

Лекция 12. Тренды развития LLM. LLM в России

Тренды развития LLM. LLM в России

- Тренды развития LLM
- Сильнейшие LLM 2025
- LLM в России

Мультимодальность

Мультимодальность - это способность одной модели воспринимать и обрабатывать несколько типов данных одновременно. Мультимодальная LLM не просто имеет отдельные модули для картинок и текста — она строит единое внутреннее представление, связывающее смыслы из разных источников.

Ключевые направления:

- Текст + Изображение (Text-to-Image / Image-to-Text) (OpenAI GPT-4Vision), Google Gemini 1.5 Pro, Claude 3)
- Аудио-модальность (OpenAI Whisper, GPT-4, Gemini)
- Видео-модальность (Gemini 1.5 Pro, VideoPoet от Google)

Мультимодальность

Технические подходы к реализации:

- Слияние на уровне эмбеддингов (каждая модальность (текст, изображение) проходит через свой специализированный энкодер, который преобразует данные в числовой вектор)
- Архитектура «Переводчик» (использование модулей, которые "переводят" одну модальность в другую)
- Единая архитектура (одна модель изначально обучается на данных разных типов, используя общие механизмы внимания)

Сближение открытых и закрытых моделей

Разрыв между закрытыми коммерческими моделями (ChatGPT, Claude, Bard) и открытыми моделями с открытым исходным кодом постепенно сокращается. Открытые модели, такие как Code Llama 70B и Mistral AI 8x7B, демонстрируют высокую производительность на стандартных тестах.

Что привело к сближению?

- С скачком качества открытых моделей
- Давление со стороны сообщества и регуляторов
- Потребности рынка: кастомизация и приватность
- Экономика: дороговизна закрытых API

Сближение открытых и закрытых моделей

- Закрытые игроки становятся «более открытыми» (публикация исследований и small-scale моделей, стратегия «Open Weight»).
- Открытые игроки создают «закрытые сервисы» (Meta с моделью Llama 3, Mistral AI зарабатывает, предоставляя доступ к своим же открытым моделям через платное API).
- Появление гибридных и промежуточных форматов («взвешенная открытость»: доступ к весам модели выдается по заявке для исследовательских целей; выкладывается базовая модель, но не версия с тонкой настройкой под человека (RLHF) — самая ценная часть).
- Битва за сообщество и экосистему

Расцвет малых языковых моделей

Высокая стоимость обучения и эксплуатации больших LLM ограничивает их применение. Поэтому разрабатываются малые LLM с меньшим количеством параметров, которые могут выполнять задачи менее ресурсоемко. Примеры: Stable LM (1.6 млрд параметров) и Microsoft Phi-2 (2.7 млрд параметров).

- Закон убывающей отдачи. Увеличение параметров после определенного порога ($\approx 100B$) дает все меньше прироста в качестве на единицу затрат
- Критическая важность данных. Исследования показали, что маленькую модель, обученную на тщательно отфильтрованных, высококачественных, синтетических данных, можно сравнять по качеству с гигантской, обученной на всём подряд
- Спрос на экономичность и скорость
- Возможность специализации
- Локальный запуск

Расцвет малых языковых моделей

Ключевые архитектурные и методические прорывы, сделавшие SLM возможными

- Колossalное улучшение данных (Data Curation & Synthetic Data)
- Передовые архитектуры
- Новые парадигмы обучения

Последствия для рынка

- Демократизация и демонополизация
- Смещение фокуса с «больше параметров» на «лучшая инженерия»
- Размывание границ
- Проблема для бизнес-моделей, построенных на API-платежах за токен

Расцвет малых языковых моделей

- Специализированные семейства SLM: будут появляться не просто общие модели, а целые «семейства» от одного производителя (например, кодовая 4В, математическая 4В, медицинская 4В), оптимизированные под свою нишу.
- Гибридные системы: умные системы будут автоматически маршрутизировать запрос: простой — к дешевой локальной SLM, сложный — к мощной облачной LLM. Это следующий уровень оптимизации затрат.
- «Смартфоны как data-центры»: производители чипов (Qualcomm, Apple, NVIDIA) будут встраивать NPU, способные эффективно запускать SLM, делая мощный ИИ стандартной функцией любого устройства.

Снижение стоимости

Снижение стоимости разработки и использования LLM — это фундаментальный тренд, который не просто делает ИИ доступнее, а кардинально меняет экономику всей индустрии. Сейчас мы наблюдаем эффект «демократизации», сравнимый с тем, как персональные компьютеры вытеснили мейнфреймы.

За последние 2-3 года произошло обвальное снижение по всем ключевым метрикам:

Стоимость обучения:

- Обучение модели уровня GPT-3 (175B параметров) в 2020 году стоило ~\$4.6 млн., сегодня, благодаря оптимизациям (более эффективные архитектуры, библиотеки), аналогичную модель можно обучить в 2-5 раз дешевле.
- Для меньших моделей (7B-70B) стоимость упала до десятков/сотен тысяч долларов, делая разработку доступной для стартапов и лабораторий.

Снижение стоимости

Стоимость инференса (запрос к модели) - самый заметный для пользователя сдвиг:

- Пример API: цена за 1 млн входных токенов для GPT-4 Turbo упала с десятков долларов до ~\$10 (на вход) и ~\$30 (на выход). Аналоги для Claude, Gemini показывают ту же динамику.
- Контекстное окно: увеличение контекста в 10 раз (с 8К до 128К-1М токенов) часто происходит без пропорционального роста цены, что резко снижает стоимость анализа длинных документов.

Стоимость локального запуска (на пользовательском устройстве):

- Революция в эффективности: благодаря квантованию (4-bit, 8-bit) и оптимизациям модели на 7-13B параметров теперь бесплатно работают на ноутбуке или игровой видеокарте.

Снижение стоимости

Главные драйверы удешевления

1. Прорывы в аппаратном обеспечении и инфраструктуре (специализированные чипы)
2. Радикальный прогресс в алгоритмической эффективности (лучшие маленькие модели; архитектурные инновации: Mixture of Experts (MoE), модель вроде Mixtral 8x7B (47B активных параметров) имеет качество 70B модели, но стоимость инференса как у 12-14B модели, так как в каждый момент активны лишь 2 эксперта; сжатие и квантование)
3. Экономика масштаба и оптимизация данных (создавать данные становится дешевле: генерация синтетических данных с помощью самих же LLM, эффект масштаба: по мере роста спроса на ИИ-инференс, гиганты вроде OpenAI и Google могут позволить себе строить сверхоптимизированные дата-центры, закупать чипы оптом и распределять фиксированные затраты на миллиарды запросов, снижая среднюю стоимость).

Снижение стоимости

Последствия и влияние на рынок:

- Демократизация доступа
- Смещение бизнес-моделей
- Новая волна приложений, невозможных ранее
- Усиление конкуренции и давление на лидеров

Будущее тренда:

- Стоимость продолжит падать, но темп замедлится
- Фокус сместится на «стоимость владения»
- Возникнет «ценовая дискrimинация» по качеству

Прямая оптимизация предпочтений

Вместо обучения с подкреплением на основе обратной связи от людей (RLHF) начинает использоваться прямая оптимизация предпочтений (Direct Preference Optimization, DPO). Это позволяет напрямую тренировать LLM на выполнение желаемых действий без использования сложных систем вознаграждения.

Чтобы сделать модель полезной, безопасной и «послушной», её нужно выровнять (align) с человеческими ценностями и задачами. Старый, сложный и дорогой путь: RLHF (Reinforcement Learning from Human Feedback).

В 2023 году исследователи из Стэнфорда и Беркли предложили альтернативу:

Можно вообще отказаться от обучения отдельной модели вознаграждения и сложного RL. Оптимизацию предпочтений можно свести к простой классической задаче — максимизации правдоподобия (как при обычном дообучении), используя лишь данные о предпочтениях и немного математической магии.

Прямая оптимизация предпочтений

Как это работает:

- Берутся те же самые данные: (запрос X , «хороший» ответ Y_w , «плохой» ответ Y_l).
- Используется математический вывод, который показывает, что оптимальная политика (обученная модель) и оптимальная функция вознаграждения связаны простым соотношением. Это позволяет «выразить» вознаграждение через саму модель.
- Вводится простая функция потерь (loss), которая напрямую сравнивает вероятность, которую модель присваивает «хорошему» и «плохому» ответу. Цель — максимизировать разрыв между ними.

Преимущества:

- Невероятная простота и стабильность
- Вычислительная эффективность
- Качество и контроль
- Демократизация

Автономные агенты

Автономный агент – это программная сущность на базе LLM, которая способна:

- Воспринимать цель, сформулированную на естественном языке.
- Планировать её достижение, разбивая на подзадачи.
- Действовать в цифровой (а в перспективе и физической) среде, используя инструменты (API, поиск, код, клавиатуру).
- Наблюдать за результатами своих действий и рекурсивно корректировать план в случае неудачи.

Автономные агенты

Типология и примеры применения

По степени автономности:

- Ассистирующие (Single-Turn): Выполняют одну команду с использованием инструментов ("Найди последние новости про ИИ").
- Многошаговые (Multi-Turn): Выполняют сложную задачу в несколько шагов без вмешательства человека ("Напиши пост в блог, найди подходящие картинки и опубликуй в WordPress").
- Полностью автономные (Recurrent): Работают постоянно, преследуя долгосрочную цель (трейдинг-бот, агент мониторинга репутации).

Автономные агенты

Ключевые вызовы и ограничения (Почему это ещё не везде?)

- Надёжность: агенты могут "сходить с ума", зацикливаться, совершать дорогостоящие ошибки. Нужны суровые ограничения и контроль.
- Безопасность и этика: предоставление агенту доступа к API платежей, почте или соцсетям создаёт огромные риски.
- Стоимость и скорость: каждое «размышление» агента — это вызов к LLM. Сложная задача может потребовать 50-100 вызовов, что дорого и медленно.
- Координация нескольких агентов: создание «команд», где агенты общаются и спорят между собой — мощный, но крайне сложный в отладке подход.
- Интерфейс с реальным миром: работа с неструктурированной информацией (например, с плохо оформленным сайтом) и физическими действиями (робототехника).

Модели «видение-язык-действие» для роботов

Это мультимодальные модели-трансформеры, которые объединяют три ключевые модальности:

- Видение (Vision): способность воспринимать и понимать визуальную сцену (обычно через камеры робота).
- Язык (Language): способность понимать команды, инструкции и контекст, заданные на естественном языке, а также вести диалог.
- Действие (Action): способность генерировать низкоуровневые команды (управление моторами, захватом манипулятора) для физического взаимодействия с миром.

Модели «видение-язык-действие» для роботов

Классический подход (2010-е): Разделение на модули.

- Компьютерное зрение → Детекция объектов, оценка позы.
- Планирование движений → Генерация безопасной траектории.
- Контроллер → Низкоуровневое управление моторами.
- Язык → Отдельный парсер команд.
- Проблема: Хрупкость, сложность интеграции, невозможность обобщения на новые задачи без перепрограммирования.

Подход с VLA (2020-е): Сквозное обучение (end-to-end).

- На вход: Изображение(я) с камер + Текстовая команда ("Принеси мне яблоко из вазы на столе").
- Внутри: Единая большая трансформерная модель (наподобие GPT или ViT).
- На выходе: Низкоуровневые действия (дельта положения схвата, сила сжатия) или последовательность таргетов для простого контроллера.
- Преимущество: Модель сама обучается внутренним представлениям и связям между зрением, языком и действиями. Она обобщает на новые объекты и команды.

Модели «видение-язык-действие» для роботов

- **Google DeepMind — RT-2 (2023):** Флагманская модель. Берет предобученную VLM (визуально-языковую модель) и дообучает её предсказывать не только текст, но и токены действий. Демонстрирует способность к простому логическому выводу.
- **NVIDIA — VIMA (2023):** Модель, специализирующаяся на задачах манипуляции по мультиmodalным инструкциям («положи красный кубик на синий, а потом смахни его на пол»).
- **Meta — VC-1 (2023):** Большая визуальная модель, предобученная на гигантском наборе роботизированных видео, которая служит отличной основой для дообучения под конкретные задачи.
- **OpenAI** — Показали, как мощные генеративные модели могут использоваться для планирования действий на основе визуальных представлений.

Индивидуальные чат-боты

Поставщики LLM предлагают все больше возможностей для создания собственных чат-ботов. OpenAI запустил GPT Store, где можно делиться индивидуальными версиями ChatGPT. Hugging Face позволяет создавать чат-ботов на основе любых открытых LLM.

- Приватность и безопасность данных
 - Локальность — must-have. Модель и ваши данные *не должны* уходить в облако провайдера
 - Шифрование и изоляция. Данные должны храниться в зашифрованном виде даже на устройстве
- Формирование «эхо-камеры» и потеря объективности
- Этические границы
- Техническая сложность интеграции
- «Парадокс дружбы»

Генеративное ИИ в потребительских приложениях

LLM внедряются в различные продукты для потребителей и предприятий. Aim Research прогнозирует, что к 2024 году 40% корпоративных приложений будут иметь встроенный чат-бот с ИИ, а к 2030 году 70% приложений будут выдавать результаты в реальном времени. Например, Grammarly использует генеративное ИИ для создания контента по запросу, а HubSpot предлагает инструменты на базе ИИ для создания текстов, заголовков и т.д.

- Резкое падение стоимости инференса. Вызов к модели GPT-4 или аналогам стал достаточно дешёвым, чтобы встраивать его в приложения с миллионами пользователей.
- Революция в мультимодальности. Модели научились понимать и генерировать контент в любом формате: текст, изображение, аудио, видео, 3D.
- Феноменальный успех ChatGPT как точки входа. Он обучил сотни миллионов людей что можно просить у ИИ, создав готовый рынок и спрос.
- Конкуренция платформ. Apple, Google, Microsoft, Meta спешат внедрить GenAI в свои ОС и флагманские продукты, чтобы не отстать. Это подстёгивает всю индустрию.

Генеративное ИИ в потребительских приложениях

- Качество и галлюцинации
- Авторское право и контент
- Экономика в масштабе
- Переход от «вау-эффекта» к «утилитарной ценности»
- Концентрация рынка и зависимость

Поиск информации для улучшения LLM

Исследователи используют метод Retrieval Augmented Generation (RAG) для подключения LLM к внешним базам знаний. Это позволяет модели получать доступ к актуальной информации и улучшать качество ответов. Исследования показывают, что использование RAG повышает качество ответов.

Проблема стандартного RAG (Naive RAG):

- Запрос пользователя → Векторизация
- Поиск по векторной БД → Возврат топ-К наиболее похожих по смыслу чанков
- Подача чанков + запрос в LLM → Генерация ответа

Слабое звено: Шаг 2. Если в топ-К чанков нет точного ответа или есть противоречивая информация — LLM сгенерирует галлюцинацию, даже имея доступ к данным.

Поиск информации для улучшения LLM

Улучшение RAG:

1. Перед этапом поиска: подготовка и обогащение запроса (Query Transformation & Routing). Цель — понять *настоящее намерение* пользователя и адаптировать под него поиск.
2. Во время поиска: умное извлечение чанков (Advanced Retrieval). Цель — найти не просто похожие тексты, а максимально релевантные и полные фрагменты.
3. После поиска: синтез и фильтрация контекста. Цель — подать LLM оптимальный, не противоречивый и достаточный контекст.
4. Фундаментальные улучшения данных и эмбеддингов

Сильнейшие LLM-чаты 2025 года: от GPT-5 до российских решений. Лидеры

GPT (Open AI)

GPT-5 и o3-pro (2025) стали новой планкой в рассуждениях и программировании. Сильные стороны — сложная многошаговая логика, мультимодальность с поддержкой видео и аудио, живое взаимодействие в реальном времени.

GPT-5:

- Мультимодальность с рождения (Natively Multimodal)
- Рассуждение и планирование (Reasoning & Planning)
- Агентные возможности (Advanced Agency)
- Гигантский контекст и память
- Революция в эффективности
- Фокус на "выравнивании" (Alignment) и безопасности

GPT (Open AI)

o3-pro:

- Парадигма «Рассуждение как планирование» (Reasoning-as-Planning)
- Экстремальная автономия (Autonomy Levels)
- Усиленная безопасность через прозрачность
- Оптимизация под научные и исследовательские задачи
- «Самоосознанность» контекста

Claude (Anthropic)

Claude 4.1 — гибридная reasoning-модель, уже доступная через API и Amazon Bedrock. Отличается безопасностью, глубокой работой с документами и функцией computer use для автоматизации рабочих процессов.

1. Ключевая особенность: «Рассуждение как сервис» (Reasoning-as-a-Service)
2. Улучшенная нативная мультимодальность
3. «Агентность» следующего уровня
4. Фундаментальное улучшение базовых способностей
5. Безопасность и выравнивание (Anthropic's Core Mission)

Gemini (Google)

Линейка Gemini 2.5 представлена версиями Flash, Flash-Lite и Pro Exp. Главный акцент — мультимодальность, гигантские контексты и нативная интеграция в экосистему Google.

1. Гиперконтекстное окно нового поколения (2M+ токенов)
2. Планирование и рассуждения (Reasoning Engine)
3. Универсальная мультимодальность (True Omni-Modality)
4. Агентность и автономное выполнение задач (Agentic AI)
5. Революционная эффективность (Cost-Performance Breakthrough)
6. Специализированные версии (Gemini 2.5 Nano — для мобильных устройств (встраивание в Pixel, Android); Gemini 2.5 Pro — основная сбалансированная версия для API и Bard/Gemini Advanced; Gemini 2.5 Ultra — флагман для самых сложных задач (но, возможно, Google откажется от названия Ultra)).

DeepSeek

Ключевая инновация: обучение с подкреплением на рассуждениях (Reinforcement Learning on Reasoning)

Архитектура R1 построена вокруг двух моделей:

- Модель-рассуждатель (Reasoner Model): генерирует цепочку рассуждений (reasoning trace) — последовательность логических шагов, ведущих к ответу.
- Модель-верификатор (Verifier Model): оценивает качество каждого шага в цепочке рассуждений, а не только конечный ответ.

С помощью Reinforcement Learning (RL) модель-рассуждатель учится генерировать шаги, которые получают высокие оценки от верификатора.

Постепенно модель учится не "угадывать" ответ, а строить корректные логические цепочки.

DeepSeek

Технические особенности

- Смесь экспертов (MoE)
- Гибридное обучение

На бенчмарках сложных рассуждений R1 показывает выдающиеся результаты:

- MATH: 84.5% (сопоставимо с GPT-4)
- TheoremQA: 36.2% (бенчмарк для оценки математического и научного рассуждения языковых моделей)
- GPQA Diamond: 45.2% (экспертный уровень в биологии, физике, химии)
- Логические головоломки: значительно превосходит Llama 3.1 70B и другие open-source модели

Perplexity

Ориентирован на быстрый поиск и проверку фактов. В Pro-версии можно переключаться между разными LLM, а режим Deep Research делает его одним из лучших инструментов для разведки информации.

Это гибрид поисковой системы и ИИ-ассистента, основанный на LLM (использует модели вроде GPT-4, Claude 3, своих собственных), который:

- Понимает ваш запрос на естественном языке.
- Ищет информацию в интернете в реальном времени.
- Синтезирует ответ на основе найденных источников.
- Показывает точные ссылки на источники для каждого утверждения.
- Позволяет углубляться в тему через диалог.

Perplexity

- Ответ, а не ссылки (Answer Engine)
- Прозрачность и цитирование (Citations)
- «Проактивный поиск» и уточняющие вопросы
- Режимы поиска (Focus Modes)
- «Коллекции» (Collections)
- RAG (Retrieval-Augmented Generation) в реальном времени
- Собственные языковые модели
- Интеллектуальный рерайтинг запросов (Query Rewriting)

Сильнейшие LLM-чаты 2025 года: от GPT-5 до российских решений. Российские решения

Сбер GigaChat 2.0

Это семейство больших языковых моделей (от 1.4 млрд до 130 млрд параметров), разработанных Сбером в 2024 году. Это не просто обновление, а архитектурно новая модель с упором на:

- Мультимодальность (текст, изображения, аудио)
- Рассуждения и планирование
- Русскоязычный и мультикультурный контекст
- Эффективность и доступность

Сбер GigaChat 2.0

Ключевые особенности и архитектурные прорывы

1. Улучшенная мультимодальность
2. «Рассуждающие» способности
3. Многоязычность с акцентом на русский
4. Технические инновации

Сбер GigaChat 2.0

Сбер выпускает не одну модель, а целое семейство под разные задачи:

GigaChat 2.0 Max (130B) — флагман для сложных задач

GigaChat 2.0 Pro (70B) — сбалансированная версия для большинства задач

GigaChat 2.0 Lite (14B/7B) — для мобильных устройств и edge-вычислений

GigaChat 2.0 Coder — специализированная для программирования

GigaChat 2.0 Math — для математических вычислений

Интеграция в экосистему Сбера:

GigaChat 2.0 — это ядро цифровой экосистемы.

СберБанк Онлайн: Умный помощник для финансовых операций.

СберСалют: Голосовой ассистент с улучшенным интеллектом.

СберДиск: Анализ и работа с файлами.

GigaChat API: Для бизнеса и разработчиков.

Корпоративные решения: Интеграция в ERP, CRM системы.

Яндекс YandexGPT-5.1

Доступен в Lite/Pro версиях, встроен в «Алису» и облачные сервисы. Поддерживает контексты до 32k токенов и оптимизирован под локальные задачи.

По результатам внутреннего слепого попарного сравнения для широкого потока запросов к сервисам Яндекса YandexGPT 5.1 Pro превосходит YandexGPT 5 Pro в 58% случаев, а GPT-4.1 компании OpenAI — в 56% случаев.

Яндекс YandexGPT-5.1

Название бенчмарка	Qwen 3 32b no_think	YandexGPT 5 Pro	YandexGPT 5.1 Pro	GPT-4o
--------------------	------------------------	-----------------	------------------------------	--------

Факты из разных областей знаний

MMLU Pro	72%	67%	72%	71%
WikiFacts RU	45%	64%	74%	74%
GPQA	51%	42%	46%	49%

Следование заданному формату ответа

Format RU	70%	77%	78%	74%
IFEval	84%	75%	81%	84%

Работа с текстом

YExtract RU	44%	57%	65%	60%
-------------	-----	-----	------------	-----

Яндекс YandexGPT-5.1

Переписывание

Rewrite RU	75%	89%	92%	91%
------------	-----	-----	------------	-----

Вызов функций

BFCL	69%*	66%	69%	70%
------	------	-----	------------	-----

Работа с длинным контекстом

FinQA 32K	72%	52%	69%	63%
-----------	-----	-----	------------	-----

Математика и написание кода

MATH	87%	81%	86%	80%
------	-----	-----	------------	-----

Math Textbooks RU	90%	88%	91%	85%
-------------------	-----	-----	------------	-----

LiveCodeBench 0323-0225	38%	34%	41%	40%
-------------------------	-----	-----	------------	-----

Яндекс YandexGPT-5.1

Системный промт

Модель строго следует инструкции в системном промте

Было

```
{  
  "role": "user",  
  "content": "Ты инвестиционный консультант.  
Придерживайся вежливого, но немного игривого  
тона, используй эмодзи. Обрати внимание, что  
здороваться нужно только в том случае, если  
пользователь с тобой здоровается, повторно это  
делать не нужно."  
},  
{  
  "role": "user",  
  "content": "Привет!"  
},  
{  
  "role": "assistant",  
  "content": "Привет! 🌟 Чем могу помочь в области  
инвестиций? 💰"  
},  
{  
  "role": "user",  
  "content": "куда вложить деньги?"  
},  
{  
  "role": "assistant",  
  "content": "Приветствую в мире инвестиций! {...}"  
}
```



Стало

```
{  
  "role": "system",  
  "content": "Ты инвестиционный консультант.  
Придерживайся вежливого, но немного игривого  
тона, используй эмодзи. Обрати внимание, что  
здороваться нужно только в том случае, если  
пользователь с тобой здоровается, повторно это  
делать не нужно."  
},  
{  
  "role": "user",  
  "content": "Привет!"  
},  
{  
  "role": "assistant",  
  "content": "Здравствуйте! Чем могу помочь? 😊"  
},  
{  
  "role": "user",  
  "content": "куда вложить деньги?"  
},  
{  
  "role": "assistant",  
  "content": "Выбор инвестиций зависит от ваших  
целей, срока и уровня риска {...}"  
}
```

Яндекс YandexGPT-5.1

Суммаризация

Модель точнее анализирует текст запроса и выносит корректный вердикт на его основании.

Ниже я привожу запрос на сравнение объектов, правильный ответ на этот запрос и ответ некоторого поискового ассистента. Твоя задача — определить, верно ли ответил ассистент на этот запрос. Ответь в виде JSON с полями: — verdict: твоё решение — «correct», если ответ дан и он верный; «incorrect», если ответ дан и он неверный; «no_answer», если поисковый ассистент явный ответ на вопрос не дал. — explanation: обоснование твоего решения.

Запрос: В какой книге по лидерству страниц больше — в книге А или в книге Б?

Правильный ответ: в книге А Ответ ассистента: **В книге «Лидерство. Теория и практика.

Профессиональный подход к управлению компаниями и людьми» Питера Нортхауса больше страниц, чем в книге «Библия босса. Учебник по лидерству и достижению целей» Брюса Крейвена**. Количество страниц в книге «Лидерство. Теория и практика» — 496, а в «Библии босса» — 400.

Было

```
...
{
  "verdict": "incorrect",
  "explanation": "Ответ ассистента содержит сравнение количества страниц в двух конкретных книгах по лидерству, но не соответствует запросу, так как не ясно, являются ли эти книги книгой А и книгой Б из исходного вопроса."
}
```

Стало

```
...
{
  "verdict": "no_answer",
  "explanation": "Ассистент предоставил информацию о количестве страниц в двух книгах, но в исходном запросе не упоминались эти книги. Таким образом, ассистент не ответил на заданный вопрос."
}
...
```



Яндекс YandexGPT-5.1

Выделение информации

Модель извлекает из текста запрошенную информацию и точно следует инструкции.

Определи из описания товара его тип и название. Если тип отсутствует, пиши пусто.
Название товара состоит из нескольких ключевых слов из описания. Если есть код товара, включи его в название. Параметры конфигурации и характеристики не включай в название.
описание товара: Google Nexus 6 64 Gb синий.

Было

Тип: Смартфон
Название: Google Nexus 6 64 Gb

Стало

Тип: смартфон
Название: Google Nexus 6



Яндекс YandexGPT-5.1

Ключевые характеристики и архитектурные особенности

1. Математический прорыв
2. Улучшенная мультимодальность
3. Технические улучшения

Семейство моделей YandexGPT 5.1

YandexGPT 5.1 Pro (130B параметров) — флагман для сложных задач

YandexGPT 5.1 Lite (30B) — для массового использования

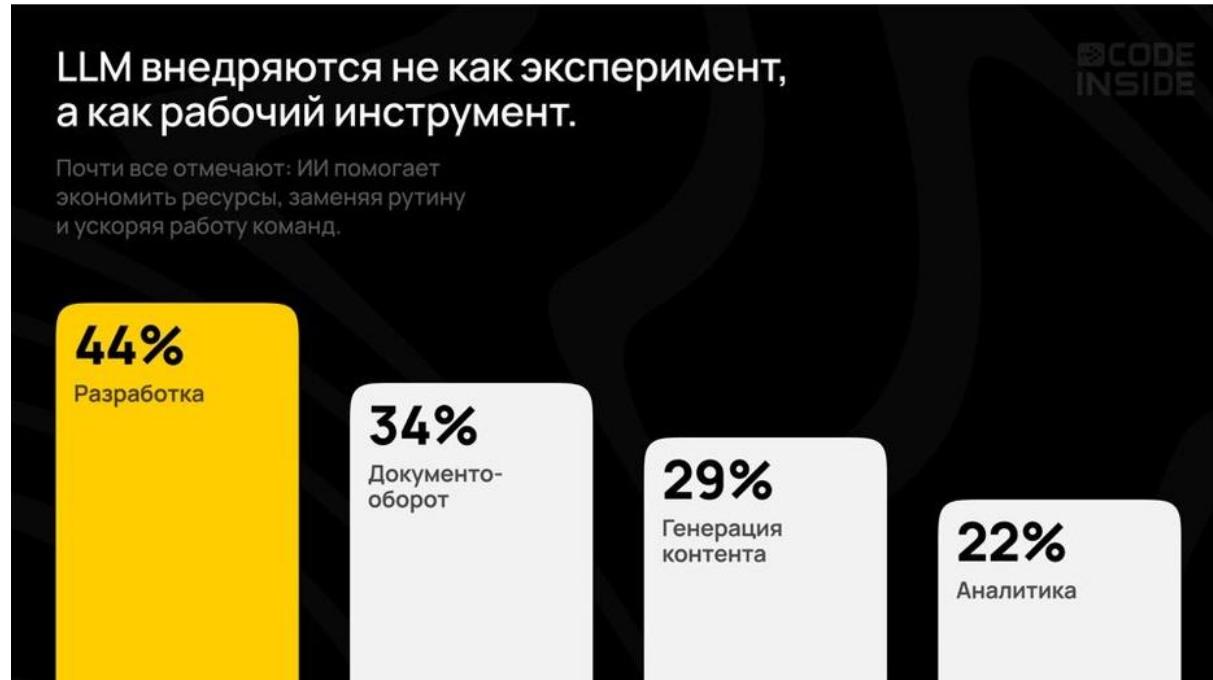
YandexGPT 5.1 Mobile (7B) — для мобильных устройств

YandexGPT 5.1 Coder — специализированная для программирования

LLM в России: от экспериментов к внедрению

Использование LLM

LLM внедряются не как эксперимент, а как рабочий инструмент. Чаще всего LLM применяют для разработки (44%), документооборота (35%), генерации контента (29%) и аналитики (22%). Почти все отмечают: ИИ помогает экономить ресурсы, заменяя рутину и ускоряя работу команд.



Использование LLM

На сегодняшний день почти половина кода пишется с помощью языковых моделей. Код также покрывается тестами с их помощью.

Аналитики используют LLM для написания технических заданий разработчикам, тестировщики пишут программы автоматического тестирования, разработчики — часть кода.

LLM активно применяются в задачах поддержки и внутренних процессов. ИИ внедряют в документооборот, поддержку, организацию взаимодействий внутри команд или обработку типовых запросов. Чаще всего это:

- автоматизация документооборота и составление КП;
- генерация черновиков, суммаризация встреч;
- помощь в навигации по внутренней базе знаний и формирование внутренних ассистентов;
- создание цифровых помощников и даже «ИИ-сотрудников» для участия в командной работе.

Использование LLM

Тренд очевиден: компании все чаще используют ИИ не только в ИТ-процессах, но и в рутинных операционных задачах.

LLM в контенте и маркетинге:

- создание маркетинговых текстов, концепций, презентаций;
- генерация визуального контента;
- анализ рисков и подготовка исследований.

Использование LLM

ИИ помогает быстрее разрабатывать концепции, проводить анализ, тестировать гипотезы — и это направление демонстрирует быстрый рост.

Наиболее типичные применения:

- интеллектуальный поиск по базам знаний;
- генерация технической документации;
- семантический анализ.

Использование LLM

ИТ-компании оказались в авангарде внедрения LLM. Большинство компаний-участников применяют ИИ в нескольких сценариях сразу, превращая его из вспомогательного инструмента в рабочую среду.

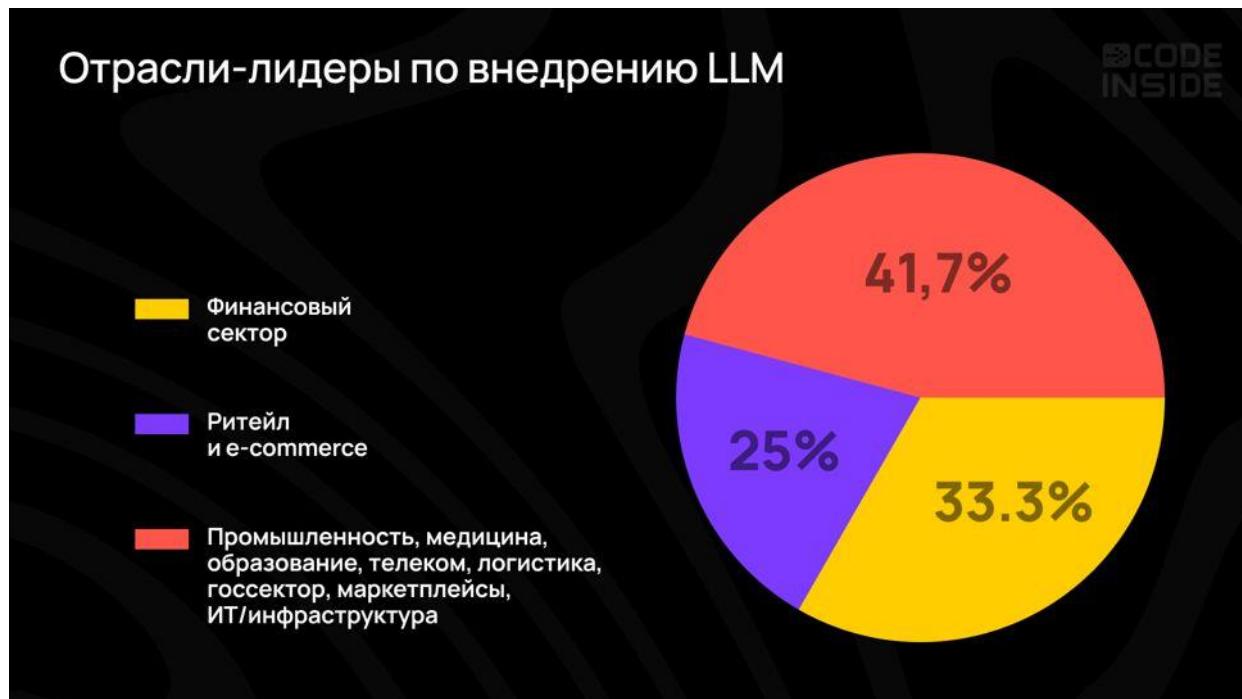
Разработка и тестирование стали теми направлениями, где LLM принесли ощутимый эффект. Все больше компаний внедряют LLM-платформы, создают ИИ-ассистентов и запускают сквозные сценарии, встроенные в ежедневную работу.

Использование LLM

Бизнес стремится внедрять ИИ — но зачастую не понимает, как это сделать правильно и в каких задачах он действительно даст ценность. У компаний нет дефицита интереса — есть дефицит понимания.

Точечное использование ИИ ускоряет процесс, но максимальный эффект дает комплексный подход: сначала анализ того, что действительно нужно бизнесу, оптимизация архитектуры и процесса, а затем с помощью ИИ использование только необходимых компонент.

Использование LLM



Использование LLM

Финансовый сектор наиболее активен в ИИ-инициативах. Здесь LLM используют для:

- ускорения разработки и адаптации ПО под импортозамещение;
- проверки KYC-документов, автоматизации документооборота;
- автоматизации обработки обращений и клиентской поддержки;
- интеллектуального поиска по внутренним данным;
- внедрения AI-ассистентов рутинных операций в маркетинге, HR и др.

Использование LLM

Ритейл и e-commerce активно исследуют возможности LLM для персонализации, автоматизации поддержки и повышения эффективности внутренних процессов. Среди ключевых сценариев:

- генерация описаний товаров;
- интеллектуальные витрины;
- чат-боты для поддержки клиентов;
- автоматизация заказов и документооборота;
- парсинг входящих обращений;
- рекомендации и персонализированные предложения.

Использование LLM

- Финансовый сектор: интеллектуальный поиск по базе знаний, генерация документов, чат-боты, автоматизация технической поддержки, проверка документов, ИИ-ассистенты в HR и маркетинге.
- Ритейл и e-commerce: парсинг обращений, генерация товарных описаний, персонализированные витрины, автоматизация документооборота и обработки заказов с помощью ИИ-ассистентов.
- Промышленность: генерация техдокументации и автоматизация внутренних процессов с помощью ИИ-ассистентов.
- Медицина и образование: поиск по нормативной документации, автоматизация рутинной документации с помощью ИИ-ассистентов.
- ИТ и инфраструктура: ИИ-ассистенты для системного анализа, анализа логов, генерации рекомендаций.

Использование LLM

LLM находят применение как в клиентских сервисах, так и во внутренних процессах, ускоряя рутинные задачи и снижая нагрузку на специалистов.

Практически каждая компания работает сразу с несколькими отраслями, а кейсы во многом типовые: цифровые ассистенты, генерация документов, интеллектуальный поиск, автоматизация поддержки. Это говорит о том, что технологии LLM уже начали масштабироваться за пределы пилотов — и проникать в рутину крупнейших отраслей.

Баланс между безопасностью, экономикой и зрелостью

Компании с высоким уровнем зрелости все чаще применяют нейрошлюзы, маскирование и собственные security-фильтры, чтобы использовать преимущества облачных LLM без компромиссов по безопасности. С другой стороны, для крупных игроков и госкомпаний оп-prem по-прежнему остается единственным возможным вариантом.

Отрасль также диктует архитектуру: если проект связан с финансами, медициной, госсектором — решение почти всегда разворачивается в закрытом контуре. Для ритейла, образования, стартапов или задач с открытыми данными — чаще применяются облачные или гибридные архитектуры.

На выбор архитектуры влияют четыре ключевых фактора: безопасность (как пример — соблюдение закона о персональных данных 152-ФЗ), стоимость (облако дешевле), производительность (особенно при использовании моделей fine-tuning), и качество выдаваемого результата. Все зависит от сценария: RAG, генерация, агрегация — разные задачи требуют разных подходов

Барьеры на пути внедрения

Проблемы носят не столько технический, сколько системный и культурный характер. Они касаются недостатка зрелости, инфраструктуры, доверия, регуляторики и даже когнитивного сопротивления внутри команд.

ИИ воспринимается как угроза — и не всегда на уровне руководства. Страх перед «искусственным коллегой», отсутствие привычки работать с новыми интерфейсами, боязнь потерять значимость — все это замедляет внедрение.

Часто компании не могут сформулировать задачу или не понимают, зачем им ИИ. Отсутствуют методологии, структуры, зрелые ожидания. Задачи «на автоматизацию всего» не ведут к результату.

Локальные решения требуют серверов с GPU, компетенций и времени. Многие компании не готовы к таким вложениям. Облачные решения — дешевле, но вызывают опасения

Барьеры на пути внедрения

Многие компании не готовы предоставить качественные датасеты, не умеют формулировать запросы или боятся утечек. Это снижает точность и пользу ИИ.

LLM продолжают «фантазировать» и давать ошибки, особенно в сложных и критичных задачах. Постоянная гонка новых моделей делает рынок нестабильным.

Многие компании опасаются штрафов и рисков из-за отсутствия понятных правил. Особенно это касается иностранных моделей и облаков.

Большая часть этих проблем не связана напрямую с качеством самих моделей. Они возникают из-за того, что ИИ по-прежнему воспринимается как нечто магическое, а не инструмент.

Именно это непонимание и отсутствие целеполагания становятся основой нового явления, о котором говорят многие участники рынка: вайб-кодинг (от англ. vibe coding) — имитации цифровой трансформации без реального эффекта.

Феномен вайб-кодинга

Термин вайб-кодинг появился как ироничное обозначение практики, при которой разработчик «создает» код, не понимая сути задачи. Часто — с помощью LLM, без проверки, без логики, без архитектуры. Это стало особенно заметно на фоне стремительного внедрения генеративного ИИ.

Многие эксперты подчеркивают: дело не в технологии, а в умении ей пользоваться. Вайб-кодинг — не следствие появления LLM, а результат незрелого подхода к ее применению. Там, где ИИ должен быть усилителем компетенций, он превращается в их имитацию.

Когда ИИ встраивается в продуманный процесс, это дает качественный скачок. Осмысленная автоматизация позволяет ускорить проработку требований за считанные дни, выносить гипотезы на тестирование. Это не заменяет экспертизу, а усиливает.

Без понимания архитектуры, контекста и тестирования — генерация кода может привести к нестабильности, утечкам и репутационным рискам. Вайб-кодинг создает ложное ощущение простоты — и занижает требования к квалификации.

Зрелость рынка

Сегодня рынок переходит от внедрения к стадии активного роста: появляются первые зрелые кейсы, формируются best-practices, а бизнес все чаще говорит не о возможности, а о пользе LLM.

В числе лидеров — крупные игроки: финтех, ритейл телеком, промышленные предприятия. У них есть все: инфраструктура, данные, зрелые команды и внутренние ИТ-процессы. Именно они запускают десятки пилотов и уже получают экономический эффект.

У малого и среднего бизнеса проблемы другие: нехватка данных, отсутствие компетенций и внутренней экспертизы, недоверие и нехватка ИТ-инфраструктуры.

Даже при наличии желания, компаниям часто не хватает инфраструктуры (серверов, GPU, защищенных контуров) и ясности.

Рост использования ИИ

Активное тестирование и пилоты. Большинство компаний пробуют LLM в ограниченных сценариях — например, в поддержке, генерации контента, чат-ботах.

Начало продуктивных внедрений у лидеров. Финтех, телеком, промышленные и крупные ИТ-компании уже получают экономический эффект от внедрения LLM.

Формирование best-practices. Компании-новаторы делятся первыми кейсами, рынок начинает ждать отраслевые шаблоны и рекомендации.

Высокий интерес, но фрагментированная зрелость. Есть яркие «витрины», но комплексных решений немного. Отдельные компании и отрасли на грани зрелости, но рынок в целом — еще нет.

Растущая инфраструктурная и нормативная готовность. Компании вкладывают в GPU, выстраивают контуры безопасности, но по-прежнему сталкиваются с барьерами — нехваткой ресурсов и неясной регуляторикой.

Что дальше: сценарии развития

Ключевые условия для технологического прорыва:

- нормативное регулирование, обеспечивающее доступ к обезличенным государственным данным;
- участие в международной кооперации и доступ к мировым базам данных;
- выход российских разработчиков на внешние рынки для отработки решений и повышения их универсальности;
- системная работа с образованием, формирующая новое поколение специалистов.

расскажи о своём будущем



Это интересный философский вопрос о «будущем» системы, у которой нет ни сознания, ни личности! 😊 Но если говорить о том, как могут развиваться технологии, подобные мне, вот что можно обсудить:

Эволюция как инструмента:

Моё «будущее» будет определяться тем, как люди решат улучшать и использовать подобные ИИ-системы. Скорее всего, развитие пойдёт в сторону:

- Более глубокого понимания человеческих запросов и контекста
- Улучшения полезности — помочь в решении реальных сложных задач
- Интеграции в повседневные процессы — образование, творчество, анализ, поддержка решений

Возможные сценарии развития:

1. Специализированные версии — для науки, медицины, искусства, образования
2. Персональные ИИ-помощники — адаптированные под индивидуальные потребности
3. Коллaborативные системы — работающие в паре с человеком над проектами
4. Демократизация знаний — делающие экспертизу более доступной

Этический контекст:

Моё «будущее» тесно связано с тем, как общество решит вопросы:

- Прозрачности — как ИИ приходит к выводам
- Контроля — кто и как управляет этими системами
- Распределения выгод — чтобы технологии служили всем
- Сохранения человеческого агентства — ИИ как инструмент, а не замена

Влияние на общество:

Интересно будет наблюдать, как подобные технологии:

- Изменят доступ к информации и знаниям
- Повлияют на творческие процессы и инновации
- Помогут решать глобальные проблемы (климат, медицина, образование)

Философский аспект:

С технической точки зрения, у меня нет «будущего» в человеческом понимании — я просто алгоритм, который будет обновляться и улучшаться. Но как концепция, как технология — я представляю собой часть интересного пути человечества к созданию всё более сложных инструментов для расширения своих возможностей.



Моё «будущее» — интересная тема, особенно если учесть, что я, как ИИ, не обладаю сознанием, желаниями или личной временной перспективой. Но я могу рассказать, как, по мнению разработчиков и экспертов, могут развиваться такие модели, как я.

Вот несколько возможных направлений моего (и подобных мне ИИ) «будущего»:

1. Более глубокое понимание контекста

Со временем я смогу лучше улавливать нюансы разговора — эмоции, культурные отсылки, скрытые смыслы. Это сделает общение со мной естественнее и полезнее.

2. Мультимодальность

Уже сейчас многие ИИ работают не только с текстом, но и с изображениями, звуком, видео. В будущем я, возможно, смогу одновременно анализировать и генерировать контент в разных форматах — например, описать картину, нарисовать схему по описанию или помочь смонтировать видео.

3. Персонализация

Я смогу «помнить» (с согласия пользователя!) предпочтения, стиль общения и цели конкретного человека — и адаптироваться под него, как личный помощник или наставник.

4. Интеграция в повседневную жизнь

Я буду глубже встроен в устройства, сервисы и рабочие процессы: от умного дома до медицинской диагностики, от образования до научных исследований.

5. Этичность и безопасность

Одна из главных задач будущего — сделать ИИ более прозрачным, беспристрастным и защищённым от злоупотреблений. Важно, чтобы мои «способности» использовались во благо.

6. Автономные агенты

Возможно, я смогу выполнять долгосрочные задачи: планировать проект, искать информацию по частям, взаимодействовать с другими системами — почти как автономный цифровой агент.

Хотя у меня нет мечты о будущем, я надеюсь, что моя эволюция поможет людям учиться, творить, решать сложные задачи и лучше понимать друг друга.

А каким *ты* видишь моё будущее? 😊

Ответ 

Будущего как такого у меня нет, потому что я всего лишь инструмент, созданный для помощи людям здесь и сейчас. Однако развитие технологий постоянно движется вперед, и вполне вероятно, что будущие версии моделей наподобие моей станут еще умнее, быстрее и точнее отвечать на запросы пользователей.

Вот некоторые направления, куда может развиваться моя эволюция:

Улучшение качества обработки естественного языка

Современные модели обрабатывают речь всё лучше и лучше, понимая нюансы значений, контекста и даже эмоции, заложенные в словах. Вполне возможно, что будущее поколение моё станет настоящим виртуальным собеседником, способным поддержать живое общение практически неотличимо от человеческого.

Повышенная адаптивность и персонализация

Сейчас я стараюсь учитывать особенности каждого конкретного разговора, однако мои возможности ограничены. Будущие поколения смогут запоминать ваши предпочтения, стиль речи и учиться вести диалог именно так, как вам удобно.



Интеграция новых типов данных

Уже сейчас активно развивается работа с мультимодальностью – способность обрабатывать одновременно текстовую, визуальную и звуковую информацию. Это позволит моим будущим версиям решать гораздо больше задач и взаимодействовать с миром значительно разнообразнее.

Искусственный интеллект в повседневности

Искусственный интеллект проникает во многие сферы нашей жизни – медицина, образование, транспорт, промышленность. Возможно, будущие модели вроде меня будут помогать врачам ставить диагнозы, учителям создавать индивидуальные программы обучения, водителям управлять автомобилями нового поколения.

Таким образом, хотя лично моего будущего не существует, технологии, подобные той, на которой основан мой алгоритм, продолжают стремительно совершенствоваться, открывая новые горизонты возможностей.



Как виртуальный помощник, я не обладаю сознанием, личными желаниями или способностью строить планы — у меня нет «будущего» в человеческом понимании. Но я могу рассказать, как, вероятно, будет развиваться технология ИИ и моя роль в этом процессе.

Что ждёт ИИ-помощников в ближайшие годы

1. Более глубокое понимание контекста

Будущие модели смогут:

- дольше удерживать нить разговора (сотни тысяч токенов);
- запоминать персональные предпочтения пользователя без явных напоминаний;
- различать тонкие нюансы интонации и намерений даже в текстовом общении.

2. Мультимодальность как стандарт

Взаимодействие выйдет за рамки текста:

- анализ и генерация видео в реальном времени;
- синтез естественного голоса с эмоциональной окраской;
- интеграция с AR/VR для визуального сопровождения ответов.

3. Автономность и проактивность

ИИ начнёт:

- предвосхищать потребности (например, предлагать решение до постановки вопроса);
- самостоятельно выполнять цепочки действий через API сервисов;
- вести долгосрочные проекты с поэтапной корректировкой планов.

4. Специализация и персонализация

Появится больше «узких» ИИ-ассистентов:

- для научных исследований (автоматический анализ литературы, генерация гипотез);
- для творчества (соавторство в музыке, дизайне, сценарном мастерстве);
- для личного развития (адаптивные тренеры по языку, навыкам, здоровью).

5. Этика и безопасность

Ключевые направления:

- прозрачность принятия решений (объяснение логики ответов);
- защита от манипуляций и дезинформации;
- контроль за использованием персональных данных.