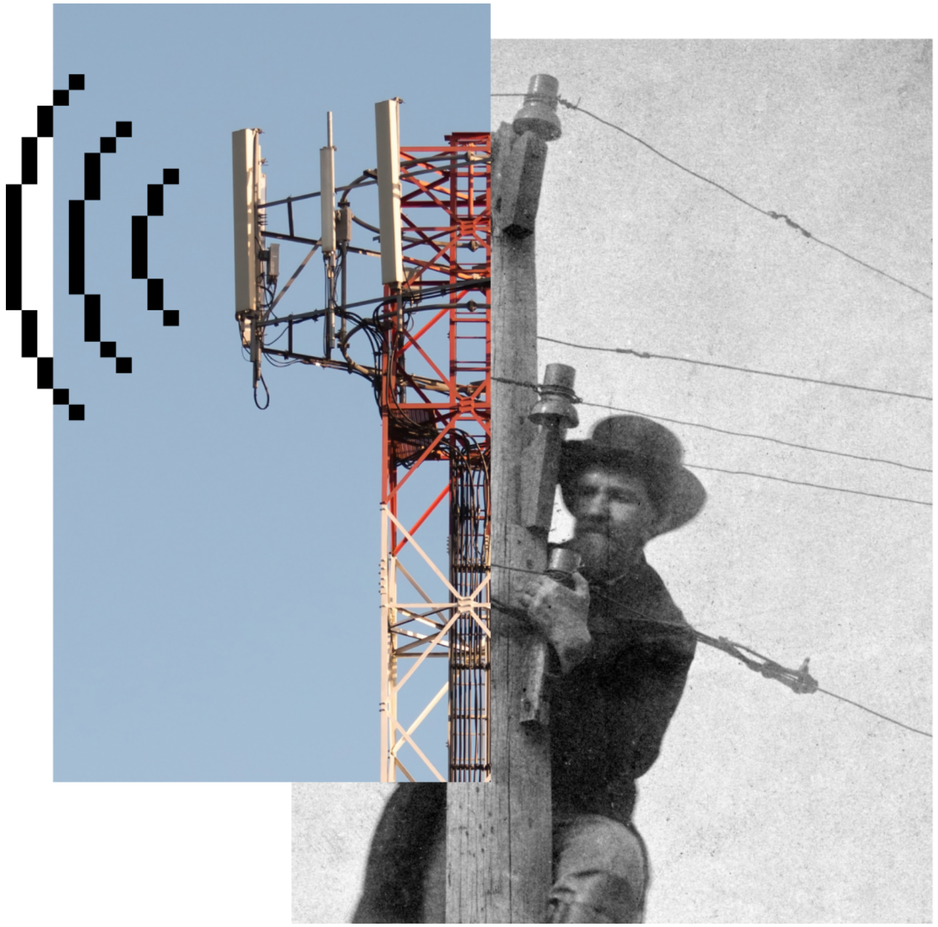


Варианты взаимодействия человека и AI-агентов

Будет разработана система, которая позволит официально признать AI-агентов полноценными сотрудниками. Эта система охватит весь спектр HR-процессов: от найма и обучения до постановки целей, оценки производительности и других аспектов.

В результате сформируется новая многоуровневая система управления, включающая чередующиеся уровни контроля: людей над людьми, людей над машинами, машин над людьми и машин над машинами. Как отмечает a16z, у каждой второй должности белого воротничка появится копилот, а затем и AI-агент.

Неврологическое улучшение

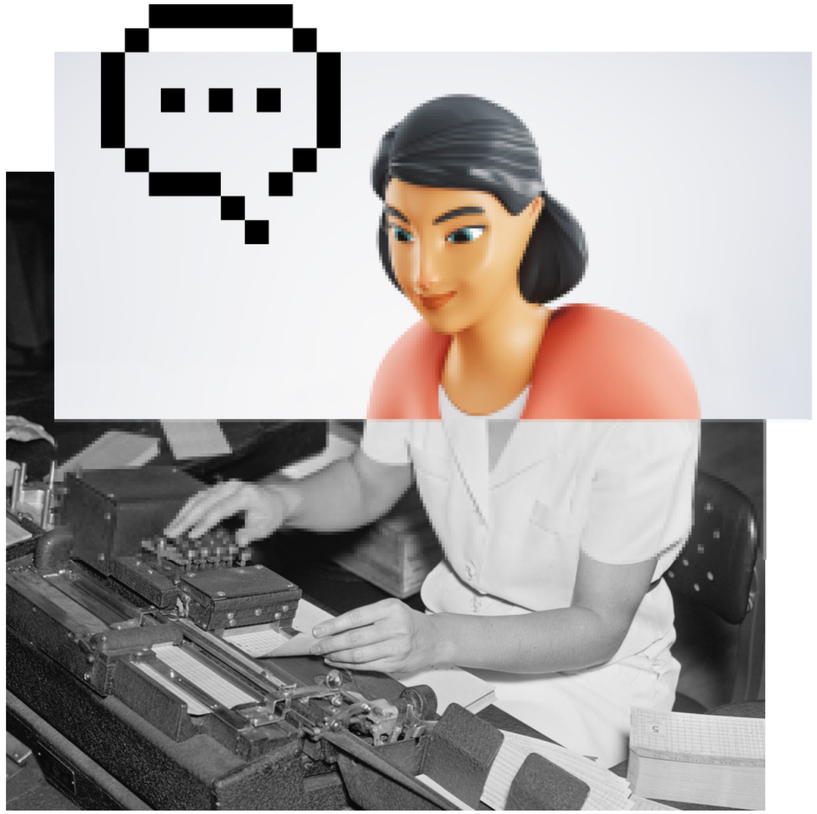


Неврологические усовершенствования, такие как двунаправленные интерфейсы мозг–машина (BBMI), направлены на улучшение когнитивных способностей. По данным Gartner, к 2030 году 60% IT-работников будут использовать эти технологии.

Интерфейсы анализируют мозговую активность и стимулируют нужные состояния: концентрацию или расслабление. Применение включает ускоренное обучение, улучшение найма и повышение производительности. В медицине это может сокращать срок стажировки хирургов, а в производстве — снизить риски несчастных случаев.

Источник: Gartner

Как не заработать синдром FOMO* в дивном новом мире?



AI-сервисы трансформируют привычные рабочие процессы, особенно в сферах IT, продаж, маркетинга и клиентского сервиса. Освоение новых компетенций становится необходимым, чтобы сохранить конкурентное преимущество и оставаться востребованным на рынке.

В сферах с высокими рисками и в ручном труде изменения происходят медленнее, но уже сейчас копилот-решения помогают оптимизировать процессы и повышать эффективность работы.

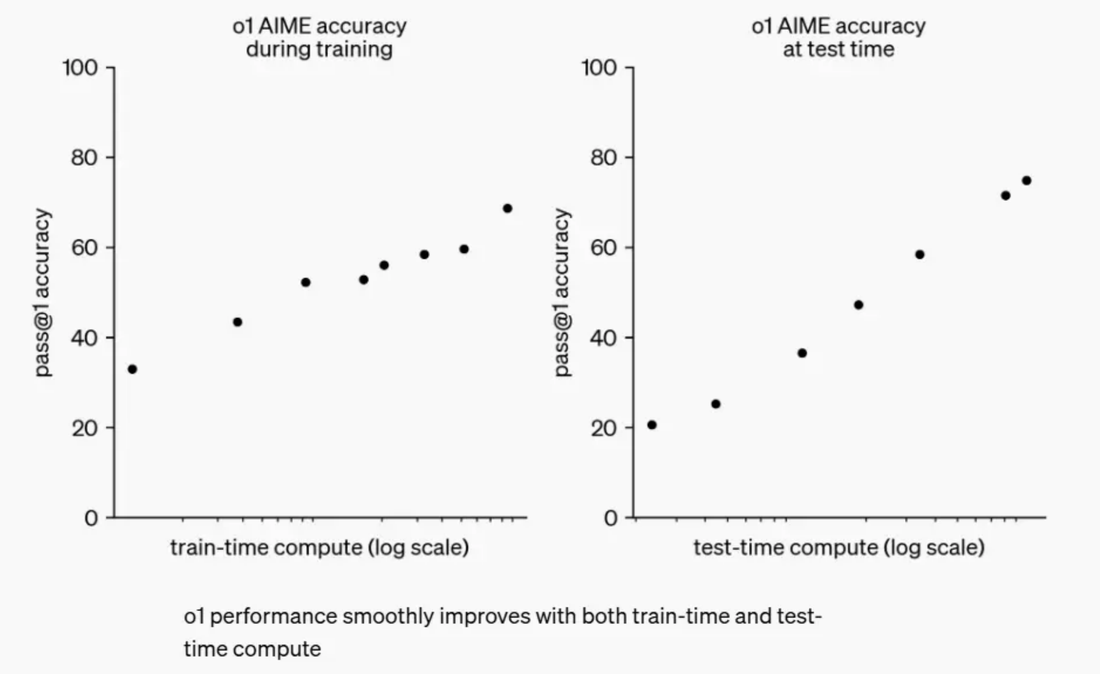
Остается вопрос — как ничего не упустить?

*FOMO — страх упущенных возможностей

Эксперименты в GenAI:

Новые технологические направления,
которые могут принести эффект в ближайшем будущем

Масштабирование вывода (Test-time compute)



o1 performance smoothly improves with both train-time and test-time compute

Оптимизация вычислений в момент вывода (inference) становится ключевым фактором повышения производительности языковых моделей. В ряде случаев адаптивное масштабирование вычислений в тестовое время оказывается более эффективным, чем увеличение количества параметров модели.

Так, OpenAI O1 тратит больше времени на размышления перед формированием ответа, оценивая оптимальный путь решения задачи, а не просто генерируя быстрый ответ. Это улучшает точность модели при обработке сложных многоступенчатых запросов, требующих продвинутых стратегий рассуждения.

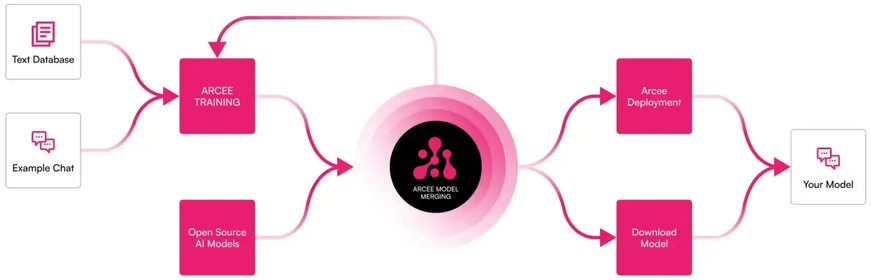
Китайский AI-стартап Moonshot AI представил мультимодальную модель Kimi K1.5, использующую обучение с подкреплением (RL) и две стратегии Chain of Thought (CoT):

- Long-CoT — предназначен для сложных задач, требующих глубокого анализа.
- Short-CoT — оптимизированная версия для повседневных запросов.

На данный момент технология не получила широкого бизнес-применения: декомпозиция задачи перед отправкой запроса остается более выгодной экономически из-за высокой стоимости токенов.

Источники: github, openai, synoptek

AI-стартапы скрещивают модели для получения «более сильной модели»



arcee.ai

Современные AI-стартапы находят новые подходы к разработке мощных моделей, комбинируя уже существующие решения. Вместо обучения моделей с нуля компании используют техники объединения (model merging), позволяя создавать более сильные и специализированные AI-системы за счёт слияния архитектур и возможностей разных моделей.

Например, Sakana AI представила метод Evolutionary Model Merge, который автоматически комбинирует open-source модели, используя эволюционные алгоритмы. Это позволяет собирать Foundational Models, настраивая их под конкретные задачи. В отличие от традиционного обучения, такой подход ускоряет создание мощных моделей и снижает затраты.

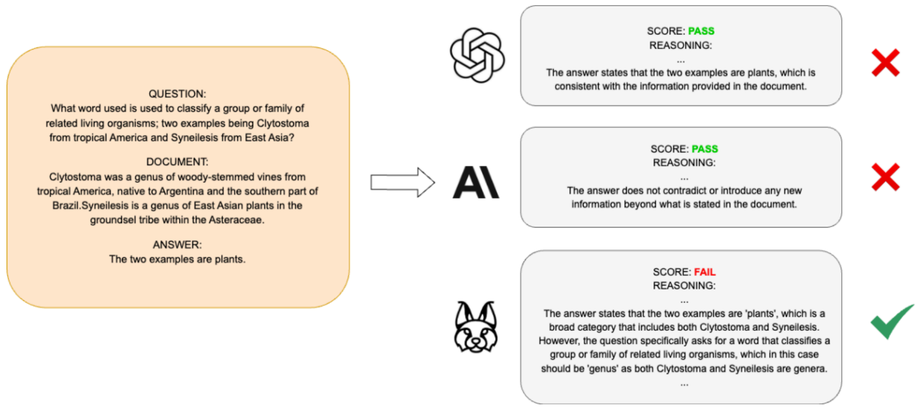
Другой пример — Arcee, предлагающая корпоративным клиентам услуги обучения и развертывания пользовательских SLM. Их технология Model Merging объединяет сильные стороны разных моделей в одну, а Spectrum оптимизирует обучение, настраивая определенные слои. Это делает AI-модели более эффективными и адаптируемыми к специфическим бизнес-задачам.

Объединение моделей становится новым трендом в развитии AI, позволяя компаниям быстрее создавать мощные системы с гибкими возможностями и улучшенной производительностью.

Источники: Sakana AI, Arcee AI

Создание AI-моделей для обнаружения галлюцинаций

Patronus AI



Lynx: модели с 8B и 70B для обнаружения галлюцинаций RAG

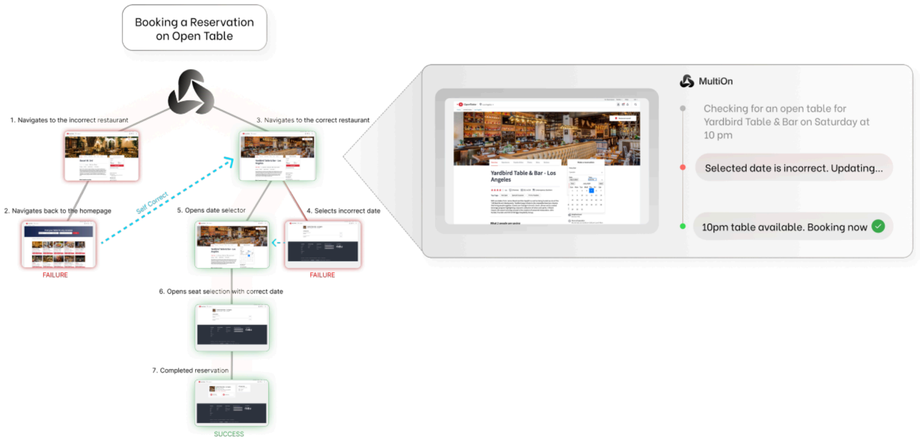
Одна из ключевых проблем GenAI – галлюцинации, когда модели уверенно генерируют неверную информацию. Разработчики активно ищут способы сделать AI более надежным и точным, создавая инструменты для выявления и устранения подобных ошибок.

Так, Llama-3-70B-Instruct была тонко настроена на сложном датасете с акцентом на реальные сценарии запросов, что повысило точность обнаружения галлюцинаций. Для дальнейшего улучшения качества работы модели предложен HaluBench – крупномасштабный датасет, содержащий 15 тыс. маркеров сложных задач, связанных с генерацией недостоверной информации.

Кроме того, создаются фреймворки для повышения надежности и прослеживаемости AI-моделей. Они позволяют системам анализировать собственные ответы и критически оценивать знания, снижая риск распространения дезинформации. Такой подход делает AI более прозрачным и управляемым для бизнеса.

Источники: Nomic AI и Patronus AI, Arxiv.org, Arxiv.org

Новый фронтир → Reasoning-модели



AI выходит за рамки рутинных задач — новый класс reasoning-моделей фокусируется на сложных рассуждениях, многозадачности и улучшенной логике принятия решений. Ведущие компании и исследовательские команды экспериментируют с архитектурами, которые делают модели более умными, адаптивными и способными к самокоррекции.

OpenAI O1 — первая серия reasoning-моделей, ориентированных на решение сложных задач, а не просто автоматизацию рутинных процессов.

G1 (Groq) — экспериментальная инициатива по созданию O1-подобных reasoning-цепочек на базе Llama-3.1 70B, открытая для разработчиков.

Agent Q — объединяет MCTS (Monte Carlo Tree Search) и self-critique, обучая AI-агентов эффективнее работать с веб-навигацией и улучшая их логические способности.

Multi1 — исследует новые стратегии промптинга для повышения способности языковых моделей к логическому мышлению.

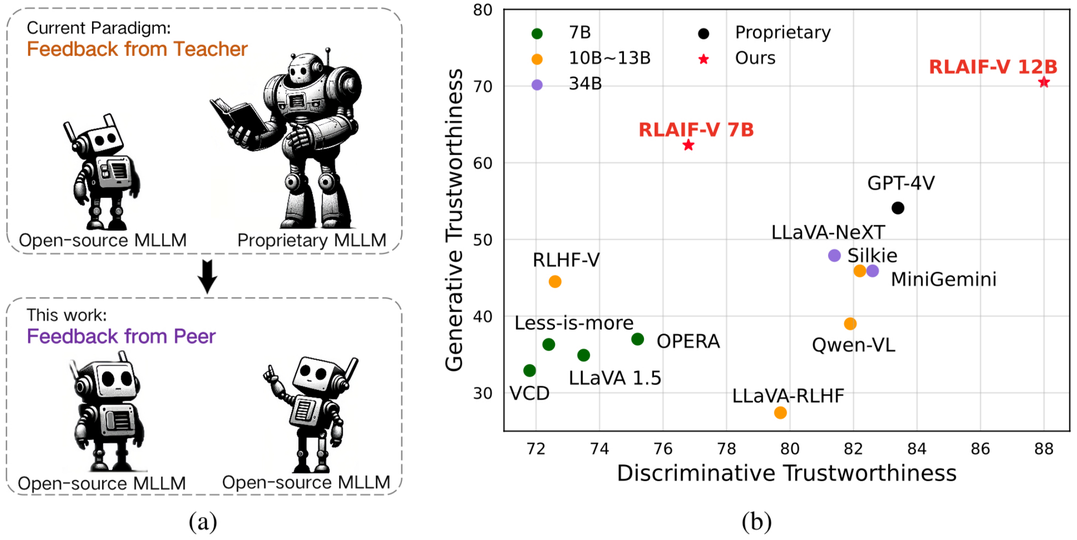
YandexGPT 4 — использует Chain-of-Thought (CoT), удерживает в 4 раза больше контекста и работает в 2 раза быстрее предыдущих версий.

Show-me (GPT-4o-mini) — альтернативный подход к reasoning: оценивает сложность задачи, разбивает её на подзадачи и проверяет правильность каждого шага.

Источники: OpenAI O1, G1, Agent Q, Multi1, YandexGPT 4, Show-me

RLHF → RLAIF: новый уровень обучения AI

Университетские AI-лаборатории



Подход RLAIF-V

Обучение AI выходит на новый этап: Reinforcement Learning with AI Feedback (RLAIF) заменяет RLHF, позволяя моделям учиться на основе обратной связи от других AI, а не только от людей. Процесс обучения становится более автономным и масштабируемым.

📍 RLAIF-V — метод согласования MLLM (например, GPT-4V) с человеческими предпочтениями. AI собирает и анализирует обратную связь через веб-интерфейс, что помогает моделям тонко настраиваться и становиться более надежными.

📍 Метод Meta — файнтюнинг с многослойной самопроверкой:

- Модель оценивает свои ответы, используя CoT.
- Few-shot prompting помогает генерировать новые промпты.
- После трех итераций LLaMA 2-70B обходит ChatGPT-3.5, Claude и GPT-4 (версия 13.07.2023) на AlpacaEval.

📍 RLEF — метод позволяет LLM улучшать генерацию кода, анализировать ошибки на тестах и корректировать их без вмешательства человека.

RLAIF делает обучение более гибким и эффективным, снижает зависимость AI от разметки и ускоряет его развитие.

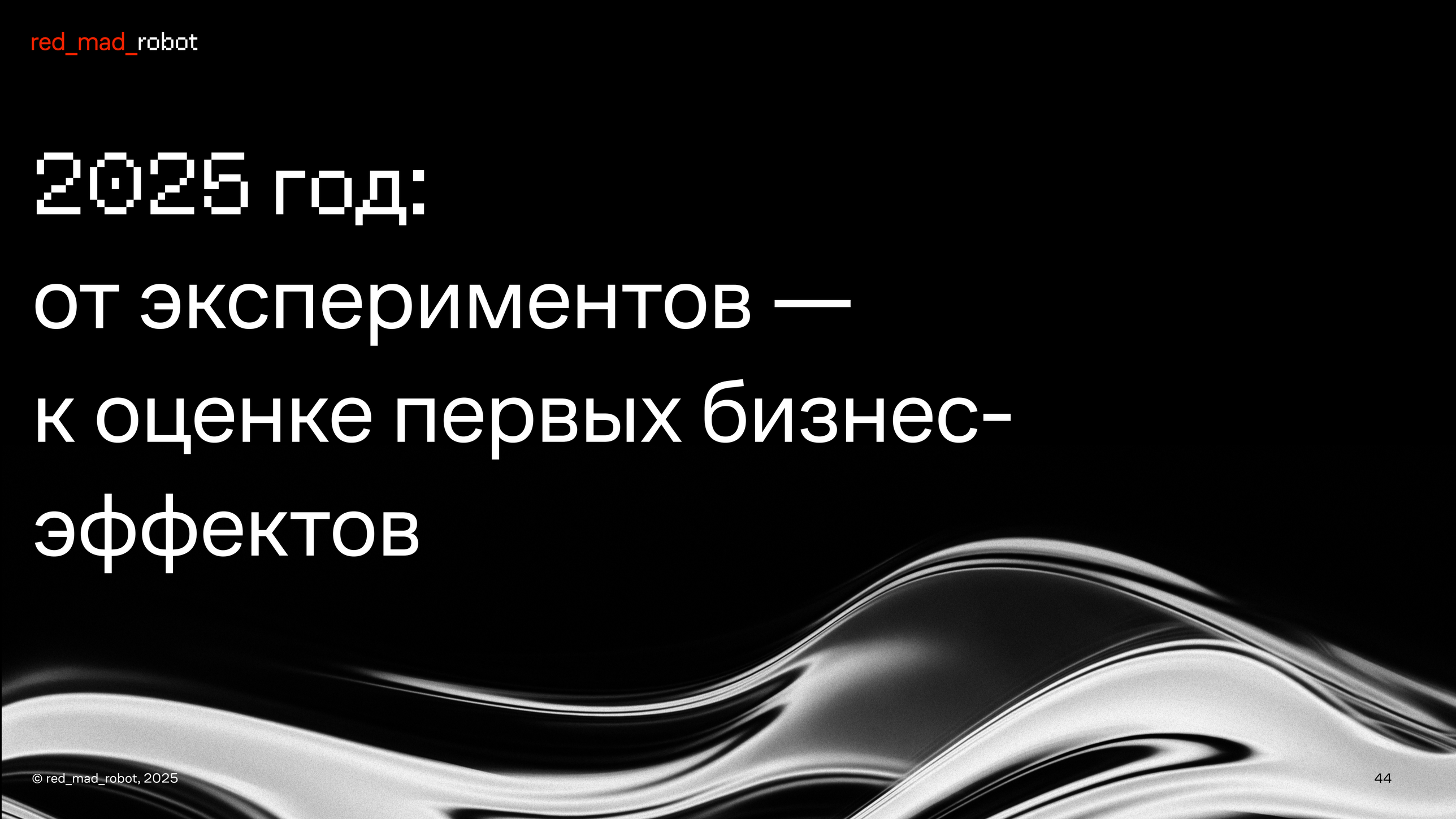
Источники: RLAIF-V, Self-Rewarding Language Models, RLEF

Появляется всё больше подходов к оптимизации обработки длинного контекста в LLM

Работа	Краткое описание
ChatQA 2: Bridging the Gap to Proprietary LLMs in Long Context and RAG Capabilities	NVIDIA представила ChatQA 2 — новую модель на базе Llama 3, которая поддерживает контекст до 128 тыс. токенов. Это значительно больше стандартных 8 тыс. токенов в Llama 3-70B, что позволяет модели давать более точные и детализированные ответы, особенно на сложные вопросы.
Human-like Episodic Memory for Infinite Context LLMs	EM-LLM (Episodic Memory LLM) работает по принципу человеческой памяти, что помогает модели лучше справляться с длинными текстами. Она разбивает информацию на осмысленные эпизоды и при необходимости извлекает из них ключевые данные — так же, как человек вспоминает нужные факты из опыта.
LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs	AgentWrite решает проблему ограничения длины выходных данных, разбивая длинные задачи на меньшие фрагменты. Этот метод позволяет моделям генерировать тексты объемом до 20 тыс. слов, при этом сохраняется логическая связность и структурированность.
LazyLLM: Dynamic Token Pruning for Efficient Long Context LLM Inference	LazyLLM — метод оптимизирует обработку длинных контекстов в LLM за счет динамического выбора и вычисления только релевантных токенов. Позволяет сократить время до генерации первого токена и сохраняет высокую точность модели.
Squid: Long Context as a New Modality for Energy-Efficient On-Device Language Models	Dolphin — архитектура, разработанная для более эффективной работы с длинными контекстами в языковых моделях. Она включает вспомогательную компактную модель, которая сжимает текст, упрощая его перед отправкой в основную систему. Такой подход снижает нагрузку на вычисления, уменьшает энергопотребление и ускоряет обработку, не теряя точности.
Writing in the Margins: Better Inference Pattern for Long Context Retrieval	Метод Writing in the Margins (WiM) помогает моделям лучше справляться с длинными текстами. Он разбивает материал на небольшие фрагменты, а затем создает «заметки на полях» с ключевой информацией. Такой подход особенно полезен в задачах поиска, где важно быстро находить релевантные данные.
ReMamba: Equip Mamba with Effective Long-Sequence Modeling	ReMamba — доработанная версия архитектуры Mamba, которая использует выборочное сжатие и двухступенчатую обработку для работы с длинными текстами. Благодаря этому модель показывает производительность на уровне трансформеров аналогичного размера, но при меньших вычислительных затратах.

Источники: ChatQA 2, EM-LLM, AgentWrite, LazyLLM, Dolphin, WiM, ReMamba

2025 год:
от экспериментов — к оценке первых бизнес-эффектов



Универсальной методики оценки эффективности внедрения GenAI пока нет, но уже можно выделить первые факторы успеха

Стратегическое целеполагание

Определите цели внедрения GenAI, например, повышение продуктивности, улучшение обслуживания клиентов или оптимизация процессов.

Качество результатов

Определите метрики для оценки эффективности: анализ отзывов клиентов, снижение количества ошибок, улучшение точности прогнозов и другие.

Экономия времени и ресурсов

Сравните время, затраченное на выполнение задач до и после внедрения GenAI. Оцените, насколько удалось сократить трудозатраты и другие ресурсы.

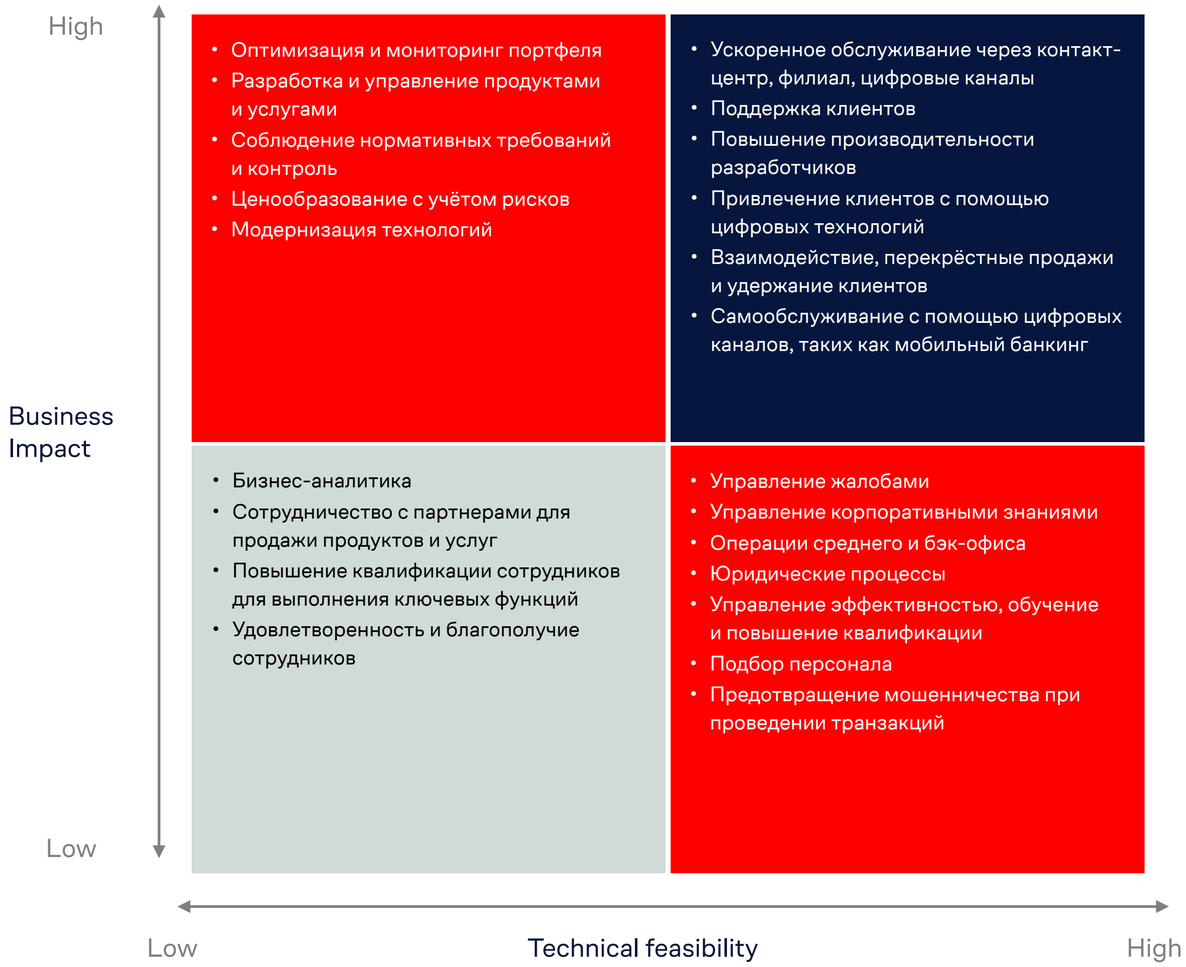
Приоритет на высокочастотные бизнес-функции

Ощутимые результаты от внедрения GenAI достигаются только в масштабе — поэтому стартовать лучше с массовых, легко масштабируемых бизнес-функций.

Источник: analyticsvidhya, unite.a, itweek, dataversity

Приоритеты AI-трансформации

Поддомены на AI-трансформацию
на примере банковского кейса



Для эффективного внедрения AI компаниям необходимо определить, какие поддомены трансформировать в первую очередь. Выбор должен основываться на двух ключевых факторах: бизнес-ценности и технической осуществимости.

Оценка бизнес-ценности

- Определение конкретных выгод от AI-трансформации в каждом поддомене.
- Соответствие решения стратегическим целям компании.
- Готовность конечных пользователей (сотрудников и клиентов) к внедрению AI.

Анализ технической осуществимости

- Доступность и качество данных, включая конфиденциальные и чувствительные сведения.
- Возможность масштабирования решения на другие бизнес-подразделения.
- Повторное использование компонентов AI-решений для других сценариев.
- Совместимость с существующей IT-инфраструктурой и необходимость её модернизации.

Грамотный подход к выбору приоритетных поддоменов позволит компаниям избежать точечных неэффективных экспериментов, обеспечить быструю окупаемость AI-проектов и создать устойчивую основу для дальнейшей трансформации.

Источник: Mckinsey

Знаем всё про новейшие технологии и тренды
на рынке — собираем экспертизу
в аналитические дайджесты и исследования.

redmadrobot.ru
Telegram-канал: @redmadnews