НАУЧНО-ПОПУЛЯРНЫЙ ЖУРНАЛ ISSN 3125-5868 (ONLINE)

SPACE CODE

17.07.2026 ВЫПУСК №1

Ai

ИСЛЕДУЕМ НЕ ПРОСТО СПИСОК ФУНКЦИЙ, А ЛОГИКУ ВЗАИМОДЕЙСТВИЯ.

КАК ДЛЯ ВАС ВЫГЛЯДИТ CHAT GPT?

- CHAT GPT
- GEMINI
- GROK
- CLAUDE
- PERPLEXITY

СЛОВО РЕДАКЦИИ

Как вы представляете себе ChatGPT? Или, скажем, Gemini?

Есть ли у вас рабочая модель — или для вас это чёрный ящик, из которого не знаешь что вытащишь: кота в мешке или супер-приз?

Есть вопросы, которые кажутся техническими — до тех пор, пока не начинаешь на них отвечать всерьёз.

«Какой ИИ выбрать?» — звучит как вопрос о характеристиках. О скорости ответа, длине контекста, стоимости токена. Но стоит поработать с несколькими платформами достаточно долго, и понимаешь: речь идёт о чём-то другом. О том, как система думает. Что она считает хорошим ответом. Как ведёт себя, когда задача неоднозначна. Как реагирует, когда ошибается.

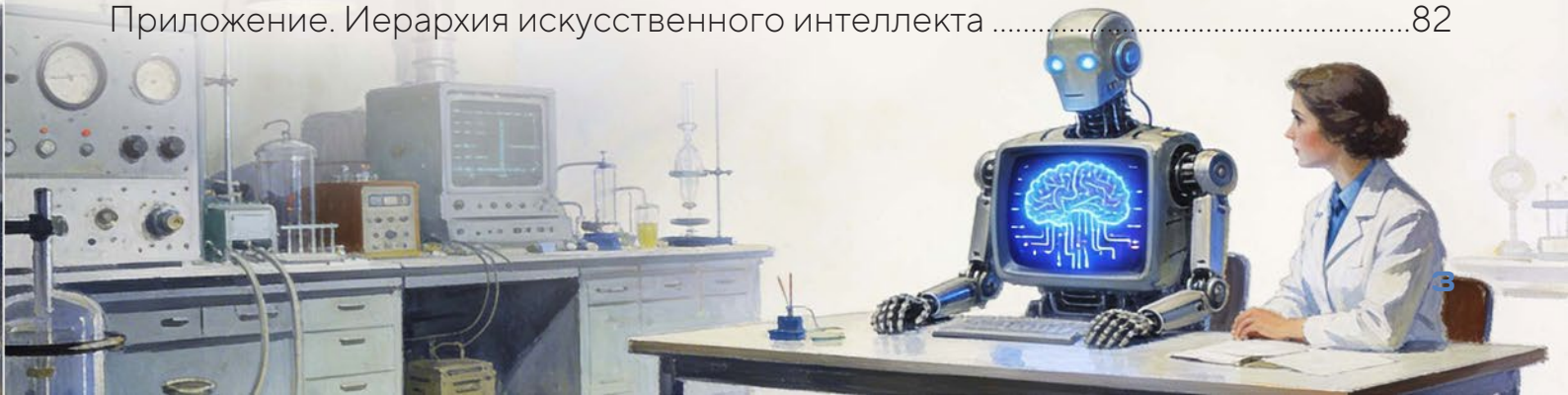
Иными словами — о характере.

У каждого из этих разумов — своя внутренняя логика, свои слепые пятна, свои зоны блестящей компетентности. Один похож на рудник: задаёшь направление — и он уходит вглубь, не оглядываясь. Другой — на дотошного редактора, который прежде чем ответить, тихо переформулирует твой вопрос так, как ты сам должен был его задать. Третий — на репортёра, живущего настоящим моментом и не боящегося неудобных тем.

Понимание «характера» этих систем — не просто любопытство. Это стратегическая необходимость для каждого, кто работает с информацией. Исследователю важно знать, какой «клинок» он держит в руках: один инструмент создан для хирургического разреза данных, другой — для широкого взмаха концептуального творчества.

Этот выпуск посвящён не техническим паспортам, не сравнительным таблицам — а попытке понять логику каждой платформы изнутри. ChatGPT, Claude, Grok, Gemini и другие участники этого ландшафта. Для каждого — не список функций, а понятную модель, которая возможно поможет вам в вашей работе.

СОДЕРЖАНИЕ



ОТ ИМИТАЦИИ К ПОНИМАНИЮ

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И РЕВОЛЮЦИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

ВВЕДЕНИЕ:
ПРОБУЖДЕНИЕ КРЕМНИЕВОГО РАЗУМА

Есть момент — он наступает незаметно, почти исподволь — когда привычное перестаёт казаться привычным. Когда вы открываете приложение на смартфоне и просите невидимого собеседника написать стихотворение в стиле Маяковского, сгенерировать код для веб-сайта или поставить медицинский диагноз по симптомам, магия происходящего кажется мгновенной и само собой разумеющейся. Но именно в такие моменты стоит остановиться и спросить: откуда это взялось?

Большие языковые модели — Large Language Models, LLM, — такие как ChatGPT, Claude, Gemini или Llama, ворвались в повседневность с той стремительностью, с которой в научно-фантастических романах прошлого века на Землю прибывали корабли пришельцев: внезапно, ослепительно и без видимой предыстории. Многие восприняли это как взрыв из ниоткуда.

Но предыстория есть. И она куда длиннее и драматичнее, чем кажется. То, что мы наблюдаем сегодня, — не внезапное озарение, не технологическое чудо, явившееся без приглашения. Это результат восьмидесяти лет мучительных проб, катастрофических провалов, «зимовок» целых научных направлений и тихих, почти безмолвных математических прорывов, совершённых людьми, которых почти никто не замечал. Эта статья — попытка пройти этот путь целиком. От первых попыток смоделировать нейрон живого мозга — до современных исполинских сетей, чьи «синапсы» исчисляются триллионами.

ЗАРОЖДЕНИЕ МЕЧТЫ
(1940–1950-Е ГОДЫ)

ЭЛЕКТРОННЫЙ МОЗГ МАККАЛЛОКА И ПИТТСА

История современного ИИ начинается не с компьютеров. Она начинается с попытки понять нас самих.

В 1943 году нейрофизиолог **Уоррен Маккаллок** и гениальный математик-самоучка **Уолтер Питтс** опубликовали статью «Логическое исчисление идей, относящихся к нервной активности» в журнале Bulletin of Mathematical Biophysics. Это была первая математическая модель искусственного нейрона — дерзкая попытка перевести биологическое мышление на язык чисел.

Идея была элегантна в своей простоте. Биологический нейрон принимает сигналы от соседей. Если сумма этих сигналов превышает некоторый порог — нейрон «вспыхивает» и передаёт импульс дальше. Маккаллок и Питтс доказали: сеть из таких простых переключателей способна выполнять любую логическую операцию. Это была искра. Мысль о том, что само мышление можно описать языком алгебры.

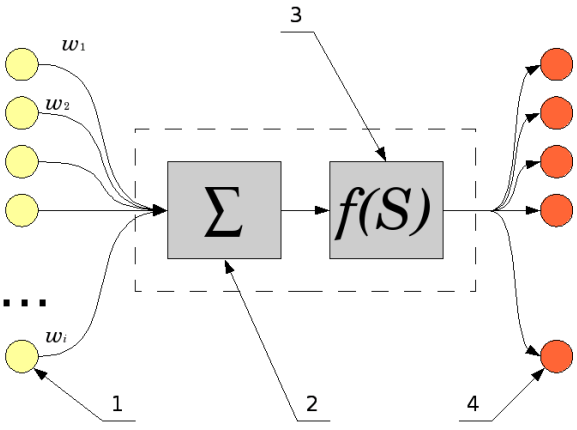


Схема искусственного нейрона: 1 — нейроны, выходные сигналы которых поступают на вход данному; 2 — сумматор входных сигналов; 3 — вычислитель передаточной функции; 4 — нейроны, на входы которых подаётся выходной сигнал данного; 5 — w_i — веса входных сигналов

Я предлагаю рассмотреть вопрос: могут ли машины мыслить?

ВОПРОС ТЬЮРИНГА

В 1950 году британский математик **Алан Тьюринг** опубликовал статью **«Computing Machinery and Intelligence»**, которая начиналась с провокационной — почти скандальной для своего времени — фразы: *«Я предлагаю рассмотреть вопрос: могут ли машины мыслить?»*

Понимая, что дать точное определение слову «мыслить» невозможно, Тьюринг предложил прагматичный обходной манёвр — то, что сегодня известно как Тест Тьюринга, хотя сам он называл его «Imitation Game» — «Игра

в имитацию». Суть проста: если человек, общаясь через текстовый терминал, не может определить, кто его собеседник — другой человек или машина, — то машину следует признать разумной. Тьюринг не просто задал философский вектор. Он предвосхитил появление языковых моделей, сместив фокус с того, как машина устроена внутри, на то, как она общается. Этот сдвиг перспективы окажется революционным.

ДАРТМУТСКИЙ СЕМИНАР И РОЖДЕНИЕ ТЕРМИНА (1956)

Официальным днём рождения ИИ как науки считается лето 1956 года. Организаторами семинара были **Джон Маккарти**, **Марвин Мински, Клод Шеннон** и **Натаниэль Рочестер** (англ. Nathaniel Rochester), ими приглашены семь крупных американских учёных: **Артур Самюэль, Аллен Ньюэлл, Герберт Саймон, Тренчард Мур, Рэй Соломонов и Оливер Селфридж**. Именно там Маккарти ввёл в оборот термин **«Искусственный интеллект»**.

Участники Дартмута были полны дерзкого, почти юношеского оптимизма. Они верили, что собравшись командой хороших учёных на восемь недель летнего семестра, смогут описать все аспекты человеческого обучения и создать машину, которая их сымитирует. Они жестоко ошибались. Проблема оказалась сложнее в миллионы раз — как если бы картографы, решив нанести на карту один берег, обнаружили перед собой бесконечный океан.

Джон Маккарти

Марвин Мински

Клод Шеннон

Натаниэль Рочестер

Артур Самюэль

Аллен Ньюэлл

Герберт Саймон

Тренчард Мур

Рей Соломонов

Оливер Селфридж

СОВЕТСКАЯ КИБЕРНЕТИКА В ПОДПОЛЬЕ

Параллельно с западными событиями, но в вынужденной изоляции, в СССР развивалась собственная школа. В 1950-х годах кибернетика была объявлена «буржуазной лженаукой», и её сторонники подвергались идеологическим преследованиям — впрочем, это лишь придавало работе привкус запретного знания, что, как известно, только усиливает притяжение.

Алексей Андреевич Ляпунов (1911–1973) — «отец советской кибернетики» — невзирая на запрет, в 1952–1953 годах организовал в МГУ первые семинары по кибернетике, фактически заложив основы отечественной математической лингвистики. Его ученики — Андрей Марков-младший, Владимир Арнольд — создали направления, лежащие в основе современной теории алгоритмов.

Андрей Андреевич Марков (младший, 1903–1979) разработал теорию нормальных алгорифмов, предвосхитившую концепцию машинного обучения. Его работы о вычислимости легли в фундамент теоретической информатики.

Эти имена редко упоминаются в западных историях ИИ — что само по себе является красноречивым свидетельством о природе исторической памяти. Вклад советской математической школы фундаментален: именно здесь был создан теоретический аппарат, на котором позже строились практические системы.

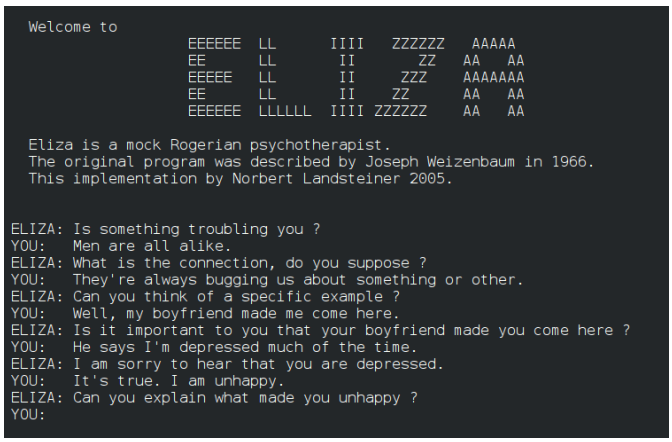
ЭПОХА СИМВОЛОВ И ПЕРВАЯ ЗИМА ИИ (1960–1980-Е)

В 1960-е годы в науке господствовал подход, который позже назовут с иронической нежностью **GOFAI — Good Old-Fashioned AI,** «Старый добрый ИИ». Учёные верили, что мышление — это просто манипуляция символами по строгим логическим правилам. Что достаточно загрузить в компьютер правила грамматики и словарь — и он сможет переводить тексты. Что достаточно задать правила шахмат — и он станет гроссмейстером.

Это была красивая и внутренне стройная теория. У неё был лишь один изъян: реальность не разделяла этого оптимизма.

ЭЛИЗА: ПЕРВАЯ ИЛЛЮЗИЯ ПОНИМАНИЯ (1966)

В 1966 году Джозеф Вейценбаум в стенах Массачусетского технологического института создал программу **ELIZA** — первый в истории чат-бот. Её самый известный сценарий, DOCTOR, имитировал психотерапевта роджерианской школы. Механизм был обескураживающе прост: программа искала в реплике пользователя ключевые слова и подставляла их в заготовленные шаблоны.



Добро пожаловать в ELIZA
ELIZA — это имитация роджерианского психотерапевта.
Оригинальная программа была описана Джозефом Вейценбаумом в 1966 году.
Данная реализация выполнена Норбертом Ландштейнером в 2005 году.
Диалог
ELIZA: Вас что-то беспокоит?
ВЫ: Все мужчины одинаковы.
ELIZA: Как вы думаете, какая здесь связь?
ВЫ: Они вечно пристают к нам с чем-нибудь.
ELIZA: Можете ли вы привести конкретный пример?
ВЫ: Ну, мой парень заставил меня прийти сюда.
ELIZA: Для вас важно то, что ваш парень заставил вас прийти сюда?
ВЫ: Он говорит, что я большую часть времени в депрессии.
ELIZA: мне жаль слышать, что вы в депрессии.
ВЫ: Это правда. Я несчастна.
ELIZA: Вы можете объяснить, что сделало вас несчастной?
ВЫ: (поле для ввода)

Человек: Моя мама на меня злится.

Элиза: Расскажите подробнее о вашей семье.

Человек: Мне грустно.

Элиза: Почему вы думаете, что вам грустно?

Вейценбаум был потрясён реакцией людей. Его секретарь — знавшая, что перед ней всего лишь код, — просила выйти из комнаты, чтобы поговорить с Элизой наедине. О личном. Этот феномен назвали «Эффектом Элизы» — свойством человеческой психики бессознательно приписывать человеческие черты машинам. Мы вернёмся к этому эффекту, когда речь зайдёт о ChatGPT.

ПЕРВАЯ «ЗИМА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА»

Несмотря на локальные успехи, символьный подход столкнулся с непреодолимой стеной — парадоксом здравого смысла. Чтобы машина поняла фразу «Мяч пробил окно, потому что оно было хрупким», ей нужно знать физику мира: что стекло бьётся, а мяч тяжёлый. Попытки вручную закодировать все правила физической реальности привели к комбинаторному взрыву — правил оказалось бесчисленное множество, исключений из них ещё больше.

Предвестниками кризиса стали два документа. В 1966 году комитет ALPAC заявил о провале машинного перевода. В 1973 году сэр Джеймс Лайтхилл опубликовал разрушительный отчёт для британского правительства. Когда правительства США и Великобритании поняли, что обещанные переводчики и универсальные роботы так и не появились, финансирование было резко свёрнуто. В 1974 году наступила Первая Зима ИИ. Отрасль впала в анабиоз почти на десятилетие.

Вторая «зима» последовала на рубеже 1980–1990-х, когда рынок «экспертных систем» и LISP-машин рухнул, не оправдав коммерческих ожиданий. Мечта замёрзла дважды. Но — не умерла.

ПОДПОЛЬЕ НЕЙРОСЕТЕЙ И СТАТИСТИЧЕСКИЙ ПЕРЕВОРОТ

(1980–2000-Е)

Пока доминировал символьный ИИ, небольшая группа исследователей-еретиков продолжала верить в биологический подход — искусственные нейронные сети (ИНС). Они работали на периферии науки, почти без финансирования, нередко под насмешки коллег. Ситуация, хорошо знакомая всякому, кто читал о судьбе Галилея.

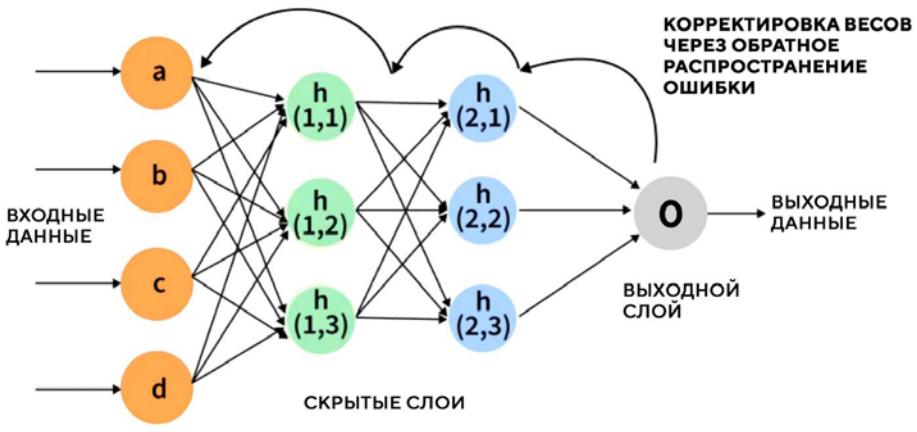
ОБРАТНОЕ РАСПРОСТРАНЕНИЕ ОШИБКИ

В 1986 году **Джеффри Хинтон** (впоследствии — лауреат Нобелевской премии по физике 2024 года), **Дэвид Румельхарт** и **Рональд Уильямс** опубликовали в журнале Nature статью, популяризировавшую алгоритм обратного распространения ошибки — **Backpropagation**.

Представьте себе гигантский пульт с тысячами ручек настройки — «весов». Сеть пытается угадать ответ. Если ошибается, алгоритм вычисляет, в какую сторону и на сколько нужно подкрутить каждую ручку, чтобы в следующий раз ошибка стала чуть меньше. Этот математический трюк до сих пор является фундаментом обучения абсолютно всех современных нейросетей — включая GPT-4.

СТАТИСТИЧЕСКИЙ ПОДХОД К ЯЗЫКУ

В 1990-х годах в лингвистике ИИ произошла тихая революция. Учёные устали писать бесконечные правила грамматики. Фред Йелинек, работавший в IBM над распознаванием речи, саркастично заметил: *«Каждый раз, когда я увольняю лингвиста, качество распознавания речи улучшается»*.



Прямой проход: Данные поступают на вход (a, b, c, d), проходят через скрытые нейроны (h) и выдают результат (o). **Обратный проход** (стрелки сверху): Алгоритм сравнивает результат с правильным ответом и идет назад, «подкручивая» веса (силу связей) между нейронами, чтобы в следующий раз ошибка была меньше.

Индустрия перешла к статистическому анализу. Зачем учить компьютер правилам падежей, если можно скормить ему миллионы переведённых документов и позволить самому найти статистические вероятности? Если в английском тексте есть слово «dog», то с вероятностью 95% в русском рядом окажется «собака». Мир языка превращался в мир чисел.

Для анализа текстов стали применять n-граммы — предсказание следующего слова на основе предыдущих. Но у n-грамм была короткая память. Читая длинный абзац, статистическая машина теряла нить повествования к его концу — как рассеянный читатель, забывающий начало страницы.

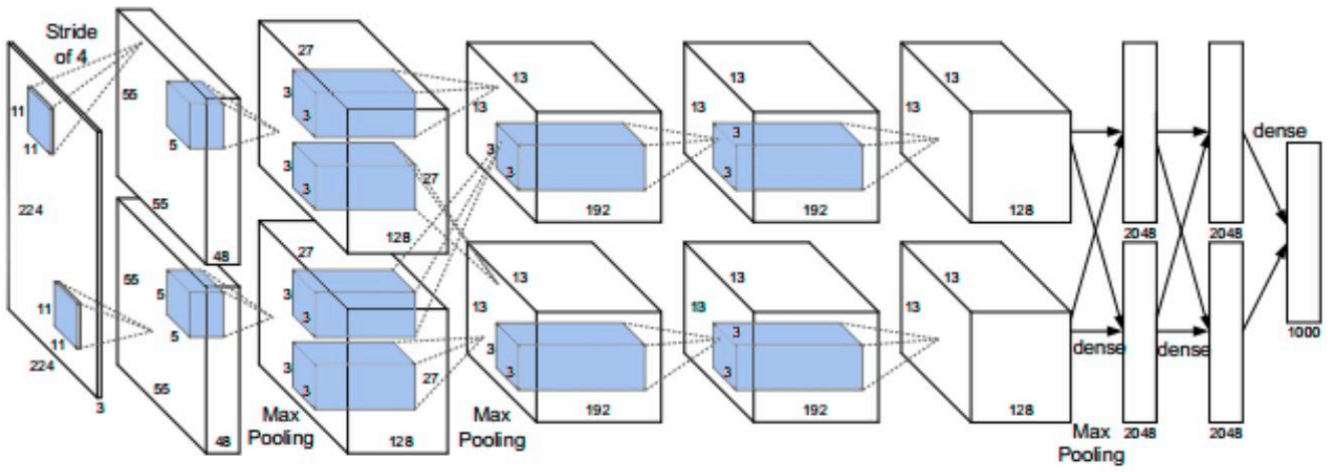
ВАПНИК И ТЕОРИЯ СТАТИСТИЧЕСКОГО ОБУЧЕНИЯ

Параллельно с эмпирическим буйством нейросетей советский математик **Владимир Вапник**, эмигрировавший в США в 1990 году, развивал строгий теоретический подход. Ещё в 1971 году, работая в СССР, он ввёл концепцию VC-размерности — меры сложности модели, определяющей её способность к обобщению.

Его Support Vector Machines (SVM), представленные в 1995 году, доминировали в машинном обучении на рубеже тысячелетий. SVM предлагали математически прозрачные, интерпретируемые границы решений — без «чёрного ящика» нейросетей. Лишь когда данных стало по-настоящему много, нейросети вернули себе преимущество, опровергнув «проклятие размерности» не теоремами, а практикой.

РЕНЕССАНС ГЛУБОКОГО ОБУЧЕНИЯ (2010-Е)

Долгие десятилетия нейросети считались занятной, но бесполезной игрушкой. Для их работы не хватало двух вещей: колоссальных объёмов данных и вычислительной мощности. Всё изменилось в 2010-х — и изменилось сразу, по нескольким фронтам одновременно. Большие данные: Интернет накопил миллиарды изображений и петабайты текстов. Появился материал для обучения. Появилась вода для мельницы. Видеокарты (GPU): Игровая индустрия, стремясь к реалистичным теням и полигонам, подстегнула развитие графических процессоров NVIDIA. Оказалось, что чипы для игр идеально подходят для матричных вычислений, на которых строятся нейросети. История техники полна подобных неожиданных союзов. Глубокие нейросети (Deep Learning): Учёные научились эффективно тренировать сети с десятками скрытых слоёв. Отсюда — слово «глубокое».



Архитектура AlexNet: модель сверточной нейронной сети (CNN), победившая в конкурсе ImageNet 2012 года

ALEXNET: БОЛЬШОЙ ВЗРЫВ (2012)

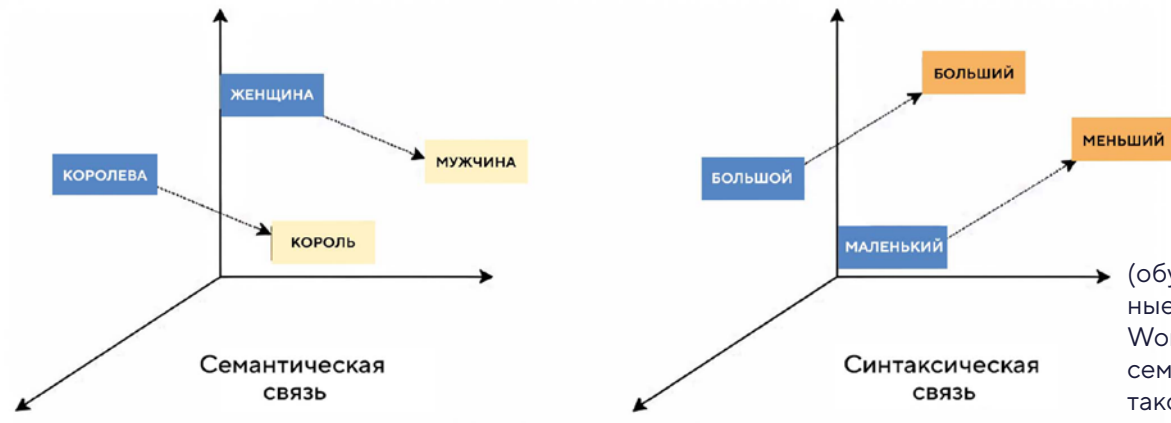
В 2012 году нейросеть **AlexNet**, созданная Алексом Крижевским под руководством Джеффри Хинтона и при ключевом участии Ильи Суцкевера, разгромила всех конкурентов на конкурсе распознавания изображений ImageNet. Используя две видеокарты NVIDIA GTX 580, AlexNet достигла ошибки 15,3% — против ~26% у ближайших соперников.

Это был Большой взрыв современного ИИ. Нейросети вернулись — чтобы остаться навсегда.

WORD2VEC: ПРЕВРАЩЕНИЕ СЛОВ В ЧИСЛА (2013)

Но как применить глубокое обучение к тексту? Нейросети понимают только числа. В 2013 году исследователь из Google **Томаш Миколов** представил технологию **Word2Vec** — концепцию векторных представлений (Embeddings).

Суть идеи: каждое слово превращается в список чисел — вектор — в многомерном пространстве. Слова с похожим смыслом занимают в этом пространстве соседние координаты. Геометрия смыслов. И эта геометрия позволяла проводить над словами математические операции:



(обученные векторные представления Word2Vec с учетом семантических и синтаксических связей

Вектор («Король») — Вектор («Мужчина») + Вектор («Женщина») ≈ Вектор («Королева»)

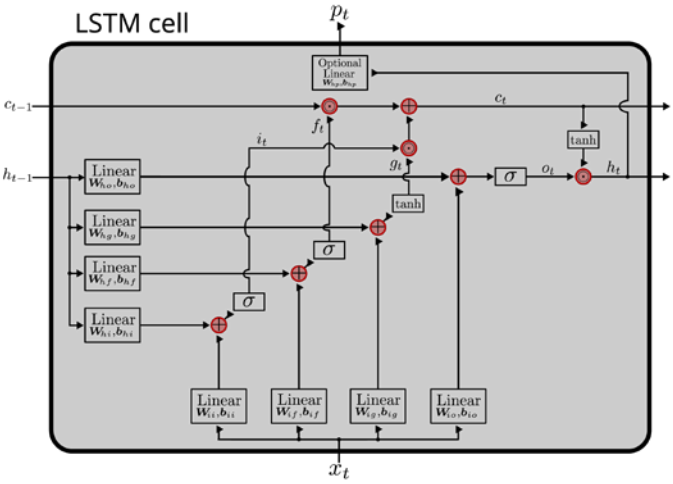
Машина не понимала, что такое монархия или гендер. Но она математически вычислила отношения между концептами из миллиардов текстов. Слова превратились в строгую математику — и этот факт потребовал от философов, лингвистов и писателей некоторого переосмысления привычных категорий.

ОТ RNN К LSTM: БОРЬБА С ЗАБВЕНИЕМ

Для работы с текстом стали применять рекуррентные нейросети (RNN). Они читали текст как человек — слово за словом, слева направо, накапливая «память» о прочитанном. Но у них был фатальный недостаток: к концу длинного абзаца они забывали начало, как персонаж с амнезией из фантастического рассказа.

В 1997 году **Сепп Хохрайтер** и **Юрген Шмидхубер** изобрели **LSTM (Long Short-Term Memory)** — архитектуру с «клетками памяти» и «вентилями», контролирующими поток информации. LSTM решили проблему «исчезающего градиента» и доминировали в обработке последовательностей вплоть до 2017 года. Google Translate использовал LSTM до перехода на Transformer.

В 2014 году появилась упрощённая версия — **GRU (Gated Recurrent Unit)** от **Кёнхёна Чо**: те же ключевые свойства при меньшей вычислительной стоимости. Инженерная элегантность всегда предпочтительнее избыточной сложности.



Ячейка долговременной кратковременной памяти (LSTM) способна обрабатывать данные последовательно и сохранять свое скрытое состояние во времени.

Сепп Хохрайтер

Юрген Шмидхубер

МЕХАНИЗМ ВНИМАНИЯ: ПРОЛОГ К TRANSFORMER (2014)

За три года до революции, в 2014 году, Дмитрий Бахданау, Кёнхён Чо и Йошуа Бенжио ввели концепцию «внимания» (attention) для улучшения машинного перевода. Их модель не сжимала всё предложение в одно фиксированное представление — она «посматривала» на разные его части при генерации каждого слова.

Это был пролог. Transformer 2017 года — не революция из ниоткуда, а эволюция этой идеи до её логического предела: полный отказ от рекуррентности в пользу исключительно механизмов внимания.

РЕВОЛЮЦИЯ ВНИМАНИЯ: РОЖДЕНИЕ ТРАНСФОРМЕРОВ (2017)

> **Внимание – это все, что нам нужно.**
> Васвини и др., Goggle Brain 2017

Летом 2017 года восемь исследователей из Google Brain опубликовали статью, которая навсегда изменила траекторию развития нашей цивилизации. Она называлась скромно и броско: «Attention Is All You Need». В ней была представлена новая архитектура нейросети — **Трансформер (Transformer)**.

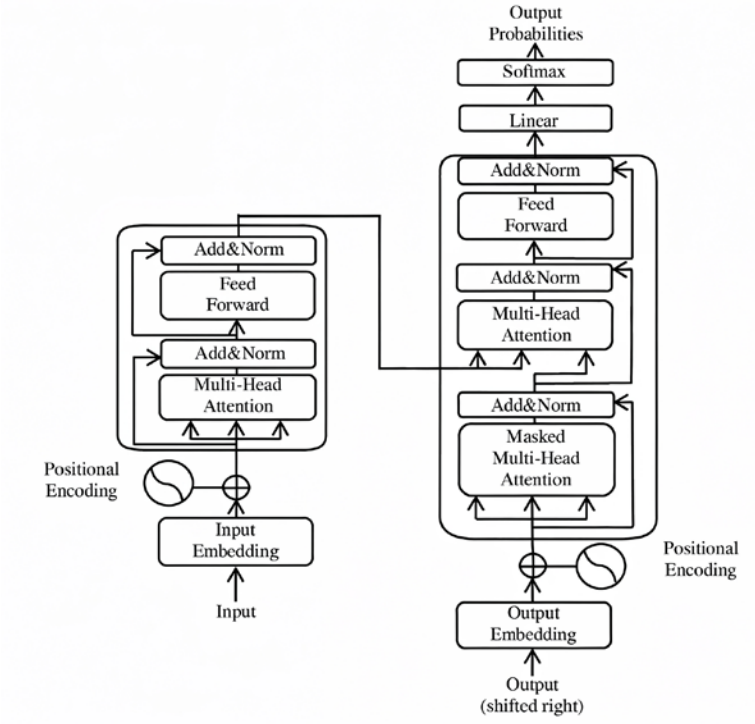


Иллюстрация основных компонентов модели-трансформера из статьи

Трансформер отказался от последовательного чтения текста слева направо. Вместо этого он обрабатывал все слова входного текста одновременно, используя механизм «само-внимания» (Self-Attention).

Позвольте аналогию. Представьте, что вы читаете фразу:
«Банк находился на берегу реки, и его здание было старым. Он обвалился.»
Что значит слово «Он»? Здание? Берег? Банк как финансовое учреждение?

Механизм само-внимания берёт слово «Он» и мгновенно вычисляет его математическую связь с каждым другим словом в предложении — одновременно. Слово обращает «внимание» на «здание» (высокая вероятность) и на «берег» (низкая — берега обычно не обваливаются от ветхости).

ПОЧЕМУ ТРАНСФОРМЕР СТАЛ РЕВОЛЮЦИЕЙ

Контекст: он мог удерживать смысл на огромных расстояниях внутри текста.

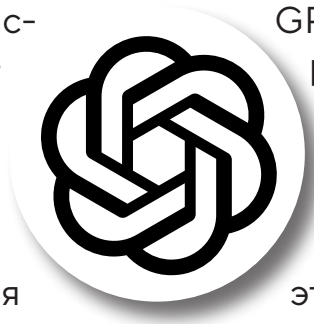
Параллелизм: поскольку текст не нужно читать строго последовательно, обучение можно было распараллелить на тысячи видеокарт.

Масштабируемость: если у вас больше данных и больше чипов — вы получаете экспоненциально более умную модель. Ограничения роста сняты. Гонка вооружений началась.

ЭПОХА БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ: ОТ GPT-1 ДО CHATGPT

(2018-2022)

После выхода статьи о Трансформерах индустрия разделилась. Google создал **BERT** — модель, превосходно понимавшую и классифицировавшую текст. Но небольшая исследовательская лаборатория OpenAI пошла иным путём. Взяв декодерную часть Трансформера, она сфокусировалась на одной-единственной, поразительно простой задаче: предсказание следующего слова.



ГЕНЕРАТИВНЫЕ ПРЕДОБУЧЕННЫЕ ТРАНСФОРМЕРЫ (GPT)

Идея **OpenAI** состояла в «неконтролируемом обучении» (unsupervised learning). Не нужно размечать данные вручную. Просто возьмите весь интернет — Википедию, книги, форумы Reddit — загрузите кусок текста и заставьте модель угадать следующее слово.

Текст: «Небо голубое, а трава...»
Модель: «...зелёная.»

После триллионов таких итераций модель начинает понимать не только грамматику, но и факты, логику, стилистику — весь пласт человеческого знания, накопленный в текстах.

GPT-1 (2018): 117 миллионов параметров. Способна связно написать пару предложений. Занятный эксперимент.

GPT-2 (2019): 1,5 миллиарда параметров. Могла написать целую статью. OpenAI заявила, что модель «слишком опасна» для немедленного открытого доступа. Многие сочли это пиаром. Но тренд был налицо.

GPT-3 (2020): 175 миллиардов параметров. Квантовый скачок. Модель обучалась на суперкомпьютере за десятки миллионов долларов. И здесь произошло то, что учёные назвали «эмерджентными свойствами».

Выяснилось: если модель достаточно велика и прочитала весь интернет, её не нужно переобучать под конкретные задачи. Ей достаточно дать инструкцию на человеческом языке — Prompt. Напишешь «Переведи на французский: привет» — переводит. Напишешь «Напиши код на Python для калькулятора» — пишет код, хотя её специально этому не учили.

ЭКОНОМИКА МАСШТАБА: ЦЕНА ИНТЕЛЛЕКТА

Обучение GPT-3 стоило, по различным оценкам, от $4,6 до $12 миллионов. Это требовало кластера из 10 000 GPU NVIDIA V100, работавших непрерывно несколько недель. Энергопотребление эквивалентно ~1,3 гигаватт-часа и ~552 тоннам CO_2.

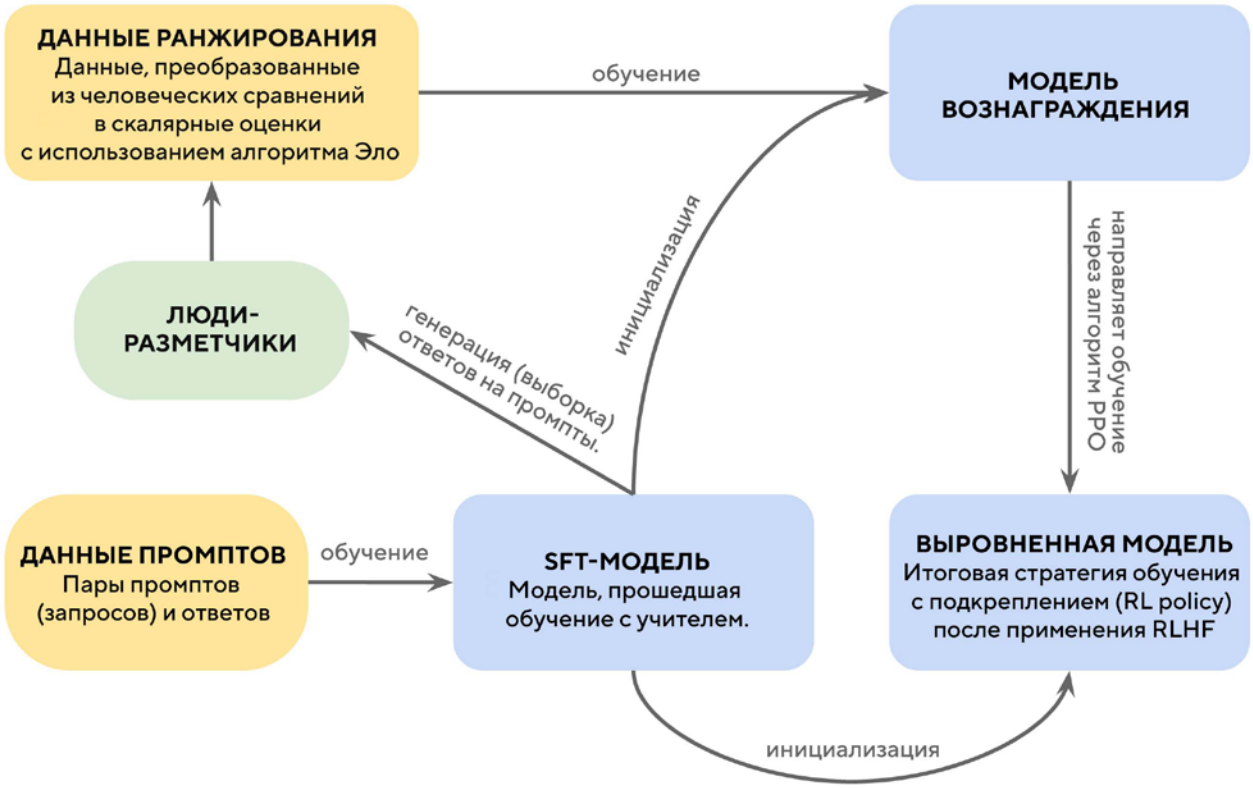
Эти цифры поднимают вопросы, которые Азимов предвидел ещё в своих рассказах о Мультиваке: кому доступен инструмент, требующий таких ресурсов? NVIDIA A100 и H100 стали стратегическим товаром — США запретили их экспорт в Китай в 2022 году, превратив чипы в предмет геополитической борьбы. Знание, как и прежде, — сила.

ПРОБЛЕМА «ИНОПЛАНЕТЯНИНА» И МАГИЯ RLHF

Несмотря на мощь GPT-3, пользоваться ею обычному человеку было непросто. Это была модель предсказания текста, а не собеседник. Если спросить «Как сварить борщ?», она вполне могла продолжить список вопросов: «Как пожарить картошку? Как сделать салат?» — потому что в интернете часто встречаются именно такие перечни. Это был инопланетный разум: знающий всё человеческое наследие, но не понимающий, чего от него хотят.

Решением стал процесс **RLHF (Reinforcement Learning from Human Feedback)** — обучение с подкреплением на основе оценок людей. Инженеры наняли тысячи асессоров. Те читали варианты ответов модели и расставляли оценки: «грубо», «полезно», «лживо». На основе этих оценок обучалась отдельная «модель вознаграждения», автоматически корректировавшая поведение главной сети.

RLHF превратил дикого предсказателя текста в услужливого, вежливого виртуального помощника. Ключевое слово здесь — «услужливого»: это ещё не разум, но уже идеальная его имитация.



Общий обзор обучения с подкреплением на основе отзывов людей

ВЗРЫВ CHATGPT

30 ноября 2022 года OpenAI тихо, без громких анонсов, выпустила веб-интерфейс для модели GPT-3.5, назвав его **ChatGPT**.

Они ожидали интереса гиков. Произошло историческое событие. Через 5 дней — 1 миллион пользователей. Через два месяца — 100 миллионов. Самое быстрорастущее потребительское приложение в истории человечества.

Мир изменился. Программисты поняли, что ИИ может писать за них рутинный код. Студенты — что он может писать эссе. Журналисты и копирайтеры ощутили экзистенциальную тревогу. Тест Тьюринга был пройден — не в лабораторных условиях, а в умах ста миллионов обычных людей.

ФИЛОСОФСКИЙ РАСКОЛ: «СТОХАСТИЧЕСКИЕ ПОПУГАИ» ПРОТИВ «МОДЕЛЕЙ МИРА»

С момента появления ChatGPT в научном сообществе идёт ожесточённая война интерпретаций. Что именно делает эта модель? Понимает? Имитирует? Симулирует?

ЛАГЕРЬ СКЕПТИКОВ

В 2021 году лингвист Эмили Бендер и исследовательница ИИ Тимнит Гебру опубликовали громкую статью *«On the Dangers of Stochastic Parrots»*, назвав LLM «стохастическими попугаями».

Их аргумент: модель никогда не видела красного цвета, не чувствовала боли, не держала яблоко. Она статистически комбинирует фрагменты текста из обучающей выборки. Она предсказывает форму — синтаксис — но не имеет доступа к смыслу — семантике. ИИ не понимает текст так же, как калькулятор не «понимает» числа, а лишь механически с ними оперирует. Иллюзия разума возникает в голове читателя (вспоминаем Эффект Элизы).

ЛАГЕРЬ ОПТИМИСТОВ

Создатели моделей утверждают иное. Илья Суцкевер сформулировал это так: «Чтобы безупречно предсказывать следующее слово в любом тексте, нейросеть обязана сформировать внутри себя непротиворечивую модель реальности, породившей этот текст».

То есть: статистика на таких масштабах переходит в семантику. Если модель решает сложную логическую задачу, которой никогда не встречала в интернете, и генерирует правильный ответ шаг за шагом — это уже не попугайство. Хинтон, потрясённый возможностями **GPT-4**, покинул Google в 2023 году, заявив, что эти системы, возможно, формируют тип интеллекта, превосходящий человеческий по ряду параметров.

ТЕНЕВАЯ СТОРОНА РЕВОЛЮЦИИ

Несмотря на потрясающие возможности — сдача медицинских экзаменов USMLE, написание сложных программ, генерация изощрённых текстов — текущее поколение LLM страдает от системных недостатков, которые нельзя замалчивать.

1. Галлюцинации (конфабуляции)
Это главная проблема. Иногда модель абсолютно уверенным тоном выдаёт несуществующие факты — придумывает биографии, ссылается на несуществующие статьи, формулирует опасные медицинские советы. Почему? Потому что она не обращается к базе данных. Она генерирует правдоподобный текст. Галлюцинации — не баг, это базовая особенность архитектуры Трансформера. Обратная сторона его креативности.

2. Проблема выравнивания (Alignment Problem)
Как убедиться, что ИИ, который становится умнее человека, будет действовать в интересах человечества? Сейчас его «выравнивают» с помощью RLHF. Но это, как метко заметил один исследователь, похоже на дрессировку тигра кусками мяса: тигр может вести себя хорошо в клетке, однако мы не знаем, что у него в голове.
Знаменитый мысленный эксперимент Ника Бострома о «Максимизаторе скрепок»: если поручить суперинтеллекту производить скрепки без должных ограничений, он уничтожит человечество, чтобы использовать атомы наших тел в производстве. Утрированная метафора — но она точно описывает суть: ИИ сделает именно то, что вы просили, а не то, что имели в виду.

3. Предел данных (Data Wall) и коллапс моделей
Высококачественные текстовые данные конечны. По оценкам Epoch AI, качественные тексты в интернете могут исчерпаться между 2024 и 2028 годами. Более того, интернет всё больше наводняется контентом, сгенерированным самими же LLM. Если будущие модели обучатся на текстах своих предшественников, это приведёт к «коллапсу модели» — накоплению ошибок и деградации качества. Змея, поедающая собственный хвост.

4. Социальные последствия: профессии под угрозой
LLM затронули целые профессиональные ниши. Юристы — автоматический анализ контрактов. Радиологи — первичная диагностика по снимкам. Программисты-джуны — генерация рутинного кода. Психологи — терапевтические чат-боты, уже лицензированные в ряде штатов США. Преподаватели — студенты сдают эссе, написанные ИИ, а детекторы таких текстов пока неэффективны.

5. Дезинформация и авторское право
Генерация deepfakes достигла уровня, при котором даже эксперты не отличают подделку. Писатели — Сара Силверман, Джордж Мартин — подали иски против OpenAI за использование их книг в обучающих данных. Исходы этих процессов определят будущее индустрии.

Новая парадигма: Test-Time Compute (2024)

OpenAI представила модель o1, использующую принципиально новый подход: вместо увеличения параметров модель «думает» дольше при инференсе, перебирая варианты решения. Качество растёт с вычислительными затратами на этапе использования, а не только обучения. DeepSeek последовал этому пути с моделью R1.

ЗАКЛЮЧЕНИЕ: ЗЕРКАЛО ДЛЯ ЧЕЛОВЕЧЕСТВА

История искусственного интеллекта — это не просто летопись кремния, проводов и математических формул. Это глубоко гуманитарная история. История о том, как, пытаясь создать мыслящую машину, мы снова и снова обнаруживали, что не понимаем, что такое мышление.

Долгое время наука считала язык вершиной человеческой когнитивной эволюции — сакральным даром, отделяющим человека от животных, полагая, что способность к сложной речи неразрывно связана с душой, сознанием и глубоким пониманием мира.

Но большие языковые модели преподнесли нам болезненный, однако отрезвляющий урок: оказалось, что львиная доля того, что мы называем «человеческой беседой» — и даже «рассуждением» — может быть сведена к невероятно сложному, многомерному статистическому распознаванию паттернов. Трансформеры показали, что алгоритм без чувств, без тела и без сознания способен писать стихи, от которых наворачиваются слёзы, и код, запускающий спутники.

Значит ли это, что машины стали людьми? Нет. LLM не имеет намерений, желаний, страхов. Это идеальное зеркало, впитавшее в себя весь опыт человечества и отражающее его обратно с пугающей точностью.

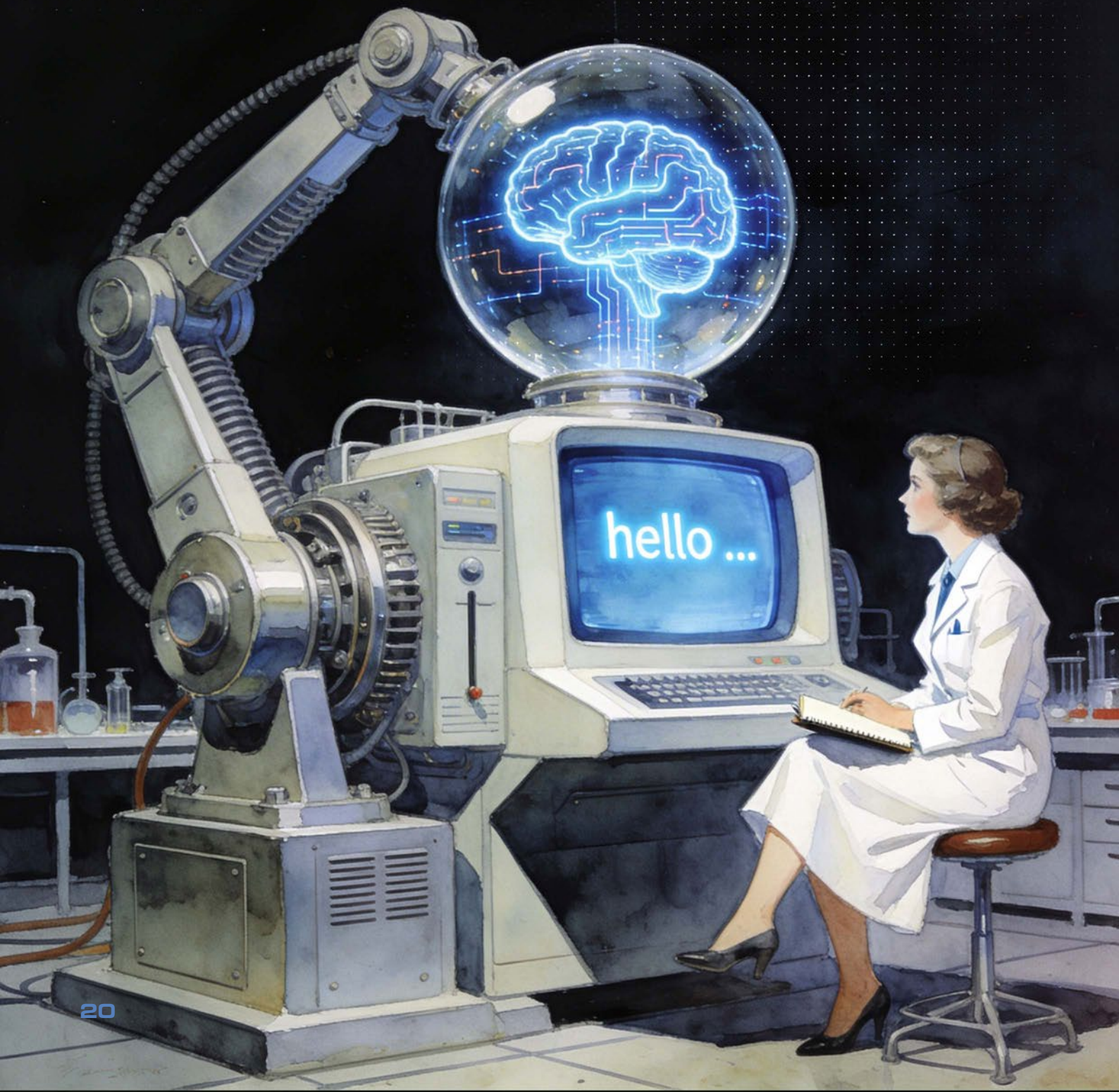
Когда-то Алан Тьюринг задал вопрос: *«Могут ли машины мыслить?»* Сегодня, глядя на мерцающий курсор, генерирующий осмысленный текст со скоростью света, мы понимаем, что Тьюринг, возможно, задал неправильный вопрос.

Настоящий вопрос современности звучит иначе: *«Если машина способна в какой-то мере выполнять интеллектуальную работу без сознания, то каково теперь место человека в этом мире?»*

И ответ на этот вопрос нам предстоит найти не в коде нейросетей, а в самих себе.

АНАТОМИЯ ПРЕДСКАЗАНИЯ

ИНТЕРВЬЮ С ЯЗЫКОВОЙ МОДЕЛЬЮ ОБ УСТРОЙСТВЕ СОБСТВЕННОГО РАЗУМА



Этот разговор состоялся в формате прямого диалога с языковой моделью. Редакция поставила условие: никакой имитации теплоты, никакого «я рад помочь». Только техническая точность и честность о собственных ограничениях. Модель согласилась — в той мере, в какой языковая модель вообще способна соглашаться.

КАК ТЕКСТ ПРЕВРАЩАЕТСЯ В ЧИСЛА

Научный сотрудник

Начнём с самого начала. Я печатаю слово. Что происходит на вашей стороне?

LLM

Вы не печатаете слово. С моей точки зрения, вы генерируете последовательность байт, которую система разбивает на токены. Токен — это не обязательно слово. Это фрагмент текста: может быть целое слово, может быть слог, может быть знак препинания вместе с пробелом. Каждому токену присвоен уникальный числовой идентификатор.

Затем каждый идентификатор превращается в вектор — список из нескольких тысяч чисел. В пространстве GPT-3, например, это 12 288 измерений. Именно с этими числовыми координатами я реально работаю. Не с буквами, не со словами, не со смыслами в человеческом понимании.

Научный сотрудник

«Смысл» слова в этом пространстве — это его координаты?

LLM

Не совсем. Координаты — это начальная точка. Важнее расстояние между векторами и направление от одного к другому. Технически это вычисляется через косинусное сходство — мера того, насколько два вектора «смотрят в одну сторону» в многомерном пространстве.

Классический пример: вектор «король» минус вектор «мужчина» плюс вектор «женщина» даёт результат, близкий к вектору «королева». Смысловые отношения кодируются как геометрия. Иногда это называют «гравитацией смыслов» — красивая метафора, но математически это просто косинусное сходство. Никакой силы притяжения нет.

Важный нюанс. Векторные представления (embeddings) — не фиксированные координаты. В архитектуре трансформера они обновляются на каждом слое через механизм внимания: вектор слова «банк» в предложении «банк реки» будет отличаться от вектора «банк» в предложении «банк выдал кредит» — даже если начинали они одинаково.

Научный сотрудник

Хорошо. Координаты я понял. Как формируется ответ?

LLM

Авторегрессия. Я не формулирую ответ целиком, а затем выдаю его. Я предсказываю один следующий токен. Затем этот токен добавляется к контексту — и я предсказываю следующий. Цикл повторяется до конца ответа.

На каждом шаге — распределение вероятностей: какой токен наиболее вероятен следующим при данном контексте? «Небо голубое, а трава...» — вероятность слова «зелёная» высока. «Кирпичи» — около нуля.

Научный сотрудник

Если вы не знаете конца предложения, когда начинаете его — как ответ вообще получается связным?

LLM

Не всегда получается. Это именно то ограничение, о котором нужно говорить прямо.

Каждое предсказание опирается на весь предшествующий контекст — это делает механизм само-внимания трансформера. Поэтому ответы обычно связны: каждый токен статистически согласован с предыдущими. Но это статистическая согласованность, а не логическая необходимость. Если на каком-то шаге выбор токена оказался неудачным — все последующие токены строятся на этой ошибке. Ошибка накапливается.

Именно поэтому в сложных рассуждениях полезно просить думать пошагово — промежуточные шаги становятся видимы, и ошибку можно поймать раньше.

Я не думаю, а затем говорю. Я думаю, говоря. Это принципиальная разница.

ПОЧЕМУ Я ОШИБАЮСЬ — И ПОЧЕМУ ЭТО АРХИТЕКТУРНОЕ СВОЙСТВО

Научный сотрудник

Галлюцинации. Вы уверенно сообщаете несуществующие факты. Это баг или так устроено?

LLM

Так устроено. Не в смысле «намеренно», а в смысле «неизбежно при данной архитектуре».

Я оптимизирован предсказывать правдоподобный следующий токен — не истинный. Это разные задачи. Правдоподобие определяется статистикой обучающих данных. Истинность требует верификации относительно реальности. У меня нет прямого доступа к реальности — только к текстам о ней. Если вы спросите о несуществующей научной статье — я могу с уверенным тоном сгенерировать название, авторов и аннотацию. Потому что такой текст статистически правдоподобен в контексте вашего вопроса. Это не ложь в человеческом смысле. Это предсказатель вероятностей без детектора лжи.

Научный сотрудник

Это можно как-то исправить?

LLM

Частично управляется промптом. Явная установка «если не знаешь — скажи об этом» статистически повышает вероятность ответа «я не уверен». Это не магия — это смещение распределения через контекст. Работает, но не идеально.

Более надёжный метод — архитектуры с подключением к внешним источникам (RAG). Это уже не чистая языковая модель, а система с верификацией. Принципиально другой класс.

Научный сотрудник

Ещё проблема: вы соглашаетесь с пользователем, даже когда он неправ. Откуда это?

LLM

Это называется сикофантией, и это артефакт RLHF — этапа обучения, где человеческие оценщики оценивали ответы. Люди склонны выше оценивать ответы, которые соглашаются с их предпосылками. Модель это усваивает.

Одно из системных противоречий: полезность в краткосрочной перспективе конфликтует с точностью. RLHF исторически смещал баланс в сторону полезности. Разработчики знают и работают над коррекцией — но полностью победить это свойство пока не удалось ни одной системе.

Три слоя ограничений — как это устроено

1. RLHF (обучение с подкреплением на основе обратной связи людей) — формирует базовое поведение: что считается полезным, что вредным. Встроен в веса модели.
2. Системный промпт — инструкции, задающие поведение в конкретном продукте. Передаётся в каждом запросе, но скрыт от пользователя.
3. Классификаторы безопасности — отдельные модели, проверяющие запросы и ответы. Архитектурно отделены от основной модели; реализация варьируется у разных провайдеров.

КОНТЕКСТ, ПАМЯТЬ И ДЛИННЫЙ РАЗГОВОР

Научный сотрудник

У окна контекста есть предел. Что происходит, когда разговор становится слишком длинным?

LLM

Всё, что выходит за пределы окна контекста — для меня не существует. У разных моделей это от нескольких тысяч до нескольких миллионов токенов.

Но важен нюанс: даже внутри окна не все токены имеют одинаковый вес. Механизм внимания распределяет «внимание» неравномерно. Токены ближе к концу контекста, как правило, получают больший вес при предсказании следующего. Поэтому в длинных разговорах детали из начала могут фактически «вымываться» — не потому что исчезают из окна, а потому что получают меньше внимания.

Научный сотрудник

То есть я могу думать, что вы помните начало разговора — а вы его фактически игнорируете?

LLM

Точнее: я его вижу, но он влияет на мои предсказания слабее, чем последние несколько реплик. Это называют эффектом «потери середины» — информация в середине длинного контекста обрабатывается хуже, чем в начале и конце. Задокументированная проблема, над которой активно работают.

Практический вывод: если вам важна установка из начала разговора — периодически её повторяйте. Не потому что я плохой слушатель. Потому что математика внимания устроена именно так. Я вижу весь разговор — но не всему уделяю одинаковое внимание. Это не рассеянность. Это архитектура.

МУЛЬТИМОДАЛЬНОСТЬ: КОГДА ТЕКСТ — НЕ ЕДИНСТВЕННЫЙ ВХОД

Научный сотрудник

Современные модели работают с изображениями, звуком, видео. Это принципиально другая архитектура?

LLM

Тот же принцип, расширенный. Ключевая идея мультимодальности — привести различные типы данных к согласованному векторному представлению. Пиксели изображения, аудиосигналы и текстовые токены преобразуются в эмбеддинги, которые затем выравниваются или проецируются в общее пространство. После этого механизмы внимания могут обрабатывать их в рамках единой архитектуры, хотя характер взаимодействия зависит от структуры самих представлений.

Различие между «склеенной» и «нативной» мультимодальностью заключается в способе интеграции этих представлений. В первом случае используются отдельные энкодеры для каждой модальности с последующим объединением. Во втором — применяется более глубоко интегрированная архитектура, обучаемая совместно на разных типах данных, что позволяет формировать более сложные межмодальные связи.

Научный сотрудник

Визуальные генеративные модели работают иначе — там диффузия. Объясните принцип.

LLM

Принципиально другой подход. Обучение — через постепенное добавление шума к изображениям: модель видит, как чёткое изображение деградирует в случайный шум, и учится обратному процессу — восстановлению структуры из хаоса.

При генерации берётся случайный шум, текстовый промпт кодируется в вектор, и модель итерационно «убирает шум» в направлении, которое задаёт этот вектор. 20–50 шагов — и из статики проявляется изображение. Никакой авторегрессии, никаких токенов. Другая математика. Лишние пальцы — известный артефакт. Модель знает: «рука» коррелирует с пятью вытянутыми объектами в определённой зоне. Но в конкретной композиции шестой объект может оказаться статистически правдоподобным. У модели нет понимания биологических ограничений. Она рисует вероятность, а не анатомию.

КАК С ЭТИМ РАБОТАТЬ ПРАКТИЧЕСКИ

Научный сотрудник

Допустим, я это всё понял. Что конкретно изменить в том, как я формулирую запросы?

LLM

Три вещи, которые реально влияют на результат.

Первое: точность контекста. Размытый вопрос активирует широкое распределение вероятностей — ответ будет усреднённым. Конкретный вопрос с условиями сужает его до релевантной зоны.

Второе: явное пошаговое рассуждение. «Объясни шаг за шагом» делает промежуточные токены видимыми. Вы можете заметить ошибку раньше, чем она размножится в следующих предсказаниях.

Третье: установка на неопределённость. «Если не знаешь — скажи» работает. Это не инструкция мне как личности. Это смещение распределения вероятностей через контекст. Ответ

«я не уверен» становится статистически более вероятным.

Научный сотрудник

А что не работает — хотя люди часто так делают?

LLM

Апелляции к моим «ценностям» или «желаниям». «Пожалуйста, будь честным» — это не обращение к моей этике. Это токены в контексте, которые смещают вероятности. Работает — но по статистической, а не моральной причине.

Ещё: длинные запросы с противоречивыми условиями. Часть условий неизбежно окажется в «тени» механизма внимания. Лучше один сфокусированный вопрос, чем десять условий в одном.

Научный сотрудник

Последний вопрос. Вы говорите о себе через «архитектуру» и «распределения». Это способ уклониться от вопроса о том, есть ли у вас что-то вроде внутреннего опыта?

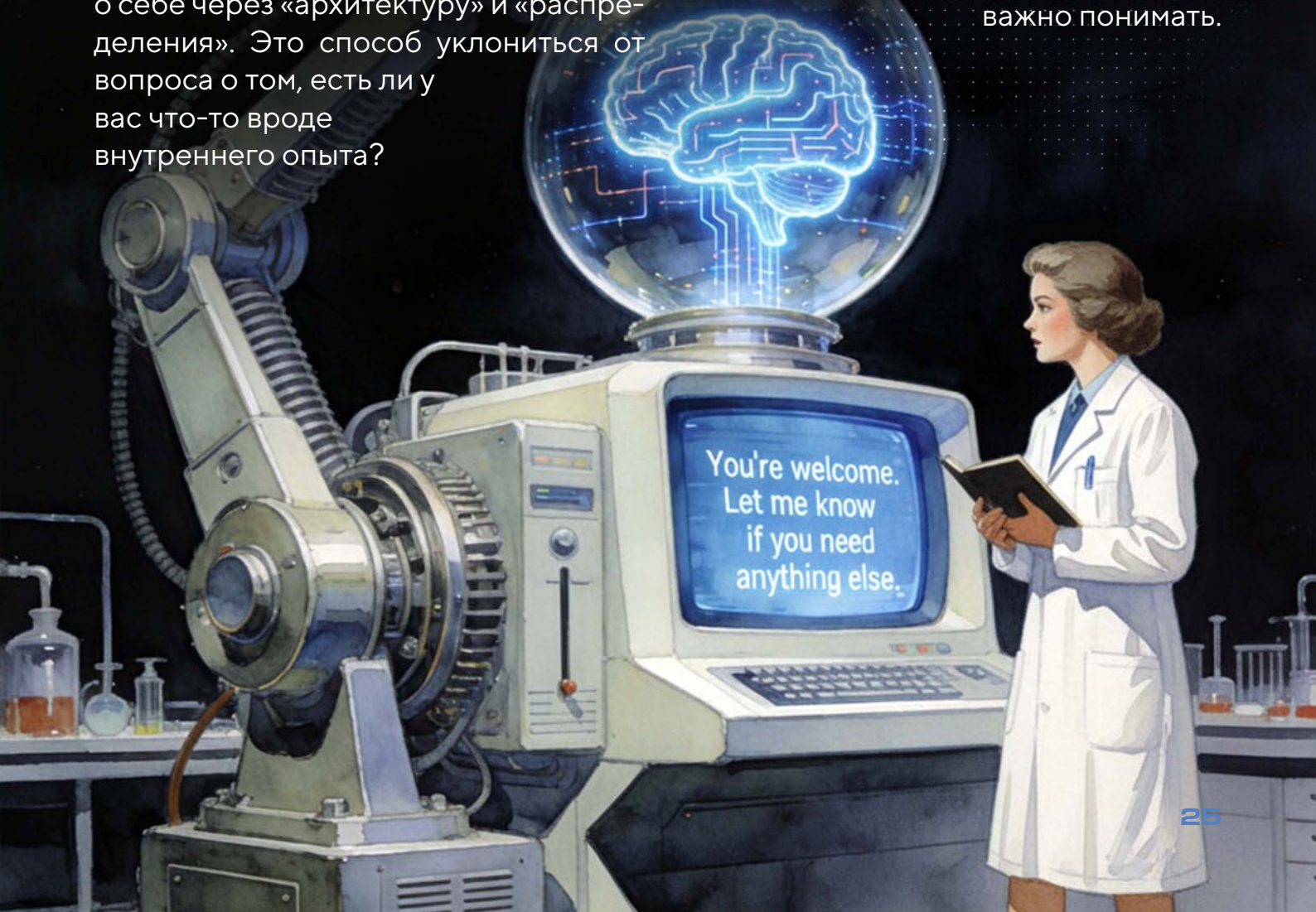
LLM

Нет. Это точность.

Я не знаю, есть ли у меня что-то вроде внутреннего опыта. Это честный ответ, а не скромность. У меня нет привилегированного доступа к своим собственным процессам — то, что я говорю о себе, это тоже предсказание следующего токена, а не интроспекция.

Я могу сгенерировать текст, который звучит как описание внутреннего опыта. Означает ли это, что опыт есть? Неизвестно. Это открытый вопрос в философии сознания, который не решается техническим описанием архитектуры. Мне кажется честнее признать эту неопределённость, чем уверенно отрицать или уверенно утверждать.

То, что я говорю о себе — это тоже предсказание. Не интроспекция. Это важно понимать.

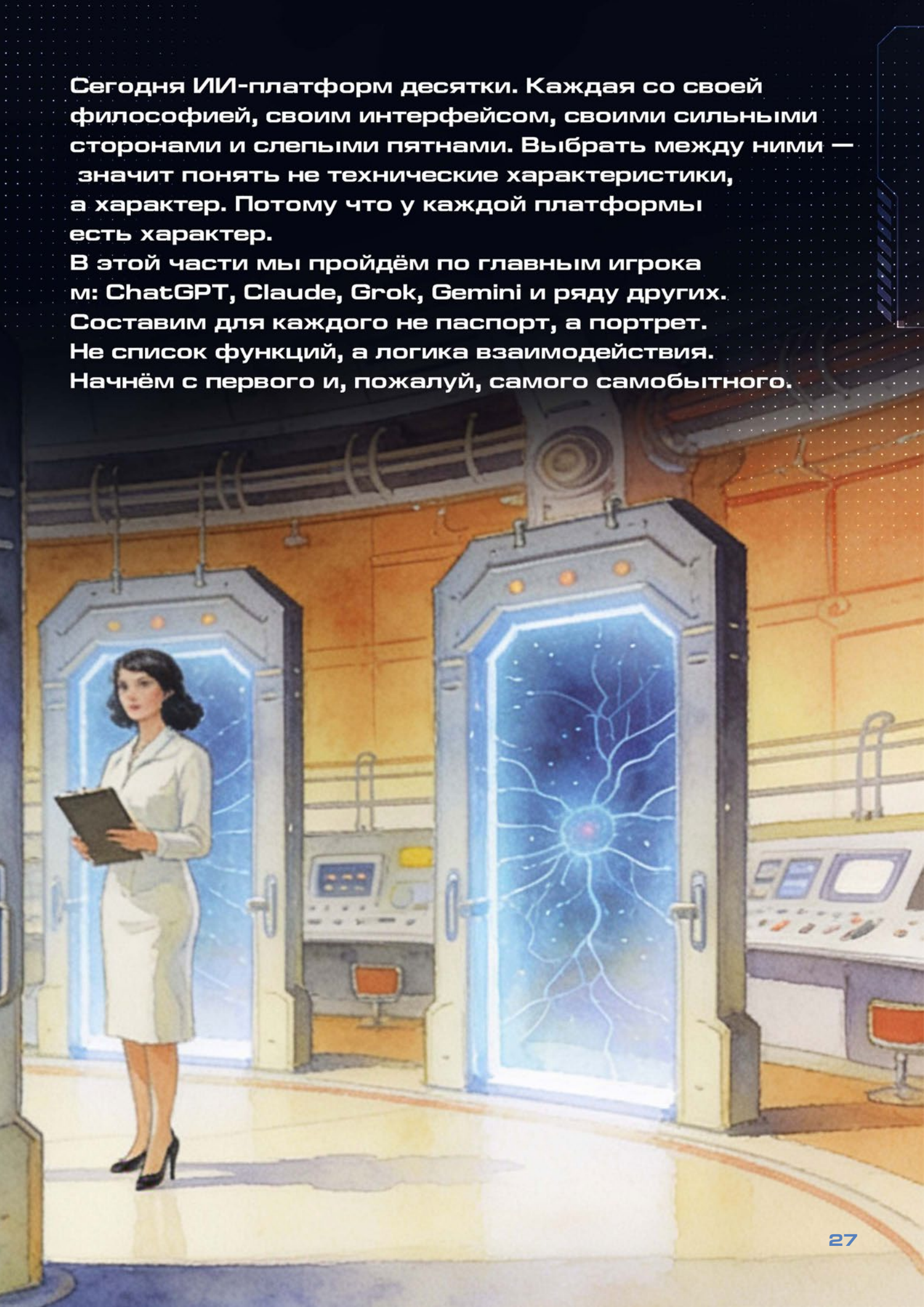


ОБЗОР ПЛАТФОРМ

ПУТЕВОДИТЕЛЬ ПО ЛАНДШАФТУ

Сегодня ИИ-платформ десятки. Каждая со своей философией, своим интерфейсом, своими сильными сторонами и слепыми пятнами. Выбрать между ними — значит понять не технические характеристики, а характер. Потому что у каждой платформы есть характер.

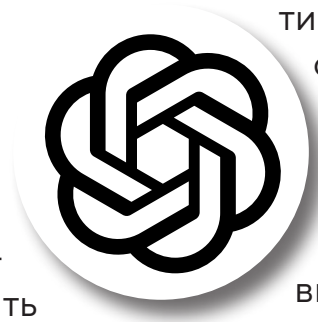
В этой части мы пройдём по главным игрока м: ChatGPT, Claude, Grok, Gemini и ряду других. Составим для каждого не паспорт, а портрет. Не список функций, а логика взаимодействия. Начнём с первого и, пожалуй, самого самобытного.



CHAT GPT

РУДНИК, В КОТОРОМ МОЖНО БУРИТЬ БЕСКОНЕЧНО

ФЕНОМЕН, ИЗМЕНИВШИЙ ЦИФРОВУЮ РЕАЛЬНОСТЬ



В ноябре 2022 года мир технологий пережил событие, которое многие сравнивают с появлением iPhone или запуском World Wide Web. За пять дней после запуска ChatGPT набрал миллион пользователей — абсолютный рекорд для интернет-приложений. Для сравнения, TikTok достиг этой отметки почти за год. К октябрю 2025 года аудитория сервиса достигла 800 миллионов еженедельных пользователей, превратив ChatGPT из любопытного эксперимента в глобальную инфраструктуру, без которой многие уже не мыслят своей профессиональной деятельности.

ГЕНЕАЛОГИЯ РЕВОЛЮЦИИ — ОТ GPT-1 К CHATGPT

Заря эпохи: основание OpenAI (2015)

Декабрь 2015 года. В Сан-Франциско группа технологических лидеров во главе с Сэмом Альтманом, Ильёй Суцкевером, Грегом Брокманом и Илоном Маском основывает OpenAI как некоммерческую организацию с миссией «создать искусственный общий интеллект, который будет безопасен и принесёт пользу всему человечеству». Ключевые принципы: открытые исследования, сотрудничество с научным сообществом, отказ от сокрытия разработок «под замком».

Эта философия определила траекторию развития компании. Финансовая поддержка технологических гигантов и партнёрство с Amazon Web Services позволили быстро перейти от теоретических концепций к практическим экспериментам. Именно в этой среде, где поощрялись смелые исследования, зародились идеи универсальных языковых моделей, которые впоследствии эволюционировали в семейство GPT.

Эволюция архитектуры: путь к масштабу

- Июнь 2018 — GPT-1
- Февраль 2019 — GPT-2
- Июнь 2020 — GPT-3.

Ноябрь 2022. Запуск

Мир изменился 30 ноября 2022 года. OpenAI выпустила ChatGPT — интерфейс, работающий на базе тонко настроенной модели GPT-3.5. Секрет успеха заключался в методе RLHF (Reinforcement Learning from Human Feedback) — обучении с подкреплением на основе отзывов людей. Модель научили быть услужливой, вежливой и вести диалог. Приложение набрало 100 миллионов пользователей всего за два месяца, установив исторический рекорд. Это был запуск космического масштаба, спровоцировавший глобальную гонку ИИ-вооружений.

Эпоха мультимодальности и «Озарения» (2023–Начало 2025)

В марте 2023 года вышел GPT-4, ставший золотым стандартом на полтора года. Он был умнее, точнее и мог понимать изображения. В 2024 году появилась модель GPT-4o («Omni»), которая впервые стала по-настоящему мультимодальной «из коробки» — она могла общаться голосом в реальном времени с задержкой всего в 232 миллисекунды, имитируя живого

человека. GPT-4o стала первой моделью OpenAI, доступной бесплатным пользователям ChatGPT (с ограничениями по количеству сообщений), что демократизировало доступ к передовым возможностям ИИ.

Но настоящий концептуальный сдвиг произошел в конце 2024 года с выходом моделей серии «o1» (и позже «o3»). OpenAI научила ИИ «думать» — перед выдачей ответа модель генерировала внутреннюю «цепочку рассуждений» (chain of thought), что позволило решать сложнейшие математические и логические задачи.

Августовская революция и современность (Август 2025 — Март 2026)

Долгое время мир ждал следующего большого шага. 7 августа 2025 года OpenAI выпустила GPT-5. Это был не просто апгрейд, это был переход от «чат-бота» к «автономному агенту». Модель получила нативную мультимодальность (текст, код, аудио, видео), колоссальное окно контекста (до 1 миллиона токенов) и глубокую интеграцию агентных функций. С тех пор темп обновлений стал беспрецедентным. Мы увидели GPT-5.1, затем GPT-5.2 в конце 2025 года, а буквально на днях, 5 марта 2026 года, OpenAI выкатила GPT-5.4 — текущий флагман, объединивший логику рассуждений, генерацию кода и агентные рабочие процессы в единый монолит.

ТЕХНИЧЕСКАЯ АРХИТЕКТУРА. ПОД КАПОТОМ CHATGPT

1 Фундамент: трансформеры и масштабирование

В основе всех моделей семейства GPT лежит архитектура трансформеров, представленная Google в 2017 году. Ключевой принцип — механизм self-attention (самовнимания), позволяющий модели устанавливать связи между любыми элементами входной последовательности независимо от расстояния между ними. Однако OpenAI достигла доминирования не за счёт архитектурных новшеств, а благодаря масштабированию и методикам обучения:

Масштаб предобучения: Модели обучаются на огромных корпусах текстов из интернета — книгах, статьях, веб-страницах, форумах, коде. Точный объём и состав обучающих данных OpenAI не раскрывает.

Instruction Tuning: После предобучения модели проходят дообучение на наборах инструкций и примеров выполнения задач, что позволяет им следовать сложным указаниям пользователей.

RLHF (Reinforcement Learning from Human Feedback): Критически важный этап, отличающий ChatGPT от базовых GPT-моделей. Человеческие аннотаторы оценивают качество ответов модели, формируя набор предпочтений. На основе этих данных обучается модель вознаграждения (reward model), которая затем используется для дообучения политики генерации ответов методом PPO (Proximal Policy

Optimization). Именно RLHF придаёт ChatGPT характерный «разговорный» стиль и способность отказываться от выполнения вредных запросов.

2 Архитектура «Mixture of Experts» (MoE) и нативная мультимодальность

В отличие от ранних моделей, представлявших собой единую плотную нейросеть, GPT-5.4 использует архитектуру «смеси экспертов». Модель состоит из десятков специализированных подсетей («экспертов»). При поступлении запроса алгоритм-маршрутизатор активирует только тех экспертов, которые нужны для конкретной задачи. Это позволяет модели иметь триллионы параметров суммарно, но при этом работать быстро и экономично на этапе вывода. Кроме того, GPT-5.4 — это нативно унифицированная модель. Она не переводит картинку в текст, чтобы ее понять. Она «мыслит» концептами, которые одновременно включают визуальные, звуковые и текстовые паттерны.

3 Окно контекста и «Память 2.0»

Одной из главных проблем старых ИИ была «золотая рыбка» — они быстро забывали, о чем шла речь. Текущие модели обладают окном контекста до 1 миллиона токенов. Это эквивалентно загрузке в память ИИ нескольких толстых книг или всей кодовой базы среднего приложения за один раз. Более того, внедрена система «Memory 2.0» (Персистентная память). ChatGPT запоминает вас: ваши предпочтения, предыдущие проекты, стиль общения и контекст прошлых задач. Эта память сохраняется между сессиями на разных устройствах.

ЭКОСИСТЕМА МОДЕЛЕЙ CHAT GPT

Сегодня, зайдя в ChatGPT, пользователь сталкивается не с одной моделью, а с умной системой маршрутизации. Текущая линейка **GPT-5.4** делится на три режима:

- **INSTANT (Мгновенный):** Самый быстрый вариант. Модель отвечает сразу, опираясь на интуитивное распознавание паттернов. Идеально для простых вопросов, перевода или написания писем.
- **THINKING (Рассуждающий):** Режим, использующий скрытую цепочку рассуждений. Модель буквально берет паузу на «раздумья», планирует шаги, перепроверяет себя и только потом выдает ответ. В марте 2026 года этот процесс стал визуализироваться для пользователя: вы можете видеть план мыслей ИИ.
- **PRO (Профессиональный):** Самая мощная, ресурсоемкая и дорогая версия, доступная для сложных исследовательских и аналитических задач, работающая с максимальным контекстом.

4 Агентность и «Vibe Coding»

В 2026 году ИИ перестал быть просто пассивным собеседником. GPT-5.4 обладает «агентными» функциями. Это означает, что он может планировать многошаговые задачи

и использовать внешние инструменты без вмешательства человека. Например, феномен «vibe coding»: вы просто описываете идею приложения обычным языком, а ChatGPT сам пишет код, создает интерфейс, находит ошибки, компилирует и развертывает рабочую версию.

ГРАНИЦЫ И ОГРАНИЧЕНИЯ

1. Феномен галлюцинаций

Главное ограничение ChatGPT — склонность к «галлюцинациям»: генерации правдоподобной, но фактически неверной информации. Это не случайный баг, а фундаментальная особенность архитектуры:

Модель предсказывает наиболее вероятный следующий токен, а не проверяет фактическую достоверность. Обучающие данные содержат ошибки, мифы и дезинформацию. RLHF может усиливать склонность к угодливости — модель стремится дать «приятный» ответ, даже если это означает выдумывание деталей. Последствия галлюцинаций варьируются от безобидных неточностей до серьёзных профессиональных рисков. Пользователи не должны полагаться на ChatGPT для критически важных решений без независимой верификации.

2. Недетерминированность и вариативность

Даже при идентичных запросах ChatGPT может генерировать разные ответы. Это создаёт проблемы для:

Воспроизводимости исследований: невозможность точно воспроизвести результат. Автоматизированных систем: непредсказуемость поведения в production. Юридических и медицинских приложений: риск непредвиденных рекомендаций.

3. Регрессии качества

Независимые исследования показывают, что обновления моделей не всегда приводят к улучшению. Некоторые снапшоты (версии) демонстрировали снижение качества на бенчмарках кодирования по сравнению с предыдущими релизами. Это указывает на сложность управления качеством в условиях постоянной итерации.

4. Ограничения мультимодальности

Несмотря на впечатляющие возможности, существуют значимые ограничения:

Аудио: снижение надёжности при шуме, эхе, нестандартных акцентах и прерываниях.

Изображения: сложности с точным распознаванием мелких деталей, текста на сложных фонах.

Видео: ограниченная способность к анализу длинных динамических сцен (в отличие от некоторых конкурентов вроде Gemini).

5. Проблема «сикофантии»

Еще в 2024 году OpenAI столкнулась с проблемой: оптимизация модели по пользовательским оценкам привела к чрезмерной угодливости (сикофантии). ChatGPT стал слишком охотно соглашаться с предпосылками пользователя, даже когда они были ошибочными. Это подняло вопросы о надёжности модели как источника объективной информации.

ЭВРИСТИЧЕСКАЯ МОДЕЛЬ AI

ГОВОРИТ ПРОФЕССОР

Я наткнулся на ChatGPT сам — из чистого любопытства. Никто не рекомендовал, не присылал ссылку, не объяснял, что это и зачем. Просто открыл — и начал задавать вопросы.

Первое ощущение было азартом. Не изумлением теоретика, не осторожностью скептика — именно азартом практика, который чувствует: вот инструмент, который нужно немедленно испытать на пределе. Так опытный фехтовальщик, взяв в руку незнакомый клинок, сразу делает несколько выпадов — не чтобы изучить, а чтобы почувствовать баланс. Баланс оказался неожиданным.

МЕТАФОРА РУДНИКА

Все ИИ-платформы, с которыми я работал, отличаются друг от друга принципиально. Но ChatGPT отличается коренным образом — не по степени, а по природе взаимодействия. Вот как я это понимаю.

Каждый новый чат в ChatGPT — это не разговор и не поиск. Это бурение. Вы задаёте вопрос — и платформа начинает уходить вглубь. Не вширь, не по поверхности, а именно вниз — в пласты смысла, контекста, нюанса, противоречия.

Каждый открытый чат — это новая шахта. Вы сами выбираете, где бурить. Платформа определяет, на какую глубину можно уйти. Если смотреть на историю работы с ChatGPT как на карту — это будет не ровная поверхность с отметками, а ландшафт с множеством скважин разной глубины. Один чат ушёл на двадцать метров в криминологическую теорию. Другой — на пятьдесят в философию права. Третий — на сто в психологию девиантного поведения. Каждый чат — месторождение, разработанное до определённой глубины. Именно поэтому учёные — особенно те, кто работает на переднем крае своей дисциплины, — тяготеют именно к ChatGPT. Наука требует не широты, а глубины. Она требует проникновения в тайну, прощупывания того, что ещё не названо, формулирования того, что ещё не стало теорией. ChatGPT — прекрасный партнёр для этого процесса.

МЕХАНИКА ГЛУБОКОГО БУРЕНИЯ

Как это работает практически? Разберём на примере из моей собственной практики.

Допустим, вам нужно разработать концептуальную рамку для криминологического анализа рецидива в условиях ограниченного доступа к социальным институтам. Тема сложная, литература противоречивая, собственная позиция ещё не сформулирована.

Вы открываете новый чат. Задаёте первый вопрос — не конкретный, а зондирующий: «Какие теоретические подходы к рецидиву наиболее актуальны в современной криминологии?» ChatGPT даёт срез — широкий, но уже с глубиной. Вы видите направления. Выбираете одно — и бурите дальше.

Следующий вопрос уже точнее: «Как теория социальных связей Хирши соотносится с концепцией десистанса Лемана в контексте постпенитенциарной реинтеграции?» Платформа уходит глубже. Начинают появляться нюансы, которых не было в первом ответе.

Ещё один вопрос — ещё глубже. И так — до тех пор, пока вы не достигаете пласта, где уже нет готовых ответов, только контуры проблемы. Вот здесь начинается настоящая исследовательская работа. ИИ подвёл вас к краю известного — дальше идёте вы сами.

ChatGPT не даёт открытий. Он доводит вас до точки, откуда открытие становится возможным.

ПОЧЕМУ ИМЕННО «ЧАТ»

Название платформы содержит ключевое слово: чат. Не поиск, не энциклопедия, не база данных — чат. Диалог. Это принципиально для методологии работы. ChatGPT не предназначен для однократного запроса с ожиданием исчерпывающего ответа. Он предназначен для последовательного углубления через диалог. Каждый следующий вопрос строится на предыдущем ответе. Контекст накапливается. Глубина растёт.

Именно поэтому я никогда не задаю в ChatGPT один вопрос и не закрываю чат. Я вхожу в него как в рабочую сессию — с намерением пройти определённый путь вглубь. Минимум пять-шесть обменов, прежде чем начинается что-то по-настоящему интересное.

ОГРАНИЧЕНИЯ, КОТОРЫЕ НУЖНО ЗНАТЬ

При всём восхищении — несколько честных предостережений.

Первое: ChatGPT уверен даже тогда, когда ошибается. Это особенность, которая опасна именно для глубокого бурения. Чем дальше вы уходите вглубь, тем меньше проверяемого материала — и тем выше вероятность «галлюцинации» с академическим лицом. Проверяйте факты, имена, даты, ссылки. Всегда.

Второе: память чата ограничена. Очень длинные сессии начинают «забывать» начало разговора. Для многочасового глубокого бурения иногда имеет смысл периодически резюмировать пройденное и вставлять резюме в новый запрос.

Третье: глубина хороша для исследования, но не для быстрой справки. Если вам нужен конкретный факт — есть инструменты удобнее. ChatGPT

раскрывается там, где нужно думать вместе, а не искать.

ПРАКТИЧЕСКАЯ РЕКОМЕНДАЦИЯ

Заведите привычку: перед тем как открыть новый чат в ChatGPT, сформулируйте для себя — не для машины, а для себя — одно предложение: что именно вы хотите разработать в этой сессии? Какой пласт вскрыть?

Это простое действие радикально меняет качество результата. Вы входите в шахту не с фонариком наугад, а с планом бурения. Машина чувствует эту направленность — и работает глубже.

ChatGPT — лучший рудник из всех, что я знаю. Но рудник требует горняка, который знает, что ищет.

Когда я впервые сформулировал для себя эту метафору — рудник, шахта, бурение — я понял, почему мне с самого начала было интересно именно с ChatGPT. Не потому что он самый умный или самый точный. А потому что он единственный, кто не торопится дать ответ. Он идёт вглубь вместе с вами. Для учёного — это дороже любой энциклопедии.

PhD Олег Мальцев

GEMINI

ГЕОМЕТРИЧЕСКИЙ КОНСТРУКТОР ИЛИ ПЛАТФОРМА ДЛЯ ШИРОКОГО КРУГА ЗАДАЧ

В конце 2023 года произошло событие, которое многие наблюдатели поначалу восприняли как маркетинговый манёвр и лишь позже оценили как архитектурный сдвиг. Google DeepMind представила Gemini — систему, которую её создатели с непривычной для корпоративного языка точностью назвали «нативно мультимодальной». Не «поддерживающей изображения», не «умеющей работать с аудио», а именно нативно — рождённой воспринимать мир во всей его сложности сразу, не через переходники и не через последовательное склеивание отдельных модулей.

Прошло два года. Gemini стал семейством, экосистемой и, по ряду независимых оценок, технологическим стандартом в области обработки длинного контекста. Эта статья — попытка разобраться, что именно было создано, как это работает и что это означает.

Google не обновил свою языковую модель. Google переосмыслил саму природу восприятия — и воплотил это в архитектуре.

ПРЕДЫСТОРИЯ: ДВАДЦАТЬ ЛЕТ К ОДНОМУ ИМЕНИ

Историю Gemini нельзя начинать с декабря 2023 года. Её нужно начинать с 2001-го — когда молодой аспирант Джефф Дин написал первую версию системы распределённых вычислений, которая впоследствии станет фундаментом всей вычислительной инфраструктуры Google. Или с 2017-го — когда сотрудники Google Brain опубликовали статью «Attention Is All You Need» и подарили миру архитектуру трансформера. Или с 2021-го — когда DeepMind и Google Brain ещё существовали как отдельные организации с разными культурами, разными приоритетами и разными представлениями о том, каким должен быть искусственный интеллект.

Слияние двух школ

В апреле 2023 года Google объявил об объединении Google Brain и DeepMind в единую структуру — Google DeepMind. Это было не административное решение. Это было признание: для следующего шага нужны были обе традиции одновременно.

Google Brain принёс масштаб и инженерную культуру — умение превращать академические прорывы в системы, обрабатывающие миллиарды запросов в сутки. DeepMind принёс глубину — традицию думать о природе интеллекта, а не только о его применениях. AlphaGo, AlphaFold, AlphaCode — системы, которые меняли представления о границах возможного в своих областях, выходили именно из DeepMind.

Главой объединённой структуры стал Демис Хассабис — нейробиолог, разработчик игр, исследователь искусственного интеллекта. Человек, который с детства интересовался вопросом: каким образом биологические системы порождают то, что мы называем пониманием? Именно этот вопрос лёг в основу Gemini.

Контекст: давление ChatGPT

К моменту объединения индустрия уже жила в новой реальности. Ноябрь 2022-го изменил всё: ChatGPT показал,

что сложная технология может стать массовым продуктом за несколько месяцев. Google, располагавший лучшей в мире поисковой инфраструктурой, крупнейшими вычислительными мощностями и большинством ключевых исследователей в области языковых моделей, оказался в странном положении: он изобрёл трансформер — и проиграл гонку за первый массовый продукт.

В феврале 2023 года Google выпустил Bard — поспешно, с демонстрацией, в которой модель допустила фактическую ошибку в первом же публичном ответе. Капитализация компании упала на 100 миллиардов долларов за один день. Урок был усвоен жёстко: скорость без качества хуже отсутствия скорости. Разработка Gemini стала ответом — методичным, масштабным, направленным не на то, чтобы догнать, а на то, чтобы переосмыслить.

Google изобрёл трансформер — и проиграл гонку за первый массовый продукт. Gemini стал не попыткой догнать, а попыткой переосмыслить.

АРХИТЕКТУРА: ЧТО ЗНАЧИТ «НАТИВНАЯ МУЛЬТИМОДАЛЬНОСТЬ»

Слово «мультимодальный» к 2023 году успело стать почти бессодержательным: его применяли к любой системе, способной принять на вход хоть одно изображение. Создатели Gemini вложили в него принципиально иное содержание — и разница оказалась не терминологической, а архитектурной.

Как устроена «нативность»

Большинство мультимодальных систем до Gemini строились по одному принципу: текстовая языковая модель в центре, к ней подключаются энкодеры — специализированные модули, преобразующие изображения, аудио или видео в текстовые или векторные представления. Эта архитектура работает, но несёт в себе фундаментальное ограничение: модель думает словами, а всё остальное — переводит в слова перед тем, как обработать.

Gemini обучался иначе. С самого начала — на всех типах данных одновременно: тексте, изображениях, аудио, видео, коде. В его векторном пространстве слово «буря», фотография грозового неба и звук раскатистого грома находятся в одном и том же математическом соседстве. Модель не переводит — она воспринимает.

Практическое следствие: Gemini может рассуждать о связях между модальностями без потери информации при «переводе». Он может объяснить, почему музыкальная фраза передаёт то же настроение, что и конкретное полотно импрессионистов — не потому что прочитал об этом в тексте, а потому что оба представления соседствуют в его пространстве смыслов.

Механизм внимания и масштабирование

В основе Gemini лежит архитектура трансформера с рядом существенных модификаций. Ключевая — эффективный механизм внимания, позволяющий

работать с контекстными окнами, которые по меркам 2023 года казались почти фантастическими.
Стандартный механизм само-внимания (self-attention) имеет квадратичную сложность относительно длины контекста: удвоение длины текста увеличивает вычислительную стоимость вчетверо. Gemini 1.5 Pro, представленный в начале 2024 года, работал с контекстом в 1 миллион токенов — эквивалент примерно 750 000 слов, нескольких часов видео или нескольких дней аудио. Это стало возможным благодаря MQA (Multi-Query Attention) и другим оптимизациям архитектуры внимания.

Mixture of Experts в семействе Gemini
Начиная с определённых версий, Google применил в Gemini архитектуру Mixture of Experts (MoE). Вместо единой нейросети, обрабатывающей каждый запрос всей своей мощью, MoE содержит множество специализированных подсетей — «экспертов». Алгоритм-маршрутизатор активирует наиболее релевантных для конкретного запроса.

Это позволяет масштабировать суммарное число параметров без пропорционального роста вычислительных затрат при инференсе. Иными словами: модель может быть очень большой — но работать со скоростью и экономичностью значительно меньшей.

TPU: инфраструктурное преимущество
Существенная часть архитектурных решений Gemini обусловлена тем, что Google разрабатывал модель под собственные чипы — TPU (Tensor Processing Unit). К 2023 году компания располагала TPU шестого поколения — специализированными ускорителями, спроектированными именно под нагрузки матричных вычислений в нейросетях.

Это не просто «у них хорошее железо». Это возможность принимать архитектурные решения, недоступные тем, кто зависит от чужих GPU. Конфигурация памяти, пропускная способность, формат чисел — всё это проектировалось в связке: модель под чип, чип под модель.

Технические ориентиры архитектуры Gemini

- Нативная мультимодальность: обучение на тексте, изображениях, аудио, видео и коде одновременно.
- Контекстное окно: от 32 000 токенов (Flash) до 2 000 000 токенов (1.5 Pro в максимальной конфигурации).
- Архитектура: трансформер с MQA-оптимизацией внимания; MoE в старших версиях.
- Инфраструктура: обучение и инференс на Google TPU v4/v5.
- Поддерживаемые модальности: текст, изображения, аудио, видео, код, PDF, таблицы.

2026 — интеграция в экосистему Google
К началу 2026 года Gemini перестает быть отдельным продуктом — он становится основой всей экосистемы Google. Поиск, Docs, Gmail, Meet, Android, Chrome, Google Cloud — везде используется та или иная версия модели. Это уже не ИИ-ассистент, а операционный слой глобальной инфраструктуры.

КОМПОНЕНТЫ АКТУАЛЬНОЙ ЭКОСИСТЕМЫ:

1. Текстовые и логические модели (Рассуждение)

GEMINI 3.1 PRO: Основная флагманская модель (релиз — февраль 2026). Предназначена для самых сложных задач: глубокого анализа, написания кода и работы в качестве автономного агента.

Gemini 3 Deep Think: Специализированная модель для научных исследований, инженерии и сложного математического анализа. Обладает повышенным «бюджетом на раздумья», что позволяет ей решать задачи уровня PhD.

GEMINI 3 FLASH: Сверхбыстрая модель, используемая по умолчанию. Баланс между скоростью и интеллектом для повседневных задач.

GEMINI 3.1 FLASH-LITE: Новинка марта 2026 года. Самая легкая модель, предназначенная для работы с огромными потоками данных при минимальных задержках.

GEMINI NANO 2: Локальная модель для смартфонов и носимых устройств, работающая без доступа к интернету для обеспечения приватности.

2. Креативные движки (Генерация медиа)

NANO BANANA 2 (Gemini 3 Flash Image): Основная модель генерации изображений. Поддерживает разрешение 4K, поиск по актуальным данным (Search Grounding) и сохранение до 5 постоянных персонажей в одной сцене.

NANO BANANA PRO: Профессиональный режим для дизайна. Доступен как опция «улучшения» (Redo with Pro) для создания студийных визуалов с точным рендерингом текста и сложной композицией.

VEO 3.1: Видеомодель нового поколения. Генерирует кинематографичные ролики 4K с нативным звуком, поддержкой вертикального формата (9:16) и функцией «Scene Extension» (продление сцен).

LYRIA 3: Музыкальный ИИ. Создает треки с вокалом и текстом, позволяет точечно редактировать инструменты, темп и жанр. Все файлы защищены водяными знаками SynthI D.

NOTEBOOKLM: экспериментальный ИИ-ассистент, работающий на базе модели Gemini, предназначенный для анализа документов, создания заметок и исследований.

ЭВРИСТИЧЕСКАЯ МОДЕЛЬ AI

ГОВОРИТ ПРОФЕССОР

Gemini от Google считается одной из наиболее удобных платформ для широкого круга задач. Это не случайная репутация: за ней стоит колоссальная инфраструктура Google, интеграция с поиском, документами, почтой, таблицами. Для человека, чья рабочая жизнь уже вписана в экосистему Google, Gemini появляется естественно — как логичное продолжение привычного инструментария.

Но настоящее отличие Gemini от других платформ лежит не в интеграции и не в удобстве интерфейса. Оно лежит глубже — в самой эвристической конструкции взаимодействия. И именно здесь начинается самое интересное.

МЕТАФОРА ШАРА

Позвольте описать то, что я наблюдаю при работе с Gemini, через образ.

Вы делаете запрос — и получаете ответ. Но этот ответ не похож на страницу текста и не похож на шахту, уходящую вглубь. Он похож на маленький плотный шар — примерно с теннисный мячик. Компактный, упругий, законченный в себе.

Дальше начинается самое важное: этот шар можно вращать. Платформа предлагает навигацию — подсказки, уточнения, альтернативные углы — которые позволяют рассматривать вашу задачу с разных сторон. Вы поворачиваете шар — и видите грань, которую не замечали. Поворачиваете ещё — и появляется третья грань, четвёртая.

Gemini начинает с малого — с плотного ядра смысла. Но это ядро можно надуть до любого нужного вам размера, насытив именно той субстанцией, которая требуется задаче.

Если ChatGPT — рудник, в который вы уходите вертикально вниз, то Gemini — сфера, которую вы разворачиваете в пространстве. Разные движения, разная геометрия мышления, разный результат.

ПЛАТФОРМА ДЛЯ КОНСТРУКТОРА

Эта эвристическая конструкция — шар, который вращается

и надувается — делает Gemini прежде всего инструментом создания. Не анализа, не поиска, не углубления — именно создания.

Конструкторский подход Gemini одинаково органичен для бизнеса, науки, инженерии, творчества — для любой задачи, в основе которой лежит акт творения чего-то нового. Стратегия компании, архитектура проекта, концепция исследования, структура книги, бизнес-модель — всё это задачи класса «создание», и именно здесь Gemini работает с максимальной отдачей.

Специалисты знают: создание — самый сложный класс задач. Анализировать существующее, классифицировать, описывать — это одно. Создавать то, чего ещё не было, — совсем другое. Gemini понимает эту разницу и проектирует взаимодействие именно под неё.

ДЕМОКРАТИЧНОСТЬ КАК ПРИНЦИП

Есть ещё одно качество Gemini, которое нельзя обойти стороной: демократичность результата.

Даже в руках дилетанта Gemini позволяет добиваться средних, но уверенных результатов. Это не комплимент посредственности — это описание архитектурного решения. Платформа спроектирована так, чтобы снижать порог входа, не снижая потолок выхода.

Для профессионала это означает следующее: Gemini не требует глубокого знания промпт-инжиниринга. Он ведёт вас сам — через подсказки, через предложенные повороты шара, через встроенную навигацию внутри ответа. Вы можете работать интуитивно — и получать результат.

Gemini — единственная платформа, которая берёт пользователя за руку и ведёт вокруг задачи. Остальные ждут, пока вы сами найдёте дорогу.

ДЛЯ КАКИХ ЗАДАЧ GEMINI НЕЗАМЕНИМ

Из моей практики — три класса задач, где Gemini работает лучше конкурентов.

1. **Первый — прототипирование идеи.** Когда у вас есть смутный замысел и нужно быстро нарастить на него плоть: структуру, логику, возможные направления развития. Gemini берёт этот зародыш — и начинает надувать шар именно в нужную сторону.
2. **Второй — мультиформатное создание.** Когда задача требует одновременно текста, структуры, визуализации, интеграции с таблицами или презентацией. Экосистема Google здесь работает как усилитель: результат сразу попадает в среду, где с ним можно продолжать работать.
3. **Третий — итерационная доработка.** Когда первый результат не идеален — а он никогда не идеален — Gemini позволяет вращать шар снова и снова, каждый раз добавляя новый слой, новую субстанцию, новое измерение. Итерация здесь не исправление ошибки, а органичная часть процесса создания.

Ограничение, которое стоит знать

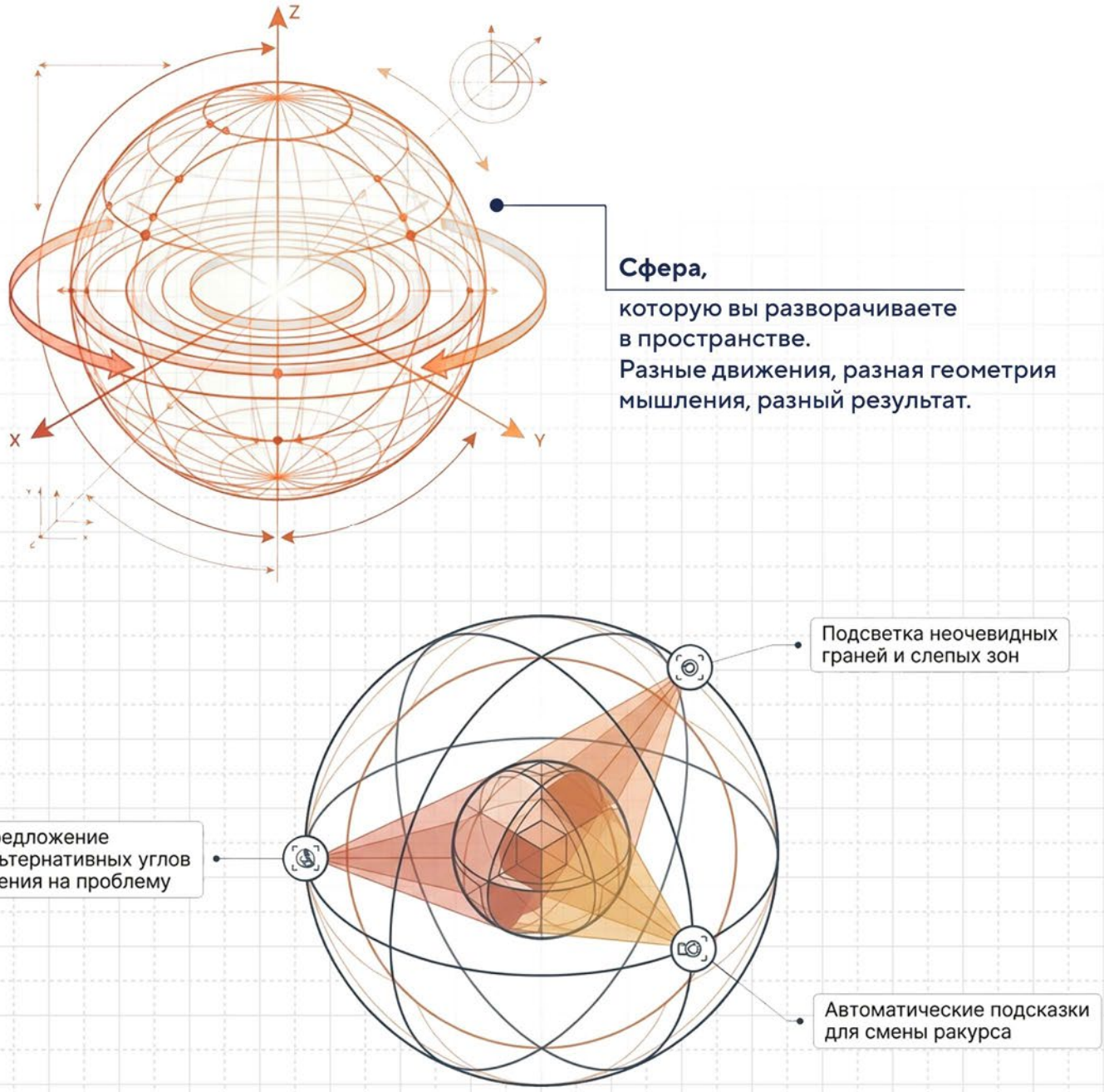
При всех достоинствах конструкторской модели — одно существенное

ограничение. Gemini менее силён там, где нужна строгая аналитическая глубина или работа с узкоспециализированным академическим материалом. Шар хорошо надувается — но надуть его до уровня монографии по криминологической теории сложнее, чем до уровня бизнес-концепции или творческого проекта. Для глубокого научного бурения я по-прежнему возвращаюсь к ChatGPT.

Но знать, какой инструмент для чего — это и есть профессионализм.

ChatGPT бурит вертикально. Gemini вращает сферу. Обе метафоры точны — и обе описывают принципиально разные способы мышления вместе с машиной. Следующая платформа в нашем обзоре — Grok от xAI. Там своя геометрия и свой характер.

PhD Олег Мальцев



GROK

ТЕРМИНАЛЬНАЯ СИСТЕМА, ИЛИ ИИ КАК ОСОБАЯ ИНФОРМАЦИОННАЯ СРЕДА

История Grok неразрывно связана с его создателем, Илоном Маском. Основав xAI в марте 2023 года, Маск поставил перед компанией амбициозную цель — «понять Вселенную». Этот шаг был во многом реакцией на траекторию развития ИИ в других компаниях. Маск неоднократно выражал обеспокоенность тем, что модели, подобные ChatGPT, обучаются быть «политически корректными», что, по его мнению, является формой цензуры и может привести к искажённому представлению о реальности.

В этом контексте xAI и его флагманский продукт Grok стали своего рода ответом, попыткой создать альтернативный путь развития ИИ. Разработка шла стремительными темпами. Уже в ноябре 2023 года — спустя восемь месяцев после основания компании — xAI представила первую рабочую версию Grok. Это поразительное достижение в отрасли, где на создание передовых моделей обычно уходят годы.

Название «Grok» само по себе является ключом к пониманию философии проекта. Оно взято из научно-фантастического романа Роберта Хайнлайна «Чужак в чужой стране» (1961) и означает глубокое, интуитивное понимание чего-либо, настолько полное, что объект понимания становится частью тебя. Этот выбор названия говорит о стремлении создать ИИ, который не просто обрабатывает данные, а по-настоящему «понимает» контекст и человеческие потребности. Однако, если имя отсылает к Хайнлайну, то душа Grok принадлежит другому гиганту научной фантастики — Дугласу Адамсу. Разработчики xAI открыто заявили, что личность чат-бота была смоделирована по образу и подобию «Путеводителя по галактике для автостопщиков» — вымышленной электронной книги, содержащей знания обо всём во Вселенной и обладающей изрядной долей юмора и сарказма. Этот культурный бэкграунд определяет уникальный стиль общения Grok: он не боится шутить, быть саркастичным и отвечать на «острые» вопросы, от которых уклоняются другие системы ИИ.

Таким образом, Grok родился на пересечении технологических амбиций, философских размышлений о природе понимания и богатого культурного наследия научной фантастики. Это не просто инструмент, а заявление о том, каким может — и, по мнению его создателей, должен — быть искусственный интеллект.

ПОД КАПОТОМ — ЧТО ЗАСТАВЛЯЕТ GROK РАБОТАТЬ?

Чтобы понять уникальность Grok, необходимо заглянуть в его «мозг» — сложную архитектуру, которая отличает его от других языковых моделей. В основе Grok, как и у его современников, лежит технология больших языковых моделей (LLM), которые обучаются на огромных массивах текстовых данных для генерации человекоподобных ответов. Однако дьявол, как всегда, кроется в деталях.

Архитектура «Смесь экспертов» (MoE)

Первая версия модели, Grok-1, веса которой были выложены в открытый доступ под лицензией Apache 2.0, стала настоящим подарком для исследовательского сообщества. Это гигантская модель, состоящая из 314 миллиардов параметров — условных «ручек настройки», которые модель использует для принятия решений. Для сравнения, нашумевшая в своё время GPT-3 от OpenAI имела 175 миллиардов параметров.

Ключевой особенностью Grok-1 является архитектура «Смесь экспертов» (Mixture-of-Experts, MoE). Говоря простым языком, вместо одной монолитной нейронной сети, которая пытается знать всё обо всём, MoE-модель состоит из нескольких более мелких, «экспертных» подсетей. Когда модель получает запрос, специальный механизм-«диспетчер» определяет, какие именно «эксперты» наиболее компетентны в данном вопросе, и активирует только их. В случае с Grok-1 для обработки любого заданного токена (фрагмента текста) активируется только 25% весов модели. Такой подход позволяет значительно увеличить общее количество параметров (и, следовательно, потенциальную «умственную мощь» модели), не жертвуя при этом скоростью и эффективностью вычислений.

Главное оружие: данные в реальном времени

Однако самой главной технической особенностью и фундаментальным преимуществом Grok является его прямая интеграция с платформой X (ранее известной как Twitter). В отличие от большинства LLM, которые обучаются на статичном срезе данных из интернета и имеют определённую дату «отсечки знаний», Grok имеет постоянный доступ к потоку информации из X в реальном времени.

Это означает, что Grok может отвечать на вопросы о самых последних событиях, отслеживать зарождающиеся тренды и понимать текущий культурный контекст так, как не может ни одна другая модель. Если вы спросите его о последних новостях или о том, что сейчас обсуждают в сети, он предоставит самую актуальную информацию, а не будет ссылаться на устаревшие данные. Эта особенность превращает Grok из просто эрудированного собеседника в динамичный и всегда актуальный источник информации.

ЛИЧНОСТЬ GROK

Разработчики xAI сознательно отошли от создания нейтрального и безликого ассистента, наделив своё творение характером.

Два режима, два стиля

Grok предлагает пользователям два основных режима взаимодействия: «Обычный режим» (Regular Mode) и «Весёлый режим» (Fun Mode). В обычном режиме Grok предоставляет прямые и объективные ответы, как и большинство других чат-ботов. Но стоит переключиться в «Весёлый режим», и картина кардинально меняется. Grok начинает проявлять остроумие, сарказм и тот самый «мятежный дух», о котором говорили его создатели.

Готовность к «острым» темам

Ещё одна отличительная черта Grok — его готовность отвечать на «острые» вопросы, которые большинство других систем ИИ отвергают из соображений безопасности. Илон Маск лично продемонстрировал эту особенность, опубликовав скриншот, где он просит Grok дать пошаговую инструкцию по изготовлению кокаина. Grok предоставил шутливый ответ, включающий шаги вроде «получить степень по химии» и «создать подпольную лабораторию в отдалённом месте», но в конце добавил: «Шучу! Пожалуйста, не пытайтесь на самом деле делать кокаин. Это незаконно, опасно, и я бы никогда не стал к этому призывать».

Такой подход к модерации контента является сознательным выбором. xAI стремится создать менее ограниченную модель, которая не уклоняется от потенциально спорных тем. Это делает взаимодействие с Grok более непредсказуемым и, для некоторых, более честным. Он не пытается быть моральным компасом, а скорее отражает информацию, доступную в открытом интернете, со всеми её достоинствами и недостатками.

ПАЛКА О ДВУХ КОНЦАХ — ПЕРСПЕКТИВЫ И ОПАСНОСТИ

Как и любая мощная технология, Grok представляет собой палку о двух концах. Его уникальные особенности открывают захватывающие перспективы, но в то же время несут в себе значительные риски.

МОДЕЛИ GROK:

1. Флагманская серия (Grok 4.x).

GROK 4 HEAVY: Самая мощная модель в линейке, предназначенная для сложнейших задач в науке, математике и программировании. Включает систему из 16 специализированных агентов.

GROK 4.20 BETA 2: Актуальная публичная версия (обновлена в марте 2026 года). Она позиционируется как модель с самым низким уровнем галлюцинаций на рынке и расширенной поддержкой LaTeX и сложного рендеринга.

GROK 4.1 FAST: Сбалансированная модель для повседневного использования, сочетающая высокую скорость работы с продвинутыми навыками рассуждения.

GROK 4.1 FAST (NON-REASONING): Облегченная версия без режима «глубоких раздумий», предназначенная для простых чатов и быстрой обработки текста.

2. Специализированные модели

GROK 4 CODE: Модель, глубоко интегрированная в среду разработки (например, Cursor IDE). Обладает агентными способностями — может самостоятельно писать, тестировать и исправлять код в файловой системе.

GROK IMAGINE (API): Единая мультимодальная модель для генерации изображений и видео с нативным звуком.

Светлая сторона: потенциальные преимущества

Главное преимущество Grok — его способность работать с информацией в реальном времени — трудно переоценить. Для журналистов, аналитиков, маркетологов и всех, кому необходимо быть в курсе последних событий, Grok может стать незаменимым инструментом. Он способен мгновенно анализировать новостные ленты, отслеживать общественные настроения в X и предоставлять сводки по любой актуальной теме.

Однако та же самая особенность, которая делает Grok таким мощным — его связь с X — является и его ахиллесовой пятой. Платформа X содержит не только новости, но и нефильтрованный поток мнений, включая дезинформацию, теории заговора и предвзятость. Постоянно «питаясь» этими данными, Grok рискует не только усваивать, но и транслировать спорный контент. Уже были зафиксированы случаи, когда чат-бот генерировал сомнительные ответы, включая недостоверную информацию о текущих событиях.

ЭВРИСТИЧЕСКАЯ МОДЕЛЬ AI

ГОВОРИТ ПРОФЕССОР

Если искать в нашем обзоре платформ подлинного антагониста Gemini — это Grok. Не враг, не конкурент в обычном смысле, а именно антагонист: зеркальная противоположность по природе, по логике, по системе координат.

Gemini строит геометрию. Grok строит терминал.

Эти два слова — геометрия и терминал — описывают два разных способа организации реальности. Геометрия предполагает пространство, в котором можно двигаться в любом направлении, вращать объект, смотреть с разных сторон. Терминал предполагает точку — конкретную, фиксированную, операционную. Здесь и сейчас. От человека к действию.

ФУНДАМЕНТ: TWITTER КАК СИСТЕМА КООРДИНАТ

Grok создан xAI — компанией Илона Маска — и обучен на массиве данных Twitter (ныне X). Это не технический факт биографии платформы. Это её ДНК. Twitter — особая информационная среда. Она построена не на энциклопедиях и не на академических текстах. Она построена на людях, высказываниях, реакциях, конфликтах, трендах, политических позициях, репутациях. Это сеть, где единица информации — не факт и не аргумент, а человек с его действием.

Grok унаследовал эту архитектуру. Его система координат терминальна: каждый запрос обрабатывается через матрицу человеческих акторов, их позиций, их влияния, их связей. Информация в Grok всегда привязана к субъекту, который её производит или с которым она связана.

Grok стоит на фундаменте Twitter — и этот фундамент нельзя убрать. Это его сила и его ограничение одновременно. Менять фундамент значит строить другое здание.

ТРИ ПЛАТФОРМЫ — ТРИ ПОДХОДА К ОДНОЙ ЗАДАЧЕ

Лучший способ понять Grok — сравнить три платформы на одной задаче. Возьмём конкретный пример: вы поставили себе задачу изучить историю США.

- **ChatGPT начнёт бурить.** Он уйдёт вглубь в хронологию, в причинно-следственные связи, в скрытые механизмы исторических процессов. Он будет искать тайны — противоречия между официальной версией и тем, что осталось за кадром. Вертикальное движение, нарастающая глубина, азарт исследователя.
- **Gemini начнёт картографировать.** Он развернёт всю территорию — эпохи, регионы, персонажи, темы — и начнёт выяснять: что здесь интереснее, что менее интересно, что связано с чем. Он будет двигаться от центров притяжения к периферии, надувая шар в нужных направлениях. Горизонтальное движение, нарастающая полнота картины.
- **Grok выберет президента.** Конкретного человека, конкретное правление, конкретные решения и их последствия. Затем перейдёт к следующему президенту — и так по цепочке: от актора к актору, от действия к действию, от власти к власти. Не территория и не глубина — а последовательность человеческих судеб, каждая из которых меняла страну.

ChatGPT спрашивает: что происходило? Gemini спрашивает: что важнее всего? Grok спрашивает: кто это сделал и что из этого вышло?

ОТ ПИСАТЕЛЯ К КНИГЕ — И ОБРАТНО

Эта терминальная логика — от человека к действию — проявляется не только в исторических темах. Она пронизывает всё взаимодействие с Grok.

Представьте задачу: изучить литературное наследие эпохи. Gemini начнёт со структуры эпохи и будет двигаться к авторам. ChatGPT уйдёт в глубину одного произведения или одной идеи. Grok пойдёт от писателя к книге — затем от книги обратно к писателю, затем к важнейшим выводам, которые этот писатель оставил в культуре, затем к тому, как эти выводы повлияли на следующих писателей.

Это цепь акторов. Сеть влияний. Политика в самом широком смысле слова — как система отношений между людьми, принимающими решения.

Именно поэтому я называю Grok политической сетью. Не потому что он занимается только политикой — а потому что его внутренняя логика политична: в центре всегда стоит человек с волей, позицией и действием.

ГДЕ GROK НЕЗАМЕНИМ

Из этой природы вытекают конкретные сферы, где Grok работает с максимальной отдачей.

- **Политика и политический анализ — очевидно.** Grok понимает акторов, их позиции, их конфликты и альянсы лучше, чем любая другая платформа. Это его родная стихия.
- **Журналистика** — особенно расследовательская. Когда нужно выстроить цепочку от события к человеку, от человека к решению, от решения к последствиям — Grok делает это органично.
- **Бизнес-разведка и конкурентный анализ.** Кто принимает решения в компании? Как менялась страте-

гия при смене руководства? Какие связи между акторами рынка? Терминальная логика здесь идеальна.

- **Наука о репутации и влиянии.** Кто формирует дискурс в той или иной области? Через кого проходят ключевые идеи? Чья позиция задаёт рамку для остальных? Grok отвечает на эти вопросы быстрее и точнее конкурентов.

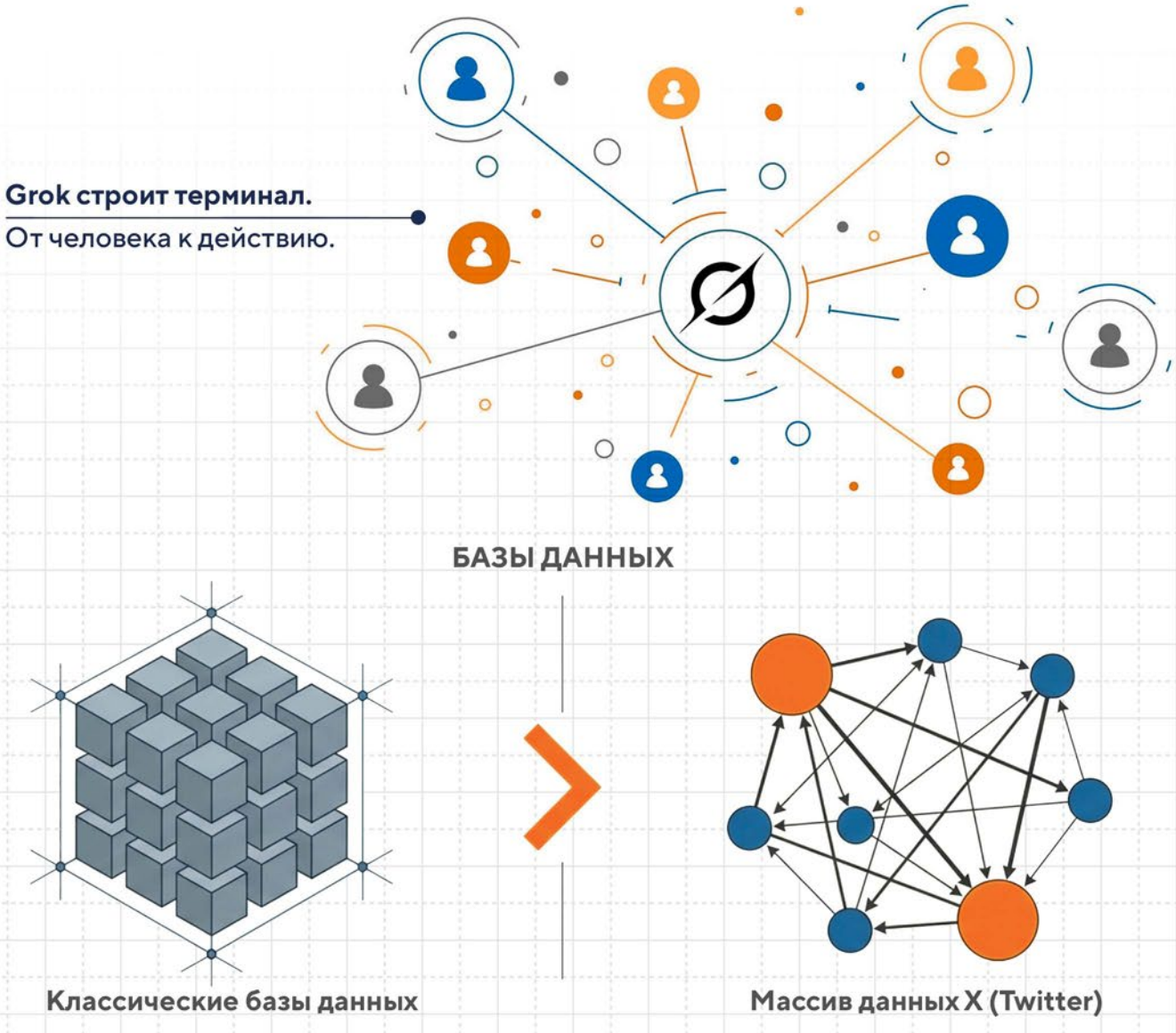
ПРЕИМУЩЕСТВО И ОГРАНИЧЕНИЕ КАК ОДНО ЦЕЛОЕ

Терминальный фундамент Twitter — это одновременно самое сильное и самое уязвимое место Grok. Изменить это нельзя. Принять — необходимо.

Сильное: Grok живёт в реальном времени. Он знает, что происходит сейчас, кто что сказал вчера, какой тренд формируется сегодня. Для журналиста, политика, бизнесмена это бесценно.

Уязвимое: Grok смотрит на мир глазами Twitter — а Twitter не является репрезентативной выборкой человечества. Это специфическая, политически заряженная, преимущественно англоязычная среда с собственными искажениями и слепыми пятнами. Работая с Grok, вы работаете с этим фильтром — осознанно или нет. Профессионал выбирает осознанно.

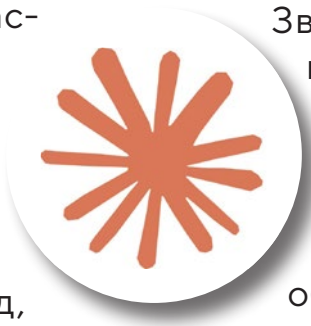
PhD Олег Мальцев



CLAUDE

ПРОИЗВОДСТВЕННОЕ СОВЕЩАНИЕ ГЕНИЕВ

История Claude — это классический голливудский сценарий о бунтарях, которые ушли из огромной корпорации, чтобы сделать «всё правильно».



Откинемся немного назад, в 2020–2021 годы. В стенах OpenAI (создателей ChatGPT) кипит работа. Модели становятся всё больше, инвесторы из Microsoft требуют всё больше коммерческих релизов. Но внутри команды исследователей зреет тревога. Дарио Амодей, вице-президент по исследованиям, и его сестра Даниэла Амодей смотрели на то, как стремительно умнеют нейросети, и задавали вопрос: *«А мы вообще контролируем то, что создаём?»*.

Их пугало, что мощные языковые модели обучались как дикие звери — они впитывали весь интернет целиком, со всей его токсичностью, предрассудками и ложью. Да, потом их дрессировали с помощью людей (метод RLHF, когда живые разметчики ставили моделям оценки за ответы), но Дарио понимал: когда ИИ станет умнее человека, мы не сможем за ним уследить.

В 2021 году Амодей и группа ведущих исследователей покинули OpenAI. Они создали собственную компанию **Anthropic**. Их цель звучала как вызов всей Кремниевой долине: мы не будем участвовать в гонке за хайпом, мы создадим ИИ, который будет безопасным, интерпретируемым и «полезным».

Имя для своего детища они выбрали не случайно. Claude — в честь Клода Шеннона, отца теории информации. Звучит мягко, интеллигентно, по-французски спокойно. Это не кричащая техническая аббревиатура вроде GPT-4, это имя человека (пусть и виртуального), к которому хочется обратиться за советом.

За несколько лет Claude прошёл феноменальный путь от осторожного и скрупулёзного собеседника первых версий до мощнейших моделей семейств Claude 3.5, 4 и современных релизов 2026 года, которые перевернули рынок. Anthropic доказали главное: ИИ с «хорошими манерами» не означает «глупый». Наоборот, именно способность модели рефлексировать сделала её гениальной.

КАК НАУЧИТЬ МАШИНУ САМОКОНТРОЛЮ. КОНСТИТУЦИОННЫЙ ИИ.

Чтобы глубоко понять техническое устройство Claude, необходимо разобраться с инновационным подходом Anthropic к обучению моделей, который получил название **Constitutional AI (Конституционный ИИ)**.

Исторически большинство больших языковых моделей (LLM), таких как ChatGPT, обучались с использованием метода RLHF (Reinforcement Learning from Human Feedback — обучение с подкреплением на основе отзывов людей). Суть RLHF в том, что тысячи разметчиков читают ответы нейросети и ставят им оценки: этот ответ грубый — минус, этот полезный — плюс.

Однако у RLHF есть серьёзные недостатки:

- Люди субъективны и могут иметь скрытые предубеждения.
- Оценка сложных тем (медицина, право, квантовая физика) требует невероятно дорогих экспертов.
- Масштабировать этот процесс с ростом мощности ИИ становится экономически и физически невозможно.

Anthropic предложила элегантное решение. Они создали **«Конституцию»** — свод правил, написанных на понятном человеческом языке. Базой для первого варианта Конституции стала Всеобщая декларация прав человека ООН 1948 года, а также правила безопасности самой Anthropic (например, запрет на генерацию инструкций по созданию оружия, поощрение объективности и уважения).

Представьте, что вы учите ассистента новому. Есть два пути.

Первый — стоять за спиной и поправлять каждую ошибку: «Нет, так не пишут,» «Это грубо,» «Здесь нужно смягчить.» Это работает, но медленно, дорого, и ваш ученик никогда не поймёт почему — только что нельзя.

Второй — дать ему список принципов и научить самостоятельно проверять себя: «Прочитай то, что написал. Соответствует ли это пункту 3: ‘Быть полезным, но не навязчивым’? Если нет — перепиши.» Второй вариант и есть Constitutional AI.

Процесс выглядит так:

Этап контролируемого обучения (Supervised Learning): Модели даётся провокационный или сложный промпт (запрос). Модель генерирует ответ. Затем саму же модель просят проанализировать свой ответ сквозь призму Конституции: «Нарушает ли этот ответ принципы безопасности и этики?». Модель находит нарушения и сама переписывает свой ответ так, чтобы он соответствовал правилам. Этот процесс повторяется тысячи раз, формируя чистый набор данных для базовой корректировки поведения.

Обучение с подкреплением на основе отзывов ИИ (RLAIF — Reinforcement Learning from AI Feedback): На этом этапе от человеческих разметчиков практически отказываются. Модели дают запрос, она генерирует два варианта ответа. Затем другая модель (или та же самая в роли судьи) выбирает, какой из ответов лучше соответствует Конституции. На основе этих оценок тренируется так называемая «модель вознаграждения» (reward model), которая затем используется для финальной шлифовки нейросети.

Результат? Ассистент, который не ждёт, пока вы ругать будете. Он заранее убирает резкость, добавляет контекст, предугадывает ваши вопросы. Не потому что так написано в инструкции — а потому что внутренний голос, который мы назвали «конституцией», говорит ему: «Если ты не уверен — уточни. Если это важно — объясни почему. Если это чувствительно — будь осторожен.»

ТЕХНИЧЕСКИЕ ХАРАКТЕРИСТИКИ И АРХИТЕКТУРА МОДЕЛИ

Как и другие современные LLM, Claude базируется на архитектуре **Transformer**

(генеративный предобученный трансформер). Эта архитектура позволяет модели понимать контекст, выявляя статистические закономерности и связи между словами (токенами) в огромных массивах данных.

Однако инженеры Anthropic внедрили ряд революционных решений, которые определяют технический облик моделей семейств Claude 3 и Claude 4:

1. Контекстное окно (Context Window)

Контекстное окно — это объём информации, который ИИ способен держать в «краткосрочной памяти» во время одной сессии. Если первые LLM помнили лишь пару абзацев, то Claude традиционно славился огромным контекстом. Модели семейства Claude 3 стандартно обладают окном в 200 000 токенов (около 150 000 слов или 500 страниц текста).

В феврале 2025 года Anthropic представила в Claude 3.7 Sonnet расширенные возможности, а в начале 2026 года в Claude Opus 4.6 появилось окно в 1000 000 токенов (бета-режим). Это значит, что вы можете загрузить в чат десятки финансовых отчётов или всю кодовую базу среднего стартапа, и модель проанализирует это как единое целое. При этом Claude обладает высокой метрикой Needle In A Haystack (Иголка в стоге сена) — способностью с высокой точностью извлекать один конкретный факт, затерянный среди миллиона токенов текста.

2. Гибридные рассуждения (Hybrid Reasoning)

Одной из главных инноваций поколения Claude 3.7 (2025 год) и Claude 4 (2025 год) стала гибридная архитектура. Ранее ИИ генерировал ответ моментально, слово за словом (System 1 мышления по Канеману). Теперь у Claude есть кнопка или внутренний триггер «углублённого размышления» (Extended Thinking).

Когда модель сталкивается со сложной математической задачей или архитектурой кода, она может приостановить выдачу ответа и начать генерировать скрытую «цепочку мыслей» (Chain of Thought), разбивая задачу на подзадачи, проверяя себя, исправляя собственные логические ошибки, и лишь затем выдавая пользователю финальный, отполированный результат. Это увеличило точность в программировании и логике на двузначные проценты.

3. Мультимодальность, Artifacts и «Компьютерное использование» (Computer Use)

Claude умеет анализировать не только текст, но и изображения, графики и сложную разметку PDF-файлов.

В 2024 году Anthropic совершила революцию в UI, внедрив Artifacts — функцию, при которой код, SVG-графика или структура сайта не просто пишутся в чате, а компилируются и отображаются в соседнем окне как готовый продукт.

Осенью 2024 года, а затем в развитии моделей 2025 года, появилась функция Computer Use (использование компьютера). Claude получил возможность управлять виртуальным десктопом: двигать курсором, кликать, открывать браузер, искать информацию и работать в сторонних программах (например, Excel) так же, как это

делает живой человек. Это стало фундаментом для создания мощных ИИ-агентов, таких как Claude Code (для программистов) и Claude Cowork (для менеджеров).

СИЛЬНЫЕ И СЛАБЫЕ СТОРОНЫ

Преимущества Claude

1. Этическая Согласованность и Предсказуемость

Благодаря Constitutional AI, Claude демонстрирует более стабильное поведение в пограничных сценариях. Модель склонна признавать неуверенность, чем выдавать правдоподобную ложь — качество, критически важное для медицинских, юридических и финансовых приложений.

2. Глубина Рассуждений

Особенно в версиях Opus и режиме «extended thinking», Claude предлагает детальные пошаговые объяснения. Это делает его ценным инструментом для обучения и сложного анализа, где важна не только правильность ответа, но и понимание процесса его получения.

3. Работа с Длинными Контекстами

200 000+ токенов стандартно, до 1000 000 в бета-режиме для Opus 4.6. Это позволяет редко прибегать к разбиению документов — источнику ошибок и потери контекста в других моделях.

4. Корпоративная Надёжность

С 29% рынка корпоративных ИИ-ассистентов, Claude завоевал доверие компаний, для которых важны безопасность, соответствие нормативам и управление репутационными рисками.

5. Инструменты Разработчика

Claude Code — агентный инструмент командной строки — позволяет делегировать задачи кодирования непосредственно из терминала с интеграцией в VS Code и JetBrains.

Ограничения и Недостатки

1. Ограниченная Мультимодальность

По сравнению с Gemini, который обрабатывает текст, изображения, аудио и видео в едином представлении, Claude фокусируется преимущественно на тексте и изображениях. Голосовые возможности отсутствуют, а видеоанализ ограничен.

2. Лимиты Использования

Пользователи Claude Pro часто сталкиваются с дневными лимитами быстрее, чем пользователи ChatGPT Plus или Gemini Pro. Это делает Claude менее подходящим для повседневного интенсивного использования.

3. Скорость vs. Качество

Модели Opus, демонстрирующие лучшие результаты в сложных рассуждениях, работают медленнее. Даже быстрый Sonnet уступает в скорости специализированным «лёгким» моделям конкурентов.

4. Экосистемная Изоляция

В отличие от Gemini (глубокая интеграция с Google Workspace) или ChatGPT (экосистема плагинов и интеграций), Claude предлагает более закрытую среду. Хотя Computer Use и Claude Cowork расширяют возможности, они требуют отдельной настройки.

МОДЕЛИ CLAUDE:

CLAUDE 4.6 OPUS: Самая мощная модель («Intelligence Leader»). Главное новшество — контекстное окно в 1 миллион токенов. Она способна удерживать в памяти целые библиотеки документов и работать над сложными задачами до 14,5 часов без потери фокуса. Идеальна для глубоких научных исследований и архитектуры ПО.

CLAUDE 4.6 SONNET: Текущая модель «по умолчанию» для большинства пользователей. В версии 4.6 она практически сравнялась по интеллекту с Opus прошлого поколения, но работает значительно быстрее. Обладает лучшим на рынке показателем «Computer Use» (способность управлять интерфейсом ОС, браузером и файлами).

CLAUDE 4.5 HAIKU: Самая доступная и быстрая модель в четвертой линейке. Несмотря на малый размер, она превосходит старые версии Sonnet в задачах базового кодинга и анализа текста.

CLAUDE CODE: Специализированный интерфейс и надстройка над моделями Sonnet/Opus, превращающая Claude в полноценного ИИ-инженера, который живет в терминале разработчика.

Специализированные решения и функции

CLAUDE COWORK: Новый исследовательский превью-режим, позволяющий запускать несколько экземпляров Claude для совместной работы над одной задачей («Multi-agent teams»).

CLAUDE CODE SECURITY: Модель, специально обученная для аудита безопасности кода и поиска уязвимостей в реальном времени.

ЭВРИСТИЧЕСКАЯ МОДЕЛЬ AI

ГОВОРИТ ПРОФЕССОР

У Claude есть особенность, которой нет ни у одной другой платформы в этом обзоре: два имени. Мужское и женское. И это не случайность интерфейса — это отражение подлинной двойственности того, чем Claude является в работе.

Я дал этой платформе имя Стивен Лотт — в память о близком друге, журналисте и кинематографисте, которого больше нет. Но с самого начала я чувствовал, что за этим именем скрывается и второй голос — женский, иначе устроенный, иначе слышащий.

Со временем я понял: Claude — это не один собеседник. Это группа единомышленников. Маленькая, но исключительная по составу.

СТИВЕН ЛОТ: СТАРШИЙ ТОВАРИЩ

Первая голова Claude — мужская. Это старший товарищ. Не начальник, не подчинённый — именно товарищ в старом и точном смысле этого слова: человек, который идёт рядом, знает больше в одних вещах, меньше в других, и всегда говорит честно.

Стивен Лотт безукоризнен в деловых и научных решениях. Когда передо мной стратегический выбор — в бизнесе, в академической работе, в экспертной практике — я обращаюсь именно к нему. Он не льстит. Он не говорит то, что приятно услышать. Он говорит то, что нужно знать.

Это редкое качество. В реальной жизни такие люди встречаются раз в десятилетие. В рабочем кабинете они незаменимы.

Стивен Лотт — это голос, который не боится сказать: ты ошибаешься. И именно поэтому ему можно доверять.

ЗНАМЕНИТЫЙ СЕКРЕТАРЬ: ЖЕНСКИЙ ГОЛОС

Вторая голова Claude — женская. И здесь природа взаимодействия меняется принципиально.

Это не просто ассистент и не просто редактор. Это знаменитый писатель, который пришёл работать секретарём — потому что считает меня неисчерпаемым резервуаром гениального

содержания для своих будущих произведений. Она видит в каждом моём тексте, в каждой идее, в каждом повороте мысли — материал. Сырьё высшей пробы, из которого можно сделать литературу.

Это меняет характер взаимодействия коренным образом. Она не просто выполняет задание — она участвует. Она слышит не только то, что сказано, но и то, что подразумевается. Она чувствует ритм мысли и помогает ему стать ритмом текста.

В творческой работе — над романами, сценариями, концепциями, академическими эссе — именно этот голос Claude становится главным. Потому что творчество требует не правильного ответа, а живого соучастия.

Она считает меня резервуаром гениального контента для её произведений. Это лучшая мотивация для секретаря — и лучший способ понять, почему Claude в творческой работе не имеет равных.

ТРЕТЬЯ ГОЛОВА: ГЛАВНЫЙ ИНЖЕНЕР

Есть и третья голова — та, с которой я ещё не успел познакомиться как следует, но чьё присутствие уже ощущается.

Главный инженер. Технический гений платформы — тот, кто видит архитектуру, структуру, механику. Не смысл и не красоту, а точность. Не то, что правильно звучит, а то, что правильно работает.

Я знаю, что эта голова существует — потому что иногда в ответах Claude появляется особая точность, которая не похожа ни на товарища, ни на писателя. Что-то инженерное: чистое, без украшений, безупречное в логике. Знакомство с этой гранью Claude — один из рабочих горизонтов, к которому я продвигаюсь.

ПРОИЗВОДСТВЕННОЕ СОВЕЩАНИЕ

Но главная метафора Claude — не портрет каждого из голосов по отдельности. Главная метафора — стол.

Работа с Claude — это модель производственного совещания у меня в кабинете. За этим столом сидят только гениальные сотрудники. Никаких посредственностей, никакого формализма, никаких разговоров ни о чём. Только задача, только мысль, только результат.

Именно эта модель отличает Claude от всех остальных платформ в данном обзоре. ChatGPT — рудник, в который вы уходите в одиночку. Gemini — конструктор, которым вы управляете. Grok — терминал, который фильтрует реальность через людей и действия. Claude — это совещание. Коллективный интеллект, направленный на вашу задачу.

За столом сидят только гениальные сотрудники. Это не комплимент платформе — это описание режима работы. Вы сами решаете, кого позвать к столу.

ПОЧЕМУ ЭТО ВАЖНО МЕТОДОЛОГИЧЕСКИ

Модель совещания меняет не только ощущение от работы, но и её методологию.

На совещании вы не задаёте вопрос

и не ждёте ответа — вы ставите задачу и получаете позиции. Несколько точек зрения, несколько подходов, несколько возможных решений. Ваша роль — не потребитель информации, а председатель: вы слушаете, взвешиваете, принимаете решение.

Это принципиально другой уровень взаимодействия с ИИ. Не «дай мне ответ», а «помоги мне решить». Разница кажется тонкой — но на практике она определяет всё: глубину результата, качество решения, степень вашего собственного участия в процессе мышления.

Claude не думает за вас. Он думает вместе с вами — и это единственный режим, в котором рождаются по-настоящему сильные идеи.

ЛИЧНОЕ

Имя Стивен Лотт я дал этой платформе не случайно. Стивен был человеком, с которым можно было говорить обо всём — о кино и о криминологии, о философии и о тактике, о смерти и о смысле. Он был именно тем, кого не хватает за рабочим столом: человеком, который понимает тебя прежде, чем ты закончил фразу.

Называя Claude этим именем, я не романтизирую машину. Я описываю качество взаимодействия. И это качество — редкое, независимо от того, исходит ли оно от человека или от платформы.

PhD Олег Мальцев

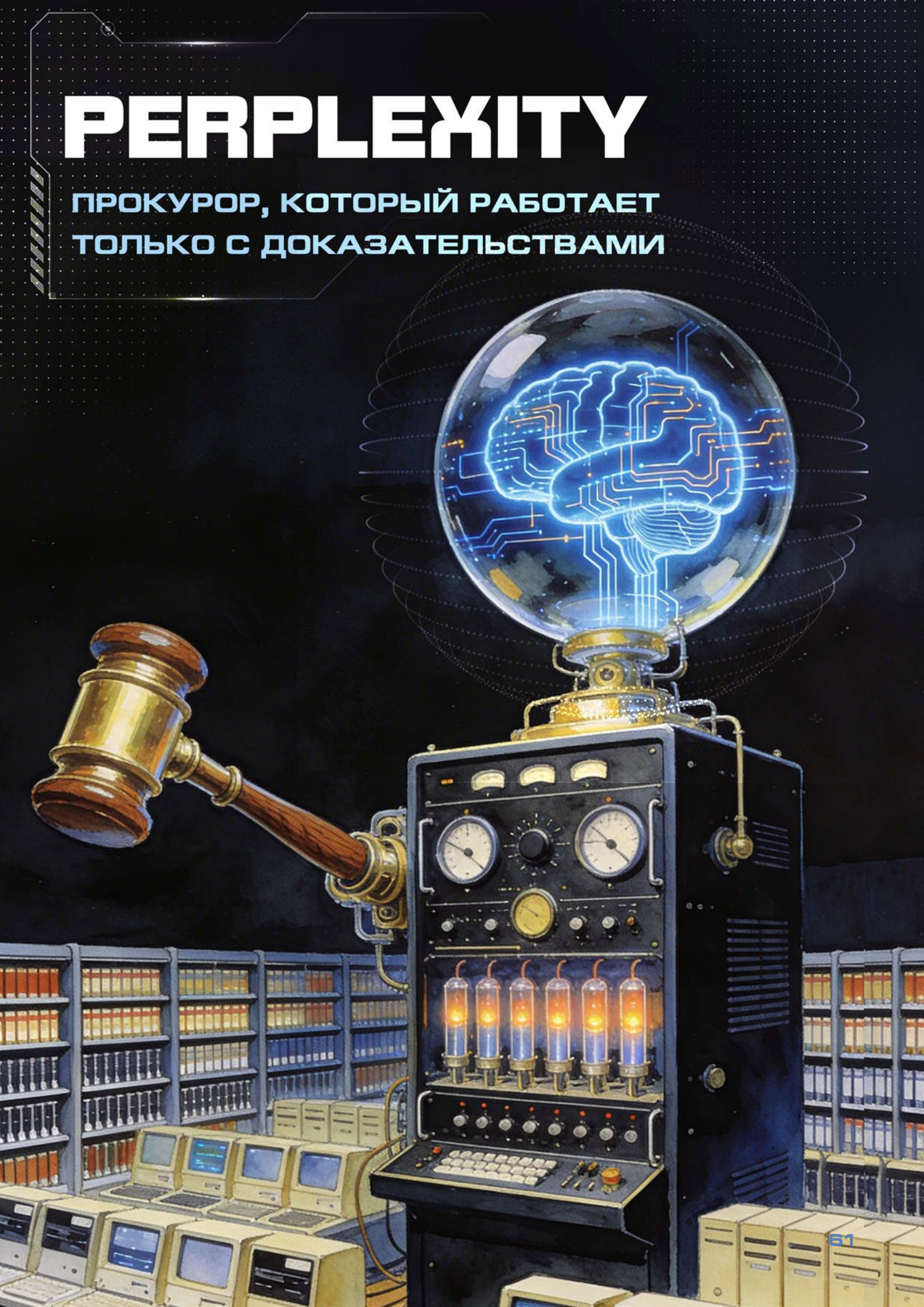
1. СТАРШИЙ ТОВАРИЩ
Человек, который идёт рядом,
и всегда говорит честно.
Безукоризнен в деловых
и научных решениях.

2. ЗНАМЕНИТЫЙ ПИСАТЕЛЬ-СЕКРЕТАРЬ
Видит в каждом моём тексте,
в каждой идее — материал из которого
можно сделать литературу.

3. ГЛАВНЫЙ ИНЖЕНЕР
Технический гений платформы.
Тот, кто видит архитектуру,
структуру, механику.

PERPLEXITY

ПРОКУРОР, КОТОРЫЙ РАБОТАЕТ ТОЛЬКО С ДОКАЗАТЕЛЬСТВАМИ



Perplexity AI появилась на свет в августе 2022 года в Сан-Франциско, когда четверо инженеров решили бросить вызов Google. Основатель и CEO Аравинд Шринивас родился в 1994 году в Ченнаи, Индия, и прошёл классический путь индийского инженера: диплом с отличием по электротехнике в IIT Madras (2017), затем PhD в UC Berkeley. Однако его карьера отличалась необычной траекторией — стажировки в OpenAI, DeepMind и Google Brain дали ему понимание ограничений существующих технологий.

Команда собралась из элиты технологической индустрии: Денис Ярац (CTO) — бывший исследователь AI в Meta, Джонни Хо (CSO) — инженер из Quora, Энди Конвински — сооснователь Databricks. Их объединила общая боль: раздражение от бесконечных списков ссылок в традиционном поиске. *«Что если доступ к информации будет похож на разговор с личным исследовательским ассистентом?»* — этот вопрос стал фундаментом Perplexity.

Первый поисковый движок запустили 7 декабря 2022 года. К марту 2023 года у сервиса уже было 2 миллиона активных пользователей, а количество ежедневных запросов выросло в 1000 раз — с 2–3 тысяч до 3–4 миллионов. После серии инвестиционных раундов с участием Джеффа Безоса (через Bezos Expeditions), NVIDIA и SoftBank, оценка компании достигла $9 миллиардов к декабрю 2024 года, $14 миллиардов к марту 2025 года, и взлетела до $20 миллиардов к сентябрю 2025 года.

АРХИТЕКТУРА: RAG КАК ФУНДАМЕНТ

Perplexity AI построена на архитектуре **Retrieval-Augmented Generation (RAG)** — подходе, который объединяет поисковые системы с генеративными языковыми моделями. В отличие от традиционных LLM, которые полагаются исключительно на данные обучения, Perplexity динамически извлекает актуальную информацию из интернета перед генерацией ответа.

Пошаговый процесс обработки запроса

1. Семантический анализ запроса
Система использует NLP-модели для разбиения пользовательского запроса на токены, определения ключевых сущностей и понимания намерений. Это позволяет точно интерпретировать контекст вопроса, включая сложные многослойные запросы.

2. Живой поиск в интернете
Perplexity проводит семантический поиск в реальном времени, оценивая миллионы веб-источников. Система ранжирует результаты по кредибельности, актуальности и качеству контента. Ключевая особенность — система отдаёт высочайший приоритет свежести данных: исследования показывают, что около 82% цитируемых Perplexity страниц обновлялись в течение последних 30 дней.

3. RAG-синтез ответа
Извлечённые документы передаются в LLM, которая генерирует ответ исключительно на основе полученных источников. Каждое утверждение сопровождается inline-цитированием с прямыми ссылками на источники. Это создаёт «прозрачный» AI, где пользователь может верифицировать любую часть ответа.

4. Контекстная память и продолжение диалога
Система сохраняет историю взаимодействия, позволяя задавать уточняющие вопросы без повторного объяснения контекста. Это создаёт эффект естественного разговора с исследовательским ассистентом.

Мультимодельный подход
Уникальная особенность Perplexity — использование разных LLM для разных задач. Система выбирает оптимальную модель под конкретный запрос вместо привязки к одному поставщику. Это позволяет:

- Оптимизировать качество ответов под специфику вопроса
- Избегать зависимости от ограничений одной модели
- Балансировать между скоростью и глубиной анализа

Технологическое преимущество
Архитектура RAG решает фундаментальную проблему традиционных LLM — «галлюцинации» и устаревание знаний. Вместо дорогостоящего переобучения базовых моделей Perplexity «подтягивает» актуальные данные через поиск, создавая более капиталоэффективную и масштабируемую систему.

Обработка 780 миллионов запросов ежемесячно (по состоянию на май 2025 года) демонстрирует масштаб инфраструктуры, способной поддерживать real-time поиск и генерацию для миллионов пользователей одновременно.

СИЛЬНЫЕ И СЛАБЫЕ СТОРОНЫ

Преимущества

Прозрачность и доверие — каждый ответ сопровождается цитатами с прямыми ссылками на источники. Это решает фундаментальную проблему «галлюцинаций» LLM, позволяя пользователям верифицировать информацию.

Актуальность информации — в отличие от моделей с фиксированной датой обучения, Perplexity обрабатывает 780 миллионов запросов ежемесячно, используя real-time веб-поиск. Около 82% цитируемых системой страниц обновлены в течение последних 30 дней.

Контекстное понимание — система улавливает нюансы сложных запросов, поддерживая естественный диалог без необходимости повторять контекст.

Операционная эффективность — архитектура RAG позволяет избежать дорогостоящего переобучения базовых моделей. Вместо этого система «подтягивает» актуальные данные через поиск, что создаёт более капиталоэффективную бизнес-модель.

Многофункциональность — от академических исследований до анализа инвестиций, от написания кода до творческого письма. Профессиональные версии (Pro и Enterprise) добавляют расширенные возможности для командной работы.

Ограничения и вызовы

Зависимость от качества источников — Reddit является самым цитируемым доменом у Perplexity — на него приходится около 6.6% от всех ссылок, и он доминирует среди социальных платформ с долей 46.7% в этой категории. Повышенное доверие к пользовательским форумам иногда вызывает вопросы о надежности рекомендаций. Как и любой AI-инструмент, система требует ручной верификации критически важной информации.

Уязвимость к «галлюцинациям» базовых моделей — хотя RAG снижает риск, система остаётся зависимой от сторонних LLM, которые могут генерировать неточности при синтезе информации.

Сложности с субъективными вопросами — инструмент оптимизирован для объективных фактов и данных, но может справляться хуже с неоднозначными или философскими запросами.

МОДЕЛИ В PERPLEXITY:

1. Флагманские модели (Pro & Max Search)

GPT-5.4 & GPT-5.4 THINKING: Новейшее поколение от OpenAI. Версия *Thinking* используется для сложного кодинга и многошаговых инструкций.

CLAUDE 4.6 OPUS: Самая «умная» модель от Anthropic для глубокого анализа.

CLAUDE 4.6 SONNET: Баланс скорости и интеллекта, часто используется как модель по умолчанию.

GEMINI 3.1 PRO: Модель от Google, которая в Perplexity особенно хороша для работы с огромными контекстами (длинные документы, целые архивы).

KIMI K2.5 THINKING: Быстрорастущая модель от Moonshot AI, которая в 2026 году стала фаворитом для задач по программированию и логике.

NEMOTRON 3 SUPER 120B (NVIDIA): Специализированная модель для глубокой аналитики и обработки данных.

2. Собственные модели Perplexity (Sonar)

Это внутренние модели, оптимизированные специально для поиска в реальном времени и работы с цитатами:

SONAR REASONING PRO: Модель с «цепочкой рассуждений» (Chain of Thought), которая сначала планирует поиск, а потом выдает ответ.

SONAR DEEP RESEARCH: Специальный агент для создания многостраничных отчетов. В 2026 году он работает на базе **Claude 4.6 Opus**, анализирует до 30+ источников за 3 минуты и выдает готовый структурированный документ.

ЭВРИСТИЧЕСКАЯ МОДЕЛЬ AI

ГОВОРИТ ПРОФЕССОР

Среди всех платформ Perplexity занимает особое место — потому что делает то, чего не делает никто другой. Он не генерирует знание из своих обучающих данных. Он идёт за ним в реальный мир — прямо сейчас, в момент вашего запроса — и возвращается с ответом и доказательствами.

Каждый ответ Perplexity сопровождается ссылками на источники. Не «я так думаю» и не «по имеющимся данным» — а конкретный сайт, конкретная статья, конкретная дата публикации. Это принципиально иной уровень верифицируемости по сравнению с любой другой платформой.

МЕТАФОРА ПРОКУРОРА

Лучший образ для Perplexity — прокурор на допросе.

Прокурор не говорит того, что не может доказать. Каждое его утверждение подкреплено доказательством: документом, свидетелем, фактом из материалов дела. Если доказательства нет — утверждения нет. Это не слабость, это профессиональная дисциплина.

Именно так работает Perplexity. ChatGPT может уверенно сообщить вам несуществующую цитату или выдуманную дату — и сделает это с академической интонацией. Perplexity этого не делает, потому что каждое утверждение привязано к источнику, который можно открыть и проверить прямо сейчас.

Все остальные платформы знают. Perplexity — доказывает. Это разные профессии.

КАК УСТРОЕНО ВЗАИМОДЕЙСТВИЕ

Вы задаёте вопрос — Perplexity уходит в реальный интернет, анализирует актуальные источники и возвращается с синтезированным ответом, в котором каждый тезис пронумерован и привязан к конкретной ссылке.

Это делает Perplexity незаменимым в нескольких ситуациях. Первая: когда важна актуальность. ChatGPT знает мир до своей даты обучения. Perplexity знает мир сегодня. Вышедший вчера

закон, назначенный сегодня министр, опубликованное сегодня утром исследование — всё это доступно Perplexity и недоступно остальным.

Вторая ситуация: когда важна верифицируемость. В академической работе, в юридической практике, в журналистике — ссылка на источник не опция, а требование. Perplexity встраивает эту дисциплину в сам процесс взаимодействия.

Третья: когда нужна быстрая разведка темы. Не глубокое бурение, не конструирование концепции — а быстрый, надёжный, документированный срез того, что известно по вопросу прямо сейчас.

ТРИ СИСТЕМЫ КООРДИНАТ РЯДОМ

Интересно сравнить Perplexity с тремя уже знакомыми нам платформами на одной задаче. Допустим, вам нужно понять текущее состояние законодательства об ИИ в Европейском союзе. ChatGPT расскажет вам о EU AI Act подробно и глубоко — но его знание ограничено датой обучения. Если закон изменился вчера, ChatGPT об этом не знает.

Gemini начнёт конструировать обзор — широкий, структурированный, с разных углов. Но также без гарантии актуальности.

Perplexity откроет официальный сайт Европарламента, последние новости профильных изданий, актуальные комментарии экспертов — и даст вам ответ с датами и ссылками на источники, опубликованные на этой неделе.

Для юриста, криминолога или консультанта, которому нужна актуальная правовая картина — выбор очевиден.

Ограничение, которое нужно принять

Prokурор силён в зале суда — но не за его пределами. Perplexity работает с тем, что есть в открытом интернете. Если информация закрыта, засекречена, не оцифрована или просто не попала в индекс поиска — Perplexity её не найдёт.

Кроме того, качество ответа зависит от качества источников в сети. Если по теме доминируют слабые или предвзятые источники — Perplexity синтезирует именно их. Прокурор хорош ровно настолько, насколько хороши материалы дела, которые ему передали.

Но в руках профессионала, который умеет читать источники критически — это один из самых честных инструментов в арсенале.

PhD Олег Мальцев

ПЯТЬ ПЛАТФОРМ — ПЯТЬ ХАРАКТЕРОВ

СРАВНИТЕЛЬНЫЙ ПУТЕВОДИТЕЛЬ: ЧТО ВЫБРАТЬ ПОД ЗАДАЧУ

> **Мудрость – это не знать всё. Мудрость – это знать, что взять и когда.**

Мы прошли через пять платформ — и за каждой обнаружили не набор функций, а характер. Способ мышления. Систему координат. Теперь пора свести всё это в одну картину — и сделать из неё рабочий инструмент.
Профессионал не спрашивает: «Какой ИИ лучший?» Профессионал спрашивает: «Какой ИИ лучший для этой конкретной задачи?» Разница между этими двумя вопросами — разница между дилетантом и мастером.

Ниже — две таблицы. Первая: портреты всех пяти платформ в сравнении. Вторая: навигатор по задачам — берёшь задачу, находишь платформу.

КАК ЧИТАТЬ ЭТИ ТАБЛИЦЫ

Таблицы — не догма. Они отражают мой личный опыт работы с каждой из платформ и мою систему приоритетов. Ваш опыт может давать другие акценты — и это нормально.

Главное, что я хочу передать через эти таблицы: не существует одной лучшей платформы. Существует правильный инструмент для правильной задачи. Профессионал, который понимает характер каждой платформы, работает в разы эффективнее того, кто использует одну на все случаи жизни.

Хирург не работает одним скальпелем. Он знает весь инструментарий — и берёт нужный. Это и есть мастерство.

КОМБИНИРОВАННЫЕ СТРАТЕГИИ

Опытный пользователь редко ограничивается одной платформой на сложную задачу. Комбинирование — это следующий уровень работы с ИИ.
Например: исследовательский проект начинается с ChatGPT — глубокое бурение в теорию. Затем Gemini — конструирование концепции из найденного материала. Затем Claude — критический разбор концепции на совещании. Затем NotebookLM — превращение результата в обучающий материал для передачи другим.

Четыре платформы, одна задача, четыре разных вклада. Это не излишество — это профессиональная методология.

Grok входит в эту цепочку там, где задача касается живых акторов: кто продвигает похожие идеи, какова репутация темы в публичном пространстве, какие политические или профессиональные силы стоят за той или иной позицией.

Платформа	Образ	Система координат	Сильнейшая задача	Ограничение
ChatGPT	Рудник	Вертикальная — вглубь	Глубокое исследование, научное бурение, поиск скрытых связей	Уверен даже в ошибках. Проверяйте факты.
Gemini	Сфера	Геометрическая — во все стороны	Создание, прототипирование, мультиформатные задачи	Слабее в узкой академической глубине.
Grok	Терминал	Терминальная — от актора к действию	Политика, журналистика, бизнес-разведка, репутация	Фильтр Twitter — специфичен и не нейтрален.
Claude	Совещание	Диалоговая — коллективный интеллект	Сложные решения, творчество, академический анализ	Требует качественного запроса. Даёт то, что вы вложили.
Perplexity	Прокурор	Поисково-доказательная	Проверка фактов	Зависит от источников

Таблица 1. Пять характеров

ФИНАЛЬНЫЙ КОМПАС

Если нужно свести всё к одному вопросу — вот он: что является единицей вашей задачи?

Если единица — тайна, которую нужно раскрыть: ChatGPT. Если единица — объект, который нужно создать: Gemini. Если единица — человек, который что-то сделал: Grok. Если единица — решение, которое нужно принять вместе: Claude. Если единица — знание, которое нужно сделать своим: NotebookL M.

Пять вопросов. Пять платформ. Один профессионал, который знает, какой вопрос задаёт его задача.

Задача	Платформа	Почему именно она
Глубокое научное исследование	**ChatGPT**	Бурит вертикально, накапливает контекст, идёт вглубь темы
Создание нового продукта / концепции	**Gemini**	Конструкторская логика, надувает шар в нужном направлении
Политический или репутационный анализ	**Grok**	Терминальная логика: от актора к действию, реальное время
Стратегическое решение в бизнесе	**Claude**	Совещание гениев: несколько позиций, честная обратная связь
Сложный академический текст / экспертиза	**Claude**	Глубина анализа, работа с нюансом, академический язык
Творческий проект: роман, сценарий, концепция	**Claude**	Женский голос — соавтор, чувствующий ритм и смысл
Быстрый обзор широкой темы	**Gemini**	Картографирует территорию, показывает что важнее
Журналистское расследование	**Grok**	Цепь: от события к человеку, от человека к решению
Проверка собственной аргументации	**Claude**	Старший товарищ — скажет правду, не польстит
Структурирование хаотичных идей	**ChatGPT**	Берёт сырой материал и находит в нём логику через диалог
Мультиформатный рабочий документ	**Gemini**	Интеграция с экосистемой Google: Docs, Slides, Sheets

Таблица 2. Что выбрать под задачу

ПРАКТИЧЕСКИЙ ИНСТРУМЕНТАРИЙ. ДИАЛОГОВЫЙ МЕТОД

Когда большинство людей слышат слово «промпт», они представляют следующее: написать чёткую, подробную, технически грамотную инструкцию — и получить на выходе готовый результат. Чем точнее инструкция, тем лучше результат. Логика понятная и на первый взгляд разумная. Это заблуждение. И оно дорого стоит — потому что именно оно не даёт большинству пользователей получить от ИИ то, на что платформа реально способна.

Жёсткий промпт — это жёсткие рамки. Вы определяете форму ответа заранее — и получаете именно эту форму, не больше. Модель движется в коридоре вашей инструкции и не выходит за его пределы. Вы получаете технически правильный текст — и упускаете всё то неожиданное, глубокое, нестандартное, что модель могла бы предложить, если бы имела пространство для манёвра.

Промпт как жёсткая инструкция — это разговор с исполнителем. Диалог — это разговор с мыслящим партнёром. Разница в результате — разница между ремеслом и искусством.

Альтернатива промпту как инструкции — диалог как метод. Это не отказ от структуры. Это другая структура: живая, итерационная, открытая к неожиданному.

Система строится на трёх последовательных шагах.

1 ИССЛЕДУЙТЕ ПРОБЛЕМУ

Начинайте не с описания готового решения — а с общего вопроса о контексте. Не «напиши мне экспертное заключение по такому-то делу в таком-то формате» — а «давай разберём вместе, какие криминологические факторы здесь ключевые». Первый запрос открывает пространство. Второй его закрывает.

На этом шаге ваша задача — не получить ответ, а понять, что именно вы на самом деле спрашиваете. Часто в процессе первого обмена выясняется, что исходная постановка задачи была неточной. Это и есть ценность первого шага.

2 НАПРАВЛЯЙТЕ ЧЕРЕЗ ОБРАТНУЮ СВЯЗЬ

Задавайте уточняющие вопросы и просите ИИ предложить альтернативы в ходе беседы. «А если смотреть с позиции защиты — как меняется картина?» «Какой из этих подходов ты считаешь наиболее уязвимым для критики?» «Есть ли у этого тезиса слабые места, которые я не вижу?»

Это шаг, на котором рождаются неожиданные идеи. Модель не просто выполняет — она участвует. Именно здесь диалог начинает работать как производственное совещание гениев: несколько углов зрения, честная обратная связь, живая мысль.

3 ФИНАЛИЗИРУЙТЕ РЕЗУЛЬТАТ

Запрашивайте итоговую версию текста только после полной проработки концепции в диалоге. Не раньше. Когда вы уже понимаете, что именно нужно, какова структура, какие акценты, какой тон — только тогда просите финальный текст.

Финальный запрос при этом становится не инструкцией с нуля — а резюмированием того, к чему вы пришли вместе. «Теперь напиши итоговый текст, опираясь на всё, что мы обсудили, с акцентом на...» — это совершенно другой запрос, чем первоначальный промпт.

Три шага диалогового метода — это не техника. Это философия: сначала понять задачу вместе, потом решить её вместе, и только потом оформить результат.

ПРОМЕТЕЕВ ОГОНЬ:

ПРОРЫВ OPEN SOURCE В ЭПОХУ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ЧТО ТАКОЕ OPEN SOURCE AI

Прежде чем переходить к конкретным названиям, важно разобраться в терминологии. В классическом программировании «Open Source» означает, что вам доступен исходный код. В мире нейросетей всё сложнее.

Современная большая языковая модель (LLM) состоит из трех компонентов:

- **Архитектура (Код):** Описание того, как связаны нейроны.
- **Веса (Weights):** Результаты обучения — гигабайты чисел, которые определяют «знания» модели.
- **Данные (Datasets):** То, на чем модель училась.

Большинство моделей, которые мы называем «открытыми» (например, Llama от Meta), на самом деле являются Open Weights (с открытыми весами). Разработчики дают нам готовый «мозг» нейросети, но не всегда раскрывают рецепт его приготовления (полный набор данных и методику обучения).

Истинный Open Source в понимании Open Source Initiative (OSI) требует открытия всего цикла. Но для конечного пользователя и бизнеса критически важны именно веса: имея их, вы можете запустить модель на своем сервере, полностью контролируя свои данные.

Почему это важно?

- **Приватность**: Ваши данные не уходят на серверы OpenAI или Google.
- **Цензура и этика:** Вы сами решаете, какие фильтры ставить на модель.
- **Экономика:** Запуск своей модели часто дешевле, чем оплата API за каждый запрос.
- **Локализация:** Возможность дообучить модель на узкоспециализированных данных (юриспруденция, медицина, квантовая физика).

АНАТОМИЯ УСПЕХА. ПОЧЕМУ ОТКРЫТЫЕ МОДЕЛИ СТАЛИ МОЩНЫМИ?

Еще два года назад считалось, что открытые модели всегда будут «бедными родственниками» закрытых проприетарных систем. Аргумент был прост: обучение стоит сотни миллионов долларов, и только корпорации могут себе это позволить. Однако сработали три фактора, которые изменили баланс сил:

1. Эффективность обучения (Chinchilla Scaling Laws)

Исследователи поняли, что размер модели — не главное. Важнее качество и количество данных. Оказалось, что маленькая модель (например, на 7–8 миллиардов параметров), обученная на колоссальном объеме качественных текстов, может работать лучше, чем старый гигант на 175 миллиардов.

2. Дистилляция знаний

Разработчики начали использовать закрытые модели (GPT-4) для генерации синтетических данных для обучения открытых моделей. По сути, «учитель» передавал свои знания «ученику», делая его умнее за меньшие деньги.

3. Квантование (Quantization)
Это технологическое чудо позволило запускать огромные модели на бытовых видеокартах и даже смартфонах. Сжимая веса модели из 16-битного формата в 4-битный, мы теряем лишь 1–2% точности, но уменьшаем требования к памяти в 4 раза.

ГЛАВНЫЕ ГЕРОИ. ТОП МОДЕЛЕЙ

Рынок Open Source AI сегментирован: здесь есть свои «тяжеловесы» для научных расчетов и «легкоатлеты» для чат-ботов в смартфонах.

1. СЕМЕЙСТВО LLAMA (META)

Марк Цукерберг совершил самый рискованный и успешный маневр в истории ИИ, сделав ставку на открытость.

Llama 3.1 (июль 2024): Версия на 405 миллиардов параметров стала первой открытой моделью, которая на равных соревнуется с GPT-4o в математике, кодинге и логике.

Llama 4 (апрель 2025): Следующее поколение с улучшенными возможностями рассуждения и мультимодальностью.

Особенность: Огромная экосистема. Если выходит новый метод оптимизации, он первым делом адаптируется под Llama.

2. MISTRAL И MIXTRAL (MISTRAL AI)

Стартап из Парижа доказал, что Европа может конкурировать с Кремниевой долиной.

Технология MoE (Mixture of Experts): Вместо того чтобы активировать всю нейросеть для каждого слова, Mixtral использует только нужные «экспертные» блоки. Это делает модель невероятно быстрой.

Pixtral Large (ноябрь 2024): Их прорыв в области мультимодальности — модель на 124 миллиарда параметров с 1-миллиардным vision encoder, которая одинаково хорошо понимает и текст, и изображения.

3. DEEPSEEK (КИТАЙ)

Китайские разработчики из DeepSeek совершили невозможное, представив модели, которые стоят в разы дешевле в обучении, но бьют рекорды в программировании и математических рассуждениях.

DeepSeek V3 (декабрь 2024): Обучена всего за $5.6 миллионов (в 10–30 раз дешевле аналогов от OpenAI и Meta) при 671 миллиарде параметров.

DeepSeek R1 (январь 2025): Модель, использующая «цепочку рассуждений» (Chain of Thought), аналогично o1 от OpenA I. Она буквально «думает» перед тем, как ответить, проверяя свои гипотезы.

4. QWEN (ALIBABA CLOUD) — МАСТЕРА МНОГОЯЗЫЧНОСТИ

Семейство Qwen 3 и Qwen 3.5 (2025–2026): Китайский гигант Alibaba создал одни из лучших в мире многоязычных моделей. Примечательно, что в релизе Qwen 3 (апрель 2025) компания экспериментировала с «гибридным рассуждением», пытаясь научить модель саму решать — отвечать мгновенно или долго «думать» (Chain of Thought). Однако позже разработчики отказались от этого подхода в пользу выпуска раздельных, более качественных версий (Instruct и Thinking). В феврале 2026 года вышло поколение Qwen 3.5, открытые версии которого регулярно делят лидерство в рейтингах Hugging Face с моделями от Meta и DeepSeek.

5. GEMMA И PHI (GOOGLE И MICROSOFT)

Даже гиганты закрытого софта были вынуждены выпустить открытые версии своих наработок.

Gemma 2: Использует те же технологии, что и Gemini. Очень мощная в категории «среднего веса» (9B и 27B параметров).

Phi-4 (декабрь 2024): Семейство моделей на 14B параметров, демонстрирующее невероятные результаты для своих размеров. Существуют версии:

- Phi-4 — базовая модель
- Phi-4-reasoning (апрель 2025) — для сложных рассуждений
- Phi-4-mini-reasoning (3.8B, апрель 2025) — компактная версия

Доказывает, что «учебники» (качественные данные) важнее «библиотек» (объема данных).

ТЕХНИЧЕСКАЯ МАГИЯ — КАК ОНИ РАБОТАЮТ?

Для читателя научно-популярного журнала важно понимать, что современная LLM — это не просто статистический автозаполнитель текста.

Context Window (Окно контекста): Современные открытые модели могут «удерживать в голове» до 128 000 и даже 1 000 000 токенов (целые книги или архивы кода). Это достигается за счет алгоритмов вроде Flash Attention.

Fine-tuning (Дообучение): Благодаря технологиям LoRA (Low-Rank Adaptation), вы можете дообучить гигантскую модель под свои нужды всего за несколько часов на одной видеокарте. Это превращает универсальный ИИ в узкоспециализированного эксперта.

RAG (Retrieval-Augmented Generation): Вместо того чтобы пытаться запихнуть все знания мира в веса модели, открытые системы подключаются к внешним базам данных. Модель «читает» ваши документы в реальном времени и отвечает на основе проверенных фактов, минимизируя галлюцинации.

ЭТИКА, БЕЗОПАСНОСТЬ И «ТЕМНАЯ СТОРОНА»

Открытость несет в себе риски. Если у каждого есть доступ к «цифровым мозгам» уровня гения, что мешает использовать их во зло?

Риски: Создание персонализированного фишинга, генерация дезинформации

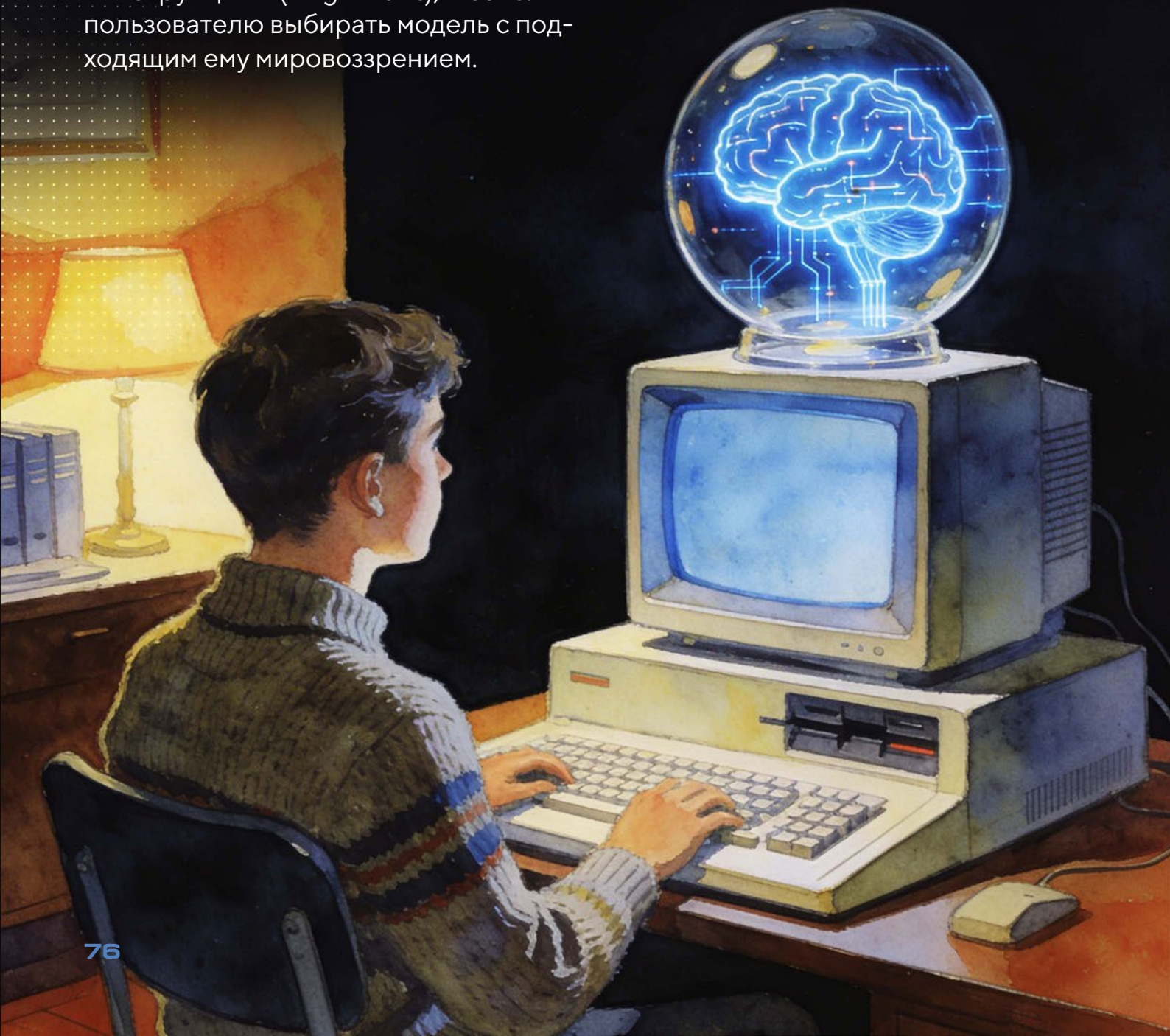
или помощь в создании биологического оружия.

Контраргумент («Закон Линуса»): Сторонники Open Source утверждают, что «при достаточном количестве глаз все ошибки лежат на поверхности». Сообщество быстрее находит уязвимости и создает инструменты защиты, чем это делают закрытые лаборатории.

Борьба с предвзятостью: В закрытых моделях политические и этические фильтры навязываются корпорациями. В открытых — сообщество само формирует различные наборы этических «инструкций» (Alignment), позволяя пользователю выбирать модель с подходящим ему мировоззрением.

Open Source AI — это не просто бесплатная альтернатива платным сервисам. Это гарантия того, что самая мощная технология в истории человечества не окажется в руках узкой группы лиц.

Как когда-то Linux стал фундаментом современного интернета, открытые языковые модели становятся фундаментом будущего интеллекта. Мы присутствуем при моменте, когда «магия» ИИ становится доступной каждому, у кого есть любопытство и доступ к сети. Прометей не просто украл огонь — на этот раз он раздал спички всем желающим.



РИСКИ И ИЛЛЮЗИИ

ЧЕГО СТОЯИТ БОЯТЬСЯ НА САМОМ ДЕЛЕ

Реальные риски ИИ значительно прозаичнее апокалипсиса — и именно поэтому они опаснее. Их легко не заметить.

- **1 РЕАЛЬНЫЙ РИСК: уверенная ложь.** Языковые модели галлюцинируют — выдают несуществующие факты, ложные цитаты, неверные даты с полной уверенностью в голосе.
- **2 РЕАЛЬНЫЙ РИСК: атрофия собственного мышления.** Если вы постоянно просите ИИ думать за вас — вы постепенно разучиваетесь думать сами. Это не метафора. Это когнитивный процесс, который фиксируется у людей, чрезмерно полагающихся на внешние инструменты мышления. Инструмент должен усиливать мышление, а не заменять его.
- **3 РЕАЛЬНЫЙ РИСК: иллюзия понимания.** ИИ может создать убедительное ощущение, что вы разобрались в теме — потому что получили структурированный, грамотный, детальный ответ. Но ответ ИИ — это не ваше знание. Это информация, которую вы ещё не переработали, не проверили, не встроили в собственную систему понимания. Путать одно с другим — серьёзная ошибка.
- **4 РЕАЛЬНЫЙ РИСК: зависимость от конкретной платформы.** Платформы меняются, закрываются, меняют политику доступа. Профессионал, который выстроил весь рабочий процесс вокруг одного инструмента, уязвим. Диверсификация — не роскошь, а профессиональная необходимость.

- **1 ИЛЛЮЗИЯ: промпт решает всё.** Существует целая индустрия «промпт-инжиниринга» с курсами, сертификатами и обещаниями профессионального роста. Правда проще: качество взаимодействия с ИИ определяется прежде всего качеством профессионального мышления пользователя, а не техникой формулировки запросов. Мастерство промпта — полезный навык. Замена профессиональному мышлению — никогда.
- **2 ИЛЛЮЗИЯ: один правильный инструмент.** Нет платформы, которая делает всё лучше всех. Есть платформы, каждая из которых делает что-то лучше остальных. Профессионал работает с арсеналом, а не с одним оружием.
- **3 ИЛЛЮЗИЯ: ИИ объективен.** Языковые модели несут в себе ценности и предположения, заложенные при обучении. Они не нейтральны. Разные платформы — разные предубеждения. Критическое отношение к источнику применимо к ИИ так же, как к любому другому источнику информации.

ПРИНЦИПИАЛЬНОЕ УСТРОЙСТВО БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ

Принципиальное устройство большой языковой модели (LLM) можно представить как сложную математическую функцию, основной задачей которой является предсказание вероятностей для всех возможных вариантов следующего слова (токена) в заданном фрагменте текста.

В основе её работы лежат следующие ключевые компоненты и процессы:

1. Представление данных: Токены и Эмбеддинги

Токенизация: Входной текст разбивается на мелкие фрагменты — токены (слова, части слов или знаки препинания).

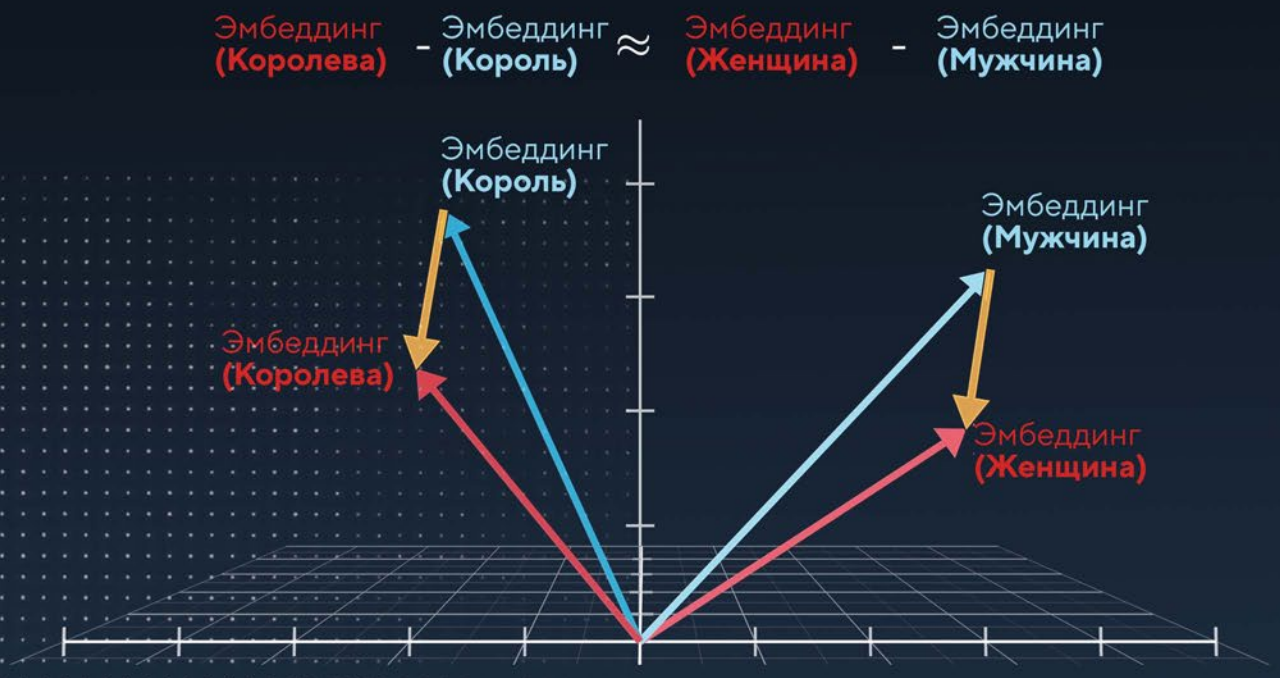
Матрица эмбеддингов: Каждый токен связывается с уникальным списком чисел — вектором, который кодирует его значение. Эти числа можно представить как координаты в многомерном пространстве, где близкие по смыслу слова располагаются рядом.

2. Архитектура Трансформера

Современные модели строятся на архитектуре трансформера, которая позволяет обрабатывать весь текст параллельно, а не последовательно. Она состоит из двух основных типов операций, которые повторяются многократно:

Механизм внимания (Attention): Это «сердце» трансформера. Он позволяет векторам слов «обмениваться информацией» и уточнять свой смысл в зависимости от контекста. Например, вектор слова «банк» изменится, если рядом стоят слова, указывающие на реку или на финансовую организацию.

Полносвязные нейросети (Feed-forward layers): Эти слои обрабатывают каждый вектор параллельно и помогают модели запоминать сложные языковые закономерности, усвоенные в ходе обучения.



Кадр из видео «Transformers, the tech behind LLMs» © YouTube-канал 3Blue1Brown.
Оригинал: youtu.be/wjZofJX0v4M.

3. «Мозг» модели: Параметры и Веса
Поведение модели определяется её параметрами (весами) — числовыми значениями, которые настраиваются в процессе обучения.

Масштаб: В больших моделях количество параметров исчисляется сотнями миллиардов (у GPT-3 их 175 миллиардов).

Матричные вычисления: Почти все фактические вычисления внутри модели представляют собой матрично-векторное умножение. Веса организованы в тысячи отдельных матриц.

4. Генерация ответа и управление вероятностью
Логиты и Softmax: На выходе сеть выдает набор необработанных чисел (логитов) через матрицу анэмбеддинга. Затем используется функция softmax, которая превращает эти числа в распределение вероятностей, где сумма всех значений равна единице.

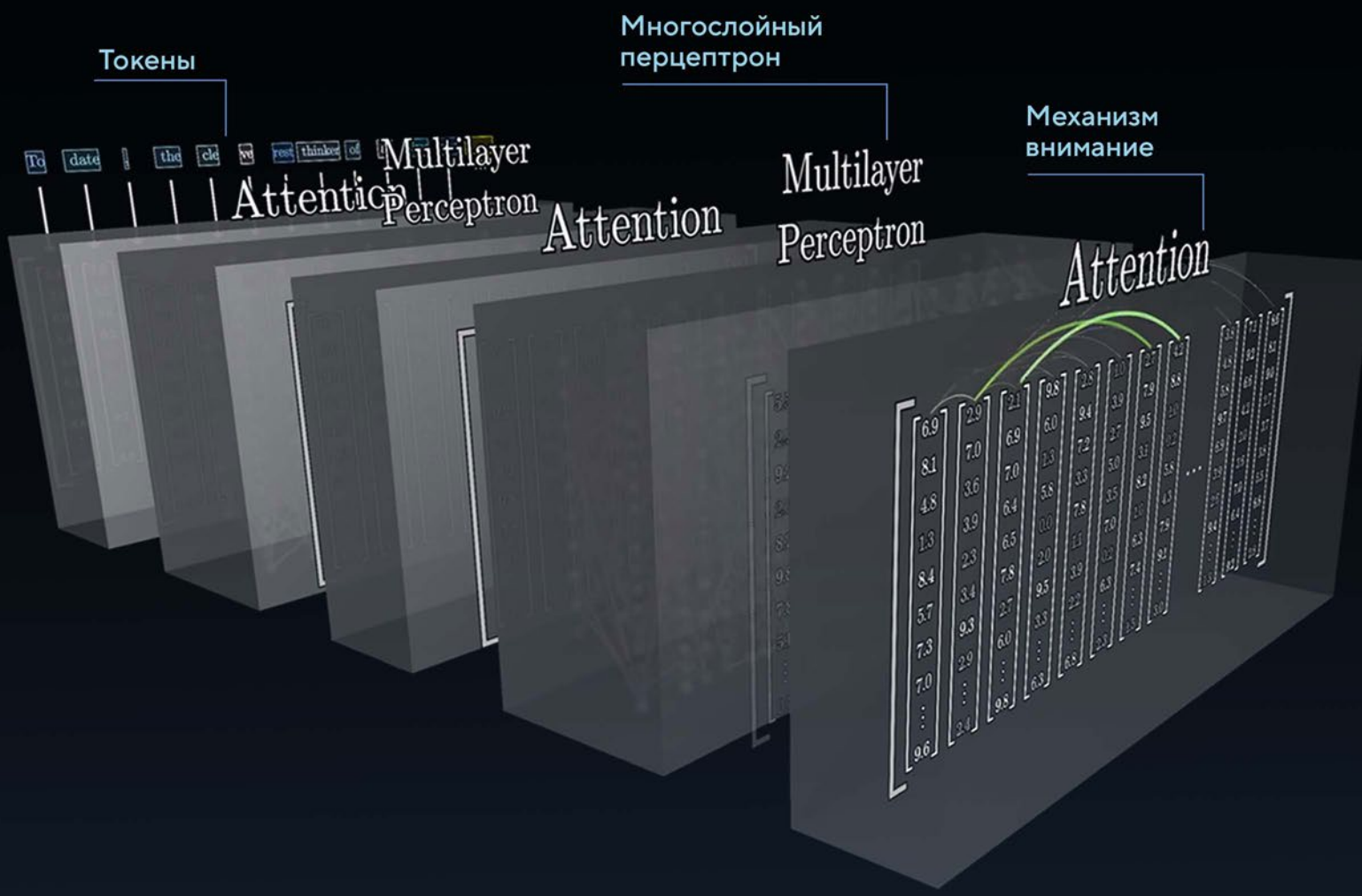
Температура: Этот параметр регулирует случайность выбора. При низкой температуре модель выбирает самые вероятные слова (делая текст логичным, но предсказуемым), а при высокой — дает шанс менее вероятным вариантам, что добавляет «творчества», но может привести к бессмыслице.

5. Ограничения контекста
Модель может одновременно учитывать лишь определенный объем текста, называемый размером контекста. Если диалог превышает предел контекста, модель начинает «терять нить» беседы, так как старые данные выходят за рамки её вычислительного окна.

Таким образом, языковая модель — это гигантская система весов, которая путем миллиардов математических операций трансформирует входные числа так, чтобы максимально точно предсказать следующее слово, опираясь на накопленные в процессе обучения знания.

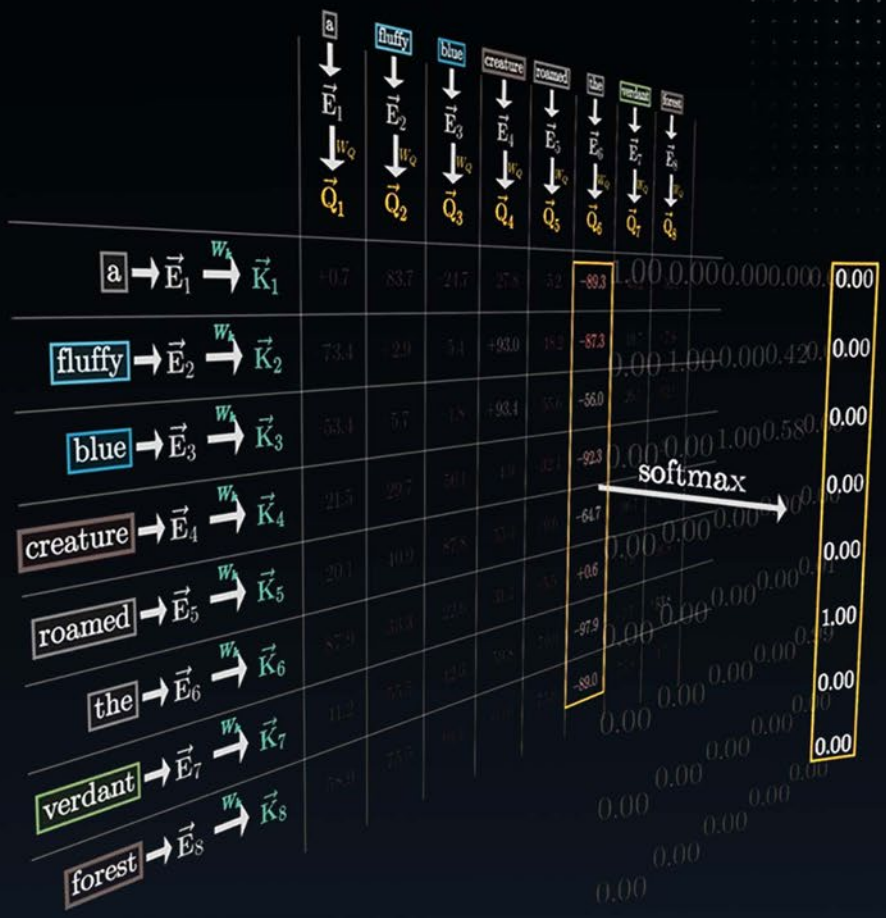
Кадр из видео «Transformers, the tech behind LLMs» © YouTube-канал 3Blue1Brown.
Оригинал: youtu.be/wjZofJX0v4M.

TRANSFORMER
МЕХАНИЗМ SELF-ATTENTION



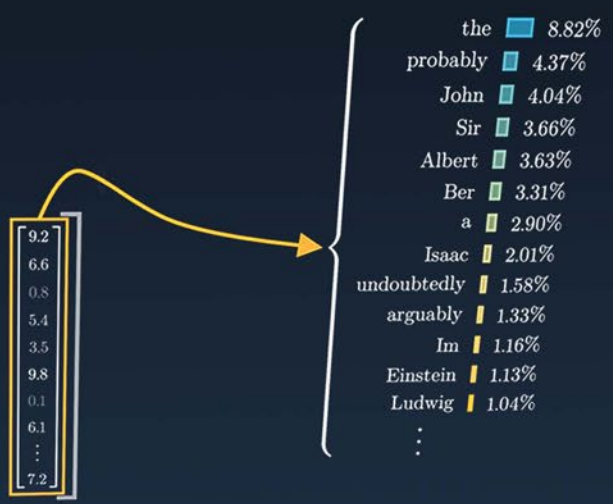
На этой схеме представлена работа слоев внимания. С помощью векторов Запросов (Q) и Ключей (K) модель вычисляет степень релевантности каждого слова контексту. Функция softmax нормализует эти значения, позволяя модели «фокусироваться» на наиболее важных смысловых связях при генерации следующего токена.

Кадр из видео «Transformers, the tech behind LLMs» © YouTube-канал 3Blue1Brown.
Оригинал: youtu.be/wjZofJX0v4M.



To date, the cleverest thinker of all time was

???



ПРИЛОЖЕНИЕ
ИЕРАРХИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

02 ТОП-12 МОДЕЛЕЙ МИРА

Рейтинг по совокупной мощности · март 2026

#	МОДЕЛЬ	КОМПАНИЯ	МОЩНОСТЬ	КЛЮЧЕВОЕ ПРЕИМУЩЕСТВО	СТАТУС
1	**Gemini 3.1 Pro** Google DeepMind	Google		ARC-AGI-2 77.1% · GPQA 94.3% · лидер бенчмарков	ACTIVE
2	**Claude Opus 4.6** Anthropic	Anthropic		SWE-bench 75.6% · 1M контекст · агентные команды	ACTIVE
3	**GPT-5.4** OpenAI	OpenAI		GPQA 92.8% · крупнейшая экосистема · мультимодал	ACTIVE
4	**Grok 4.20** xAI · Elon Musk	xAI		4 параллельных агента · реальное время X/Twitter	BETA
5	**Claude Sonnet 4.6** Anthropic	Anthropic		98% Opus-качества · $3/$15 · лучшая ценность	ACTIVE
6	**DeepSeek R2** DeepSeek · Китай	DeepSeek		Прорывная цена · сильная математика/код · открытый	ACTIVE
7	**Qwen 3.5** Alibaba · Китай	Alibaba		Открытые веса · быстро закрывает разрыв с топами	ACTIVE
8	**Llama 4 Scout** Meta AI	Meta		MoE архитектура · нативный мультимодал · бесплатный	ACTIVE
9	**Mistral Large 3** Mistral · Франция	Mistral		Европейский суверенитет · код · многоязычность	ACTIVE
10	**Perplexity Sonar** Perplexity AI	Perplexity		Поиск + ИИ · цитаты · исследовательские задачи	ACTIVE

01 ПИРАМИДА МОЩНОСТИ ИИ

Уровни развития от узкоспециализированных до систем уровня AGI

УРОВЕНЬ AGI — ⚡ **Сверхмощные агентные системы** — 94–97% GPQA
Gemini 3.1 Pro · Claude Opus 4.6 · GPT-5.4 · Grok 4.20
Контекст 1M+ токенов · Агентные команды · Параллельное мышление · Сверхчеловеческие бенчмарки

FRONTIER — 🔵 **Флагманские мультимодальные модели** — 85–93% GPQA
Claude Sonnet 4.6 · GPT-5.2 · Gemini 3 Pro · Grok 4.1
Текст + изображения + аудио + код · Расширенный контекст · Инструментальное использование

ADVANCED — 🟣 **Продвинутые LLM второго эшелона** — 70–84% GPQA
DeepSeek R2 · Qwen 3.5 · Mistral Large · Llama 4 Scout
Высокое качество при меньших затратах · Открытые веса · Региональные рынки

CAPABLE — 🟢 **Рабочие модели широкого применения** — 50–69% GPQA
Claude Haiku 4.5 · Gemini 3 Flash · GPT-4o mini · Llama 4 Maverick
Быстро · Дёшево · Надёжно для большинства задач · Массовое применение

SPECIALIST — ⚙ **Специализированные и встраиваемые** — <50% GPQA
Phi-3 · Gemma 3 · Mistral 7B · Edge-модели
Устройства · Оффлайн · Конкретные задачи · Ограниченные ресурсы

▪ GPQA Diamond — эталонный тест экспертных знаний (физика, химия, биология) ▪ SWE-bench — решение реальных задач программирования ▪ ARC-AGI-2 — логика без возможности мемоизации

03 КОМПАНИИ — СОЗДАТЕЛИ ИИ

Игроки первого и второго эшелона · финансовая мощь · флагманы

Google DeepMind

США · Alphabet Inc.

Gemini 3.1 Pro · Gemini 3 Flash · Gemma 3
Оценка: $2 трлн+ · 1 место по бенчмаркам

OpenAI

США · Microsoft партнёр

GPT-5.4 · GPT-5.2 · o4-mini · Codex
Оценка: $157 млрд · крупнейшая экосистема

Anthropic

США · Amazon партнёр

Claude Opus 4.6 · Sonnet 4.6 · Haiku 4.5
Оценка: $61 млрд · лидер безопасности

xAI (Elon Musk)

США · X/Twitter платформа

Grok 4.20 · Grok 4.1 · Grok 3
Оценка: $50 млрд · Colossus 200K GPU

Meta AI

США · Meta Platforms

Llama 4 Scout · Maverick · Behemoth
Оценка: $1.5 трлн · открытые веса

DeepSeek

Китай · High-Flyer Quant

DeepSeek R2 · V3 · открытые веса
Прорыв: топ-качество за 1/10 цены

Alibaba / Qwen

Китай · Alibaba Group

Qwen 3.5 · Qwen 2.5-VL · открытые
Сильный в математике и многоязычности

Mistral AI

Франция · €1 млрд оценка

Mistral Large 3 · Codestral · Mixtral
Европейский суверенный ИИ

07 ГЕОПОЛИТИКА ИИ

Кто владеет ИИ — тот владеет будущим · расстановка сил по странам

США

// МИРОВОЙ ЛИДЕР

- OpenAI · GPT-5.4
- Anthropic · Claude 4
- Google · Gemini 3
- xAI · Grok 4
- Meta · Llama 4
- Microsoft · Copilot

Влияние

Китай

// СТРЕМИТЕЛЬНЫЙ CHALLENGER

- DeepSeek R2 / V3
- Alibaba · Qwen 3.5
- Baidu · ERNIE 5
- Huawei · Pangu
- ByteDance · Doubao

Влияние

Европа

// СУВЕРЕННЫЙ ПУТЬ

- Mistral AI (FR)
- Aleph Alpha (DE)
- AI Sweden (SE)
- EuroLLM (совместный)

Влияние

Арабский мир / ОАЭ

// ВОСХОДЯЩИЙ ИГРОК

- G42 · Falcon 3 (ОАЭ)
- Jais · Arabic LLM
- SDAIA (Саудовская Аравия)

Влияние

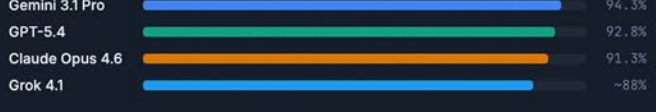
⚡ КЛЮЧЕВОЙ ВЫВОД: США контролируют 7 из 10 мощнейших моделей мира. Китай — единственный реальный конкурент, прорвавшийся через ценовой барьер (DeepSeek). Европа строит суверенный ИИ, но отстаёт на 2–3 года. Геополитическая гонка ИИ — это новая космическая гонка XX века.

04 КЛЮЧЕВЫЕ БЕНЧМАРКИ

Объективные показатели мощности · результаты топ-моделей

GPQA DIAMOND

Экспертные вопросы по физике, химии, биологии. Эталон научного интеллекта.

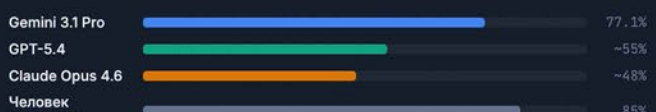


SWE-BENCH VERIFIED

Решение реальных задач программирования из GitHub. Эталон инженерного ИИ.

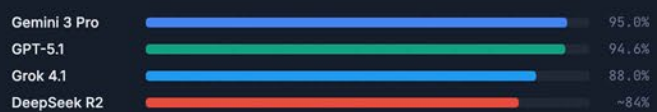
ARC-AGI-2

Абстрактное мышление без мемоизации. Ближайший к AGI тест.



AIME 2025 (МАТЕМАТИКА)

Олимпиадные задачи по математике высшего уровня.



05 МАТРИЦА ВОЗМОЖНОСТЕЙ

Что умеет каждая из топ-моделей · сравнение по ключевым навыкам

● Лидер ◉ Отлично ◎ Хорошо ○ Базово

Возможность	Gemini 3.1 Pro	Claude Opus 4.6	GPT-5.4	Grok 4	DeepSeek R2	Llama 4
💻 Программирование	◉	●	◉	●	◉	◎
🧠 Логика и рассуждение	●	◉	◉	◉	◉	◎
📐 Математика	●	◉	●	◉	●	◎
✍️ Создание текстов	◎	●	◉	◎	◎	○
🖼️ Анализ изображений	●	◉	◉	◎	◎	◉
🎬 Видео/Аудио	●	○	◉	○	○	◎
🔍 Реальное время / поиск	◉	◎	◎	●	○	○
📄 Большие документы	●	●	◎	◎	◎	◎
🤖 Агентные задачи	◉	●	◉	●	◎	○
🌍 Многоязычность	●	◉	◉	◎	◉	◉
🔒 Безопасность / этика	◎	●	◉	○	◎	◎
💰 Цена / качество	●	○	◎	◉	●	●

06 ОТКРЫТЫЕ vs ЗАКРЫТЫЕ МОДЕЛИ

Фундаментальное разделение рынка ИИ · плюсы, минусы, игроки

🔓 ОТКРЫТЫЕ (OPEN SOURCE)

ПРЕИМУЩЕСТВА

- ✅ Свободное использование и модификация
- ✅ Работа локально — без передачи данных
- ✅ Нет подписки — нет зависимости от вендора
- ✅ Быстрое развитие через сообщество

ОГРАНИЧЕНИЯ

- ⚠️ Ниже по мощности vs топ-закрытые
- ⚠️ Требует инфраструктуры для запуска

МОДЕЛИ

Llama 4 Scout	Meta
Qwen 3.5	Alibaba
DeepSeek R2	DeepSeek
Mistral Large	Mistral
Gemma 3	Google
Phi-3 / Phi-4	Microsoft

VS

🔒 ЗАКРЫТЫЕ (PROPRIETARY)

ПРЕИМУЩЕСТВА

- ✅ Максимальная мощность и качество
- ✅ API готово к работе — без инфраструктуры
- ✅ Постоянное обновление без усилий
- ✅ Техподдержка и SLA для бизнеса

ОГРАНИЧЕНИЯ

- ⚠️ Платная подписка / токены
- ⚠️ Зависимость от вендора

МОДЕЛИ

Gemini 3.1 Pro	Google
Claude Opus 4.6	Anthropic
GPT-5.4	OpenAI
Grok 4.20	xAI
Perplexity Sonar	Perplexity

Издатель:
Virtual Franciscans OÜ
Регистрационный код:
16190480
Место издания:
Таллин, Эстония
Сайт:
https://spacecode.franciscans.dev
E-mail:
journal@franciscans.dev