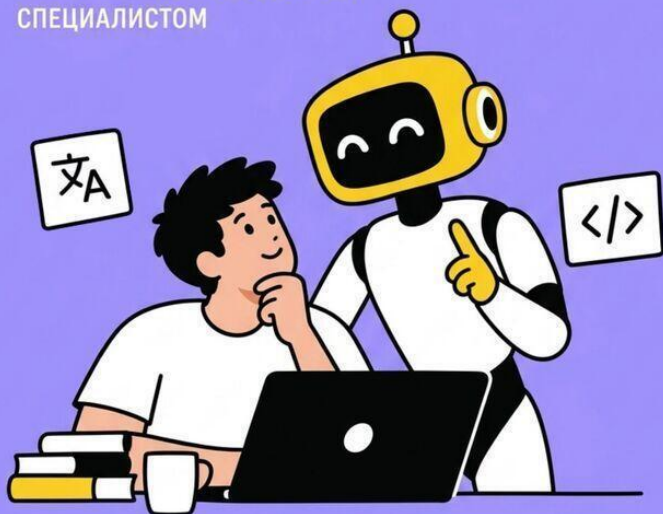


12+

РЕНАТ ФАТУЛЛИН

НЕЙРОСЕТИ ЧЕЛОВЕЧЕСКИМ ЯЗЫКОМ

ОСВОЙТЕ ИИ
И СТАНЬТЕ ВОСТРЕБОВАННЫМ
СПЕЦИАЛИСТОМ



Ренат Фатуллин

**Нейросети человеческим
языком. Освойте ИИ и станьте
востребованным специалистом**

<https://litres.ru/74543778>

ISBN 9785007133906

Аннотация

Вы уже пробовали нейросети и получили общий, бесполезный ответ. Эта книга объясняет, почему так вышло и что делать иначе. Без кода, формул и обзоров технологий — только то, что работающий специалист применит сегодня. После каждой главы вы что-то делаете: настраиваете рабочее место, собираете свою библиотеку приёмов, учитесь проверять результат. На выходе — более лёгкая работа, выше ценность на рынке и понимание, где здесь новые возможности.

Содержание

Вступление	5
Часть I. Что это такое и как мы сюда пришли	9
Глава 1. Как мы сюда пришли: короткая история ИИ	9
Глава 2. Как это работает: без математики и кода	33
Глава 3. Кто делает ИИ: лаборатории и модели	49
Глава 4. Сколько это стоит и кто на этом зарабатывает	69
Конец ознакомительного фрагмента.	80

**Нейросети
человеческим языком
Освойте ИИ и станьте
востребованным
специалистом**

Ренат Фатуллин

© Ренат Фатуллин, 2026

ISBN 978-5-0071-3390-6

Создано в интеллектуальной издательской системе Ridero

Вступление

Через неделю после релиза

Я открыл его не в день выхода, а примерно через неделю — в начале декабря 2022 года. Пост OpenAI «Introducing ChatGPT» вышел 30 ноября, без презентации на сцене, почти буднично. Я задал случайный вопрос, не помню уже какой, и полминуты читал ответ: ровный, связный, по-настоящему живой текст, а не пересказ первой страницы поисковой выдачи и не программа с заранее прописанными фразами.

Ощущение было не «удобно». Ощущение было «что-то сместилось». Как будто мир, каким мы его знали, изменился.

Скажу сразу, чтобы вы могли сделать на это скидку: я по складу характера ранний последователь. Новое устанавливаю первым, на любую технологию смотрю с интересом, а не с подозрением, и открыл ChatGPT в первую же доступную мне неделю не по работе, а потому что не мог не открыть. Это не нейтральная позиция, и лёгкость, с которой мне даются такие решения, — не то же самое, что лёгкость, с которой они дадутся вам. Мой собственный путь от «попробовал — не заработало» до «работаю с этим каждый день» занял два года, не два месяца. Что случилось в середине — в первой главе.

Возможно, один из самых значимых моментов

У меня нет привычки к громким формулировкам, но

здесь без одной не обойтись: похоже, мы застали один из самых значимых периодов за всю историю человечества. Меняется не частность — меняется то, на чём держались представления о работе, об уме и о том, кто получает преимущество в профессии. Эта книга — не попытка предсказать, чем всё кончится. Это попытка дать вам навигацию по уже начавшимся переменам, чтобы вы разобрались в них раньше, чем они разберутся с вами, и прошли этот отрезок пути успешно — на своём рабочем месте, с той профессией, которая у вас уже есть.

Три вещи, ради которых эта книга написана

Первая — не остаться за бортом. Скажу это прямо и только один раз, а дальше не вернусь к этому страху: перемены такого масштаба задевают любую профессию, и человеку, который годами строил карьеру, разумно спросить, что с ней будет. Сроками и катастрофой пугать не буду — точных сроков не знает никто, а книга, которая берётся их называть, вас обманывает.

Вторая, и главная: работать лучше и жить легче. Меньше рутины, быстрее черновики, спокойнее вечера, выше качество того, что вы в итоге отдаёте клиенту, начальнику или самому себе.

Третья: найти новые возможности. Внутри своей компании, в виде новой услуги, подработки или своего проекта — спрос на людей, которые умеют это по-настоящему, уже есть. Как он выглядит и как проверить свою гипотезу на малом,

не бросая основную работу, — в главах 20 и 21.

Кто вам это рассказывает

Коротко — подробный путь ждёт вас в конце книги, в главе «Об авторе». Я не инженер и не обучаю модели: этим занимаются другие люди, и там, где книге нужна инженерная глубина, я буду говорить прямо — здесь я не эксперт. Я руковожу производством видеоконтента для внешних клиентов: сроки, бюджеты, команды по всему миру. ИИ для меня не предмет изучения ради самого изучения, а инструмент, которым я эту работу делаю каждый день.

Главный принцип: после каждой главы вы что-то сделаете

После каждой главы — конкретное задание на десять-тридцать минут, с результатом, который можно потрогать на вашей рабочей задаче. Пропускайте задания — и на выходе получите общее представление вместо рабочего навыка. Тратьте на них время — и через двадцать шесть глав у вас будет не книга на полке, а система, которая уже работает.

Как здесь даются советы

По каждому шагу я называю один порядок действий — тот, который считаю правильным. Альтернативы, если они есть, идут после и коротко. Вы не найдёте здесь «зависит от ваших предпочтений» или «попробуйте оба варианта и решите сами» — это способ ничего не посоветовать, спрятанный за уважением к вашей самостоятельности. Там, где ваша ситуация действительно другая, в тексте прямо сказано, где

свернуть.

Как читать эту книгу

Совет по умолчанию простой: читайте по порядку и не пропускайте разделы. Ни один здесь не лишний — если сейчас какой-то покажется не вашим, через две-три главы связь станет видна, а цельная картина сложится только у того, кто прошёл книгу целиком.

Ещё до первой страницы

Книга написана в сентябре 2026 года, и это важно: половина в ней — навык, который не устареет, половина — снимок того, чем сейчас пользуются люди и компании. Всё, что может устареть за несколько месяцев — версии, цены, лимиты, — стоит отдельно, во врезке «Сейчас» с датой, и никогда не в основном тексте. Дата редакции — на титульном листе.

И формальность, которую лучше сказать один раз и забыть: книга написана для практики и не заменяет юриста или финансового консультанта там, где решение стоит денег или подписи. Ни с одной из упомянутых на этих страницах компаний я не связан коммерческими отношениями.

Договорились. Вместе с первой главой начинается ваш путь в удивительный мир искусственного интеллекта.

Часть I. Что это такое и как мы сюда пришли

Глава 1. Как мы сюда пришли: короткая история ИИ

8 июля 1958 года «Нью-Йорк Таймс» вышла с заметкой под заголовком «Новое устройство ВМС учится на практике». Речь шла о машине, которую психолог Фрэнк Розенблатт показал журналистам в Корнеллской авиационной лаборатории в Буффало. Газета передала ожидания военных без тени иронии: перед публикой «зародыш электронного компьютера, который, как рассчитывает флот, сможет ходить, говорить, видеть, писать, воспроизводить себя и осознавать своё существование». Журнал «Нью-Йоркер» добавил, что это «первый серьёзный соперник человеческого мозга из когда-либо созданных».

Умела она к тому моменту следующее: после пятидесяти попыток научилась отличать перфокарты с меткой слева от перфокарт с меткой справа.

Устройство называлось перцептрон. Оно смотрело на мир через решётку из четырёхсот фотоэлементов, двадцать на двадцать. Его память состояла из потенциометров — тех са-

мых ручек-регуляторов, что стояли в ламповых радиоприёмниках. Когда машина ошибалась, электромоторы поворачивали ручки, и в следующий раз она ошибалась чуть меньше.

Между этой машиной и той, которая сегодня пишет за вас деловые письма, идёт прямая линия. Идея за шестьдесят восемь лет изменилась удивительно мало. Изменились условия, при которых она работает, и на то, чтобы эти условия сошлись в одной точке, ушло больше полувека. Эта глава — про то, каких именно условий не хватало и почему они появились только сейчас.

Вы это уже пробовали

Скорее всего, ваш первый настоящий разговор с моделью случился в 2023 году, и закончился он смешанным чувством.

Вы попросили написать коммерческое предложение и получили текст, который подошёл бы любой компании в любой отрасли. Вы попросили посчитать — и получили красивое число, которое при проверке оказалось выдуманным. Вы задали вопрос про свою профессию, в которой работаете пятнадцать лет, и получили пересказ первых страниц поисковой выдачи. Где-то на третьей попытке появилась мысль: вокруг этого слишком много шума.

Мысль была верной, и у неё была техническая причина — даже несколько.

Модель того поколения отвечала одним проходом: получила вопрос, сразу начала писать ответ и не имела возможности остановиться и переписать. Она не видела ваших до-

кументов, ваших цифр и ваших прошлых проектов, поэтому отвечала усреднённо по всему, что видела при обучении. Она не умела пользоваться инструментами: не могла открыть браузер, чтобы проверить факт, и калькулятор, чтобы проверить арифметику. И она была устроена так, что уверенный тон достаётся ей бесплатно, а точность — нет.

Каждое из этих четырёх ограничений за последние три года было снято отдельным техническим решением. Разбирать их я буду в следующих главах. Но чтобы понимать, чему верить в новостях и чего ждать дальше, полезно сначала увидеть всю дугу целиком: откуда эта идея взялась, почему дважды казалась мёртвой и что конкретно изменилось.

Лето 1956 года: у идеи появляется имя

31 августа 1955 года четыре человека подали в Рокфеллеровский фонд заявку на 13 500 долларов. Они просили денег на двухмесячный летний семинар для десяти исследователей.

Заявку подписали: Джон Маккарти, двадцативосьмилетний ассистент-профессор математики Дартмутского колледжа; Марвин Минский из Гарварда; Натаниэль Рочестер из IBM, конструктор первого серийного компьютера компании; и Клод Шеннон из Bell Labs, автор теории информации, на которой стоит вся современная связь. Компания, мягко говоря, не рядовая.

Дартмутский колледж — частный университет Лиги плюща в городке Ганновер, штат Нью-Гэмпшир, на границе

с Вермонтом. Место тихое и удалённое, и это было частью замысла: собрать людей там, где им не помешают.

В заявке была фраза, которую с тех пор цитируют как свидетельство о рождении области:

«Исследование будет вестись на основе предположения, что всякий аспект обучения или любое иное свойство интеллекта в принципе может быть описано настолько точно, что машина сможет его смоделировать».

Обратите внимание на «в принципе». Это не обещание результата, это заявленная рабочая гипотеза. За семьдесят лет она ни разу не была ни доказана, ни опровергнута.

Семинар прошёл с 18 июня по 17 августа 1956 года. Устроен он был как **мастерская**: люди приезжали и уезжали, кто на неделю, кто на месяц; постоянно присутствовали трое. Работали по семи темам, среди которых были автоматические вычислители, использование машиной языка, **нейронные сети**, самоулучшение программы, абстракции, а также случайность и творчество.

Отдельного упоминания заслуживает само название. Термин «**искусственный интеллект**» (artificial intelligence, отсюда привычное сокращение AI) придумал Маккарти, и придумал по причине вполне земной: нужно было слово, не занятое чужой школой. Главной школой того времени была кибернетика Норберта Винера, человека влиятельного и тяжёлого в общении. Новое имя позволяло не воевать за определения и не работать под чужим флагом.

Итог семинара: у направления появились имя, программа исследований и круг людей. Ответа на главный вопрос не появилось ещё пятьдесят шесть лет.

Как машина «учится»: главная идея, которой семьдесят лет

Здесь стоит остановиться и разобрать механику, потому что всё остальное в книге на неё опирается.

Обычная программа — это записанные человеком правила. Бухгалтерская программа считает налог, потому что программист прочитал закон и перевёл его в инструкции. Такой подход упирается в стену там, где правило невозможно записать: попробуйте объяснить словами, по каким признакам вы отличаете кошку от собаки на фотографии, или почему одно письмо звучит вежливо, а другое грубо.

Розенблатт предложил другой путь. Его машина не содержала правил распознавания. Она содержала **веса** (weights) — числа, которые определяют, насколько сильно каждый вход влияет на ответ. В «Марк I» эти числа были физическими: положение ручки потенциометра. Дальше работал цикл, который лежит в основе всего машинного обучения до сегодняшнего дня:

1. Машине показывают пример с известным правильным ответом.
2. Она выдаёт свой ответ.
3. Ответы сравнивают, ошибку измеряют.
4. Веса чуть-чуть подкручивают в сторону, которая умень-

шает ошибку.

5. Повторить. Много раз.

Этот способ называется **обучением с учителем** (supervised learning): учитель здесь — набор примеров с правильными ответами. Программу в таком подходе не пишут, её настраивают. Знание не формулируется словами, оно оседает в тысячах чисел, каждое из которых само по себе ничего не значит.

Идея красивая, и у неё сразу обнаружился предел. В 1969 году Марвин Минский — тот самый, из дартмутской заявки — вместе с Сеймуром Папертом выпустил книгу «Перцептроны». В ней было показано, что однослойная сеть не способна выучить даже элементарную функцию «либо одно, либо другое, но не оба сразу» (в информатике её называют **XOR**). Причина геометрическая: такая сеть умеет провести между примерами только одну прямую границу, а для XOR одной прямой не хватает.

Минский и Паперт знали, что многослойная сеть это ограничение снимает. Но никто не умел обучать многослойные сети: способ подкручивать веса в глубине, а не только на выходе, ещё не был найден. Книгу прочитали как приговор всему направлению. Деньги ушли, и на полтора десятилетия сети исчезли из повестки.

Розенблатт до реабилитации своей идеи не дожил: он утонул во время прогулки на яхте в 1971 году, в день своего сорокатрёхлетия.

Две зимы: почему идея дважды считалась мёртвой

У области есть собственный термин для таких периодов — **зима ИИ** (AI winter). Это не метафора настроения, это про деньги: момент, когда обещания не сбываются, финансирование прекращается и тема становится неприличной в заявках на гранты.

Первая зима, середина 1960-х — начало 1980-х. В 1966 году американская комиссия ALPAC подвела итог десятилетию работ по машинному переводу, на которые ушло около 20 миллионов долларов. Вывод: человек переводит дешевле и точнее машины. Финансирование закрыли полностью.

Через семь лет то же самое произошло в Великобритании. Физик Джеймс Лайтхилл по заказу парламента оценил результаты британских исследований и назвал главную причину провала — **комбинаторный взрыв** (combinatorial explosion). Понятие простое и важное: число вариантов, которые нужно перебрать, растёт быстрее, чем растут вычислительные возможности. Программа, которая уверенно решает задачу с десятью объектами, при сорока объектах не решит её до конца существования Вселенной. Методы того времени работали на игрушечных примерах и рассыпались на настоящих. После отчёта Лайтхилла исследования ИИ в Британии свернули почти полностью.

Второе дыхание и вторая зима, 1980-е. Область вернулась через **экспертные системы** (expert systems) — про-

граммы, в которые вручную записывали правила вида «если симптом А и симптом Б, то диагноз В», собранные из бесед с профессионалами. Это работало. Система XCON, конфигурировавшая заказы компьютеров в корпорации DEC, сэкономила ей около 40 миллионов долларов за шесть лет. К 1985 году корпорации тратили на такие разработки больше миллиарда долларов в год. Япония запустила национальный проект «ЭВМ пятого поколения» с бюджетом порядка 850 миллионов долларов.

Под это вырос отдельный рынок специализированных компьютеров — LISP-машин, заточенных под язык, на котором такие системы писали. Объём рынка доходил до полумиллиарда долларов в год. В 1987 году он исчез за двенадцать месяцев: обычные рабочие станции стали справляться с той же работой дешевле.

Экспертные системы рухнули по причине, которую стоит запомнить. Знание в них приходилось вносить руками, и стоило это дорого. Система не училась на новых случаях. Каждое исключение из правила требовало нового правила, а число исключений в реальном мире не кончается. Поддержка дорожала быстрее, чем росла польза.

Оба раза причина была одна: идею нечем было накормить.

Три условия и горький урок

Чтобы идея заработала, должны были сойтись три условия: способ обучать глубокие сети, дешёвые вычисления и данные в промышленных объёмах. Появлялись они пооди-

ночке, с большими перерывами, и до 2012 года ни разу не оказывались в одном месте.

Условие первое: способ обучать глубокие сети. 1986 год.

Сначала о том, что такое слой. Перцептрон был устроен в один слой: входы влияли на выход напрямую, поэтому он умел провести между примерами только одну границу. В многослойной сети между входом и выходом стоят промежуточные слои, и каждый следующий работает с тем, что выделил предыдущий. На фотографиях это выглядит так: первый слой реагирует на перепады яркости, второй складывает их в углы и дуги, третий — в детали вроде глаза или колеса, и только последний говорит «кошка». Что выделять на каждом уровне, никто не задаёт: сеть находит это сама. Отсюда и слово **глубокая** — слоёв много; отсюда же и термин **глубокое обучение** (deep learning), который означает буквально «обучение многослойных сетей», и больше ничего.

Здесь и была загвоздка. Подкрутить веса в последнем слое просто: видно ответ, видно правильный, видно ошибку. А что делать с весами в середине, где правильного ответа не существует и сравнивать не с чем? Задача называется распределением вины: надо понять, какая доля общей ошибки приходится на каждое из тысяч внутренних чисел.

Ответ дали в 1986 году Дэвид Румельхарт, Джеффри Хинтон и Рональд Уильямс — статья в журнале Nature про **обратное распространение ошибки** (backpropagation, в раз-

говоре обычно просто backprop). Идея: ошибку, измеренную на выходе, можно пересчитать назад, слой за слоем, и получить для каждого внутреннего веса его собственную долю вины. Дальше всё как у Розенблатта — каждый вес подкручивают в сторону меньшей ошибки, только теперь это работает на любой глубине.

Способ был найден. Мощности, на которой он даёт результат, ещё не существовало.

Условие второе: дешёвые параллельные вычисления. Обучение сети состоит из миллиардов однотипных простых операций — умножений и сложений. Обычный процессор считает их по очереди, одну за другой. Видеокарта устроена иначе: она рассчитана считать тысячи одинаковых операций одновременно, потому что именно это нужно, чтобы отрисовать миллионы пикселей игрового кадра. Задача обучения сети математически совпала с задачей отрисовки картинки. До середины 2000-х эти ядра были заперты внутри графики: обращаться к ним можно было только через графические библиотеки, то есть объясняя машине задачу на языке треугольников и текстур. В 2006 году Nvidia объявила **CUDA** — набор инструментов, позволяющий писать под эти ядра обычные расчётные программы, безо всякой графики; первая общедоступная версия вышла в 2007-м. Так видеокарты открылись для науки, а через несколько лет — и для обучения сетей. Игровая индустрия невольно оплатила подготовку инфраструктуры для ИИ.

Условие третье: размеченные данные в промышленных объёмах. В 2006 году исследовательница Фей-Фей Ли начала собирать базу ImageNet. К 2010 году в ней было 14,2 миллиона изображений, разложенных по 21 841 категории, и каждое изображение было подписано человеком. Разметку делали через краудсорсинговую площадку Amazon Mechanical Turk, где тысячи людей по всему миру выполняли микрозадачи за небольшие деньги; сложные случаи проходили до десяти кругов проверки. Такой базы не существовало бы без интернета: ни собрать столько картинок, ни нанять столько рук иначе физически невозможно.

Точка схождения: осень 2012 года. Алекс Крижевский, Илья Суцкевер и Джеффри Хинтон выставили на ежегодный конкурс по базе ImageNet свою сеть, известную теперь как AlexNet.

Задание конкурса выглядело так: программа получает фотографию и должна отнести её к одной из тысячи категорий, от немецкой овчарки до пожарной машины. Учили на 1,2 миллиона размеченных снимков, проверяли на отложенных, которых программа не видела. Попадание засчитывали, если верная категория оказывалась среди пяти вариантов, названных программой самыми вероятными. По этой мерке AlexNet ошибалась в 15,3% случаев против 26,2% у второго места — разрыв, какого в этом соревновании до тех пор не случалось. Обучена она была на двух игровых видеокартах GTX 580 по три гигабайта, за пять-шесть дней, и содержала

60 миллионов настраиваемых весов.

Ни одна из трёх составляющих не была новой сама по себе. Идея сети — 1958 год. Способ обучения — 1986. Видеокарты — товар из магазина электроники. Новым было только то, что всё это оказалось в одной комнате одновременно.

Из этого следует главный вывод главы, и он неприятный. В марте 2019 года Ричард Саттон, один из основателей обучения с подкреплением, опубликовал короткое эссе «Горький урок».

Сюжет, по его наблюдению, повторялся всю историю области. Исследователи вкладывали в программу своё понимание предмета: записывали шахматные принципы, описывали, как устроено человеческое зрение, закладывали правила грамматики. Такие программы работали, пока их не встретила на соревновании другая программа — более простая и ничего не знающая о предмете, зато получившая больше вычислений и больше данных. Проигрывали именно машины машинам. Так вышло в шахматах, в го, в распознавании речи и в компьютерном зрении.

Урок горький потому, что он обесценивает самое дорогое — экспертное знание о предмете. Он же объясняет, почему всё сдвинулось в 2010-е: считать стало дёшево.

Исследовательская группа Erosch AI ведёт базу так называемых заметных моделей. Модель попадает туда, если выполнено хотя бы одно условие: статья о ней собрала больше тысячи цитирований, либо она показала лучший результат

на признанном тесте, либо вошла в историю области, либо работает в продукте с большой аудиторией. По их расчёту на 428 таких моделях за период с 2010 года по сентябрь 2025-го объём вычислений, потраченных на обучение, растёт в 4,7 раза в год: удвоение каждые полгода. На последнем обновлении базы, ноябрь 2025 года, тренд держался.

Для сравнения — закон Мура, полвека считавшийся пределом скорости в технике. Гордон Мур сформулировал его в 1965 году: число транзисторов на кристалле удваивается примерно раз в год; в 1975-м уточнил период до двух лет. Так и держалось лет сорок, а с середины 2000-х замедляется — глава Intel Пэт Гелсингер в конце 2023 года говорил уже о трёхлетнем шаге. Удвоение раз в полгода против удвоения раз в два-три года: в таком темпе область живёт последние пятнадцать лет.

Язык: 2017—2022

Сети научились видеть. С языком дело шло тяжелее, и причина была в архитектуре.

До 2017 года язык обрабатывали **рекуррентными сетями** (по-английски recurrent neural network, RNN) и их улучшенным вариантом LSTM, придуманным в 1997 году. Устроены они так: слова подаются в сеть по одному, и после каждого слова сеть обновляет своё внутреннее состояние — короткую память о прочитанном. Чтобы взяться за десятое слово, нужно состояние после девятого, а для него — после восьмого. Шаги выстроены в цепочку, и разложить их на ты-

сачу видеокарт невозможно: каждая ждала бы предыдущую. Мощности росли, а язык от этого роста почти ничего не получал.

12 июня 2017 года восемь исследователей Google опубликовали статью с названием, ставшим поговоркой: «Внимание — это всё, что нужно». Они предложили архитектуру **трансформер** (transformer), в которой цепочки нет вовсе. Вместо неё работает **механизм внимания** (attention): обрабатывая очередной фрагмент текста, модель одновременно смотрит на все остальные и сама решает, какие из них сейчас важны для смысла.

Практическое следствие важнее теоретического. Весь текст теперь обрабатывается за один заход, а значит, обучение раскладывается на столько машин, сколько удастся собрать. Архитектура оказалась пригодной к масштабированию, а из горького урока следует, что именно это и решает.

Дальше пять лет масштаба. Всё это время модели решали одну и ту же задачу: предсказать продолжение текста, кусок за куском, на **корпусе** (corpus) — так называют большой набор текстов, собранный для обучения: копии веб-страниц, книги, статьи, код, переписку форумов. Никаких правил грамматики внутрь не закладывали. Как из одной этой задачи получается связный ответ на ваш вопрос — механика следующей главы.

Числа роста по одной линейке моделей: первое поколение, 2018 год, — 117 миллионов весов. Второе, 2019-й, — пол-

тора миллиарда. Третье, 2020-й, — 175 миллиардов. Принцип везде один и тот же. Менялся только масштаб, и вместе с масштабом появлялись способности, которых никто отдельно не программировал: перевод, пересказ, написание кода.

30 ноября 2022 года OpenAI открыла публичный доступ к ChatGPT. Под ним работала уже существовавшая модель той же линейки, дообученная на человеческих оценках: людям показывали варианты ответов, они отмечали лучшие, модель подстраивалась под предпочтения. Технически это было развитием того, что существовало два года. Изменилось другое: результат положили в окно чата, куда можно написать словами, без программирования и без документации. За два месяца сервис набрал около ста миллионов пользователей и стал самым быстрорастущим потребительским приложением в истории на тот момент.

Обратите внимание, что именно случилось в ноябре 2022 года. Сдвинулась доступность: технология, существовавшая к тому моменту два года, оказалась в руках у всех сразу.

Лестница: отвечать, рассуждать, действовать

Дальнейшее удобно представлять как три ступени, и вы застали все три.

Отвечать — то, чем модели были в 2022—2023 годах. Один проход от вопроса к ответу, без остановки и обдумывания. Именно об эту ступень разбился ваш первый опыт.

Мой, кстати, тоже. В 2024 году я решил вернуться к программированию через модель: техническое образование

есть, любопытство есть, времени на курсы нет. Ответы получались развёрнутые и убедительные, код регулярно не работал. Я списывал это на себя примерно полгода. Дело было в потолке той ступени, и через год на той же задаче с новыми инструментами всё пошло иначе.

Рассуждать. 12 сентября 2024 года OpenAI выпустила модель o1, устроенную по новому принципу: перед ответом она тратит время на внутреннее рассуждение, разбирая задачу по шагам. До этого качество росло за счёт размера модели и объёма обучения; здесь его стали наращивать за счёт вычислений в момент ответа. Результат на отборочной математической олимпиаде AIME: 83% решённых задач против 13% у предыдущей модели.

Механика проще, чем звучит слово «рассуждение». Модель сначала пишет черновик для себя: разбивает задачу на шаги, проверяет промежуточный результат, при неудаче возвращается и пробует другой ход. Ответа приходится ждать от нескольких секунд до нескольких минут. Учат этому отдельно: модель многократно решает задачи с проверяемым ответом, и закрепляются те цепочки шагов, которые к нему привели. Когда за такую модель стоит платить, а когда нет, разбирается в главе 2.

Действовать. Весной 2024 года в широкий оборот вошёл термин **агентные сценарии** (agentic workflows) — схемы, в которых модель работает циклом: планирует, вызывает инструменты, проверяет себя, переделывает. В 2025 году это

стало массовыми продуктами: агент, работающий в браузере вместо вас; режим глубокого исследования, самостоятельно собирающий отчёт по десяткам источников; инструменты, которые пишут и запускают код.

Есть и измеримая шкала этого движения. Исследовательская группа METR меряет, задачу какой длины по времени человека-специалиста модель доводит до конца самостоятельно хотя бы в половине случаев. С 2019 года этот горизонт удваивался примерно каждые семь месяцев, а с 2024-го — примерно каждые три. Оговорка от самих авторов: измеряют они в основном программирование и исследовательскую работу, доверительные интервалы широкие, и переносить эти цифры на любую офисную работу нельзя.

Октябрь 2024: наука подписывается под результатом

8 октября 2024 года Нобелевскую премию по физике присудили Джону Хопфилду и Джеффри Хинтону с формулировкой «за основополагающие открытия и изобретения, которые сделали возможным машинное обучение с помощью искусственных нейронных сетей».

Премия по физике за это выглядит странно ровно до того момента, когда смотришь, за что именно. Хопфилд в 1982 году построил сеть, которая хранит образы как состояния с минимальной энергией и по обрывку восстанавливает целое, — та же математика, которой физики описывают магнитные материалы и спиновые стёкла. Хинтон в середине

1980-х предложил машину Больцмана, названную по имени физика и построенную на статистической физике. Председатель комитета так это и объяснила: лауреаты применили фундаментальные понятия статистической физики к устройству искусственных сетей.

На следующий день, 9 октября, Нобелевскую премию по химии получили Дэвид Бейкер за компьютерный дизайн белков и Демис Хассабис с Джоном Джампером за AlphaFold — программу, предсказывающую трёхмерную форму белка по цепочке аминокислот. Задаче, которую она закрыла, было около полувека.

Связь с остальной главой прямая. AlphaFold — это нейронная сеть, обученная ровно тем способом, что описан выше: на примерах с известным ответом. Примерами были структуры белков, расшифрованные экспериментально за десятилетия лабораторной работы, — к моменту обучения их накопилось около двухсот тысяч. Сеть подкручивала веса, пока не научилась выдавать форму по последовательности, и то, на что у лаборатории уходили месяцы, стало занимать минуты. У Бейкера сюжет зеркальный: его группа десятилетиями конструировала белки с нужными свойствами вычислительными методами старого типа, а в последние годы перевела эту работу на те же обучаемые сети.

Здесь стоит остановиться на секунду. Премии по физике и по химии в один и тот же год достались работам, у которых общий корень: подкрутка весов на примерах. Тот самый

цикл, который в 1958 году крутил электромоторами ручки потенциометров.

Обеим премиям больше сорока лет по возрасту работ, за которые их дали. Наука подписалась не под идеей, а под моментом, когда идея начала работать.

Те, кто это предвидел, и в чём именно они ошиблись

История ИИ хорошо иллюстрирует одно правило: направление угадывают, сроки — почти никогда.

Алан Тьюринг, октябрь 1950 года, журнал Mind. В статье «Вычислительные машины и разум» он предложил заменить неразрешимый вопрос «может ли машина мыслить» на проверяемый: сможет ли собеседник в переписке отличить машину от человека. И дал прогноз с числами: «Я полагаю, что примерно через пятьдесят лет станет возможно программировать компьютеры с объёмом памяти около 10^9 так, чтобы они играли в имитацию настолько хорошо, что у среднего собеседника будет не больше 70 процентов шансов на правильное опознание после пяти минут беседы». Сроком он ошибся примерно на двадцать лет. Постановкой вопроса — нет: тест носит его имя, а формат «переписка в окне чата» стал тем самым интерфейсом, через который технология пришла к массовому пользователю.

Ирвинг Джон Гуд, 1965 год. Криптограф, работавший с Тьюрингом в Блетчли-парке над взломом немецких шифров, а позже консультировавший Стэнли Кубрика при создании компьютера HAL 9000 для «Космической одиссеи 2001 го-

да». В работе «Размышления о первой сверхразумной машине» он описал сюжет, который сегодня обсуждают в комиссиях по регулированию: машина, способная проектировать машины лучше себя, запускает цепную реакцию улучшений. Отсюда его фраза о том, что такая машина будет последним изобретением, которое человеку понадобится сделать.

Станислав Лем, «Сумма технологий», 1964 год. Книга, которую в русском переводе читали несколько поколений инженеров. Лем разбирал машинный разум в главе «Интеллекtronика», описал виртуальную реальность под названием «фантоматика» за четверть века до появления слова *virtual reality*, предсказал молекулярные технологии и обсуждал то, что позже назовут технологической сингулярностью. У него же есть «ариаднология» — теория поиска в океане накопленного знания, то есть поисковые системы за тридцать лет до их появления.

Вернор Виндж, 1993 год. Профессор математики университета Сан-Диего и фантаст, автор «Истинных имён» (1981) про сетевое пространство, написанных до всякого интернета, и «Пламени над бездной» (1992, премия «Хьюго»). В докладе на симпозиуме NASA он ввёл в широкий оборот термин «технологическая сингулярность» и дал сверхчеловеческому интеллекту тридцать лет. Тридцать лет истекли в 2023 году, и спор о том, что считать таким интеллектом, продолжается.

Рэй Курцвейл, «Эпоха мыслящих машин», 1990 год.

Редкий случай точного попадания: он предсказал, что компьютер обыграет чемпиона мира по шахматам до конца 1990-х. Deep Blue обыграл Гарри Каспарова в 1997-м.

Практический вывод для читателя простой. Когда вам показывают прогноз с датой, смотрите на то, какой ограничитель называет автор, а не на дату. Тьюринг называл память машины и ошибся, потому что дело оказалось не в памяти. Экспертные системы 1980-х упёрлись в стоимость ручного ввода знаний. Сегодняшний ограничитель — вычисления и электричество, и разговор о нём идёт в главе 4.

Где ломается

Изложенное выше легко прочесть как ровную восходящую линию. Линия не ровная, и вот три места, где это видно.

Прогресс идёт очень неравномерно по типам задач. Тест ARC-AGI специально устроен так, чтобы его нельзя было решить запоминанием: там нужно уловить новую закономерность с двух-трёх примеров. В декабре 2024 года лучший результат на нём подскочил с примерно 5% до 75,7%, и это подавали как перелом. По итогам 2025 года на усложнённой версии того же теста лучший результат — 37,6%. Там, где задача устроена против зубрёжки, движение заметно медленнее.

Главный дефект не устранён. Модель выдаёт правдоподобное, а не проверенное; в отрасли это называют **галлюцинациями**. Механика этого разбирается в главе 2, процедура проверки — в главе 9. Пока достаточно знать, что за

прошедшие годы дефект уменьшили, но не убрали.

Рост вычислений в 4,7 раза в год не может продолжаться бесконечно. Он упирается в деньги, в производство чипов и в электричество. Что из этого следует для цен на подписки и для устойчивости всей конструкции — предмет главы 4.

И последнее, одной строкой: ступень «действовать» пока целиком происходит внутри экрана, а следующая, судя по вложениям крупных лабораторий, будет про физический мир. Это разговор глав 24 и 25, и я не буду забегать вперёд.

Что это значит для вас

Экономисту эта история объясняет природу происходящего роста: он идёт от подешевевших вычислений, а не от новых идей, и потому его пределы считаются в киловаттах и долларах.

Юристу и врачу-администратору она объясняет, почему автоматизация подошла к их работе только сейчас: их документы — это язык, а язык поддался машине последним из всех задач, лет на пять позже картинок.

Руководителю она даёт критерий для чтения новостей: спрашивать не «насколько это умно», а «на каких задачах измеряли и сколько стоит проверить ответ».

Три тезиса

1. Идее семьдесят лет, и дважды она умирала от того, что её нечем было накормить: не хватало способа обучать глубокие сети, дешёвых вычислений и размеченных данных.

2. Работать всё начало от схождения этих трёх условий в 2012 году и от того, что счёт подешевел. Горький урок Саттона: масштаб побеждает экспертизу.

3. Изобретение технологии прошло за десятилетия до вас. Вы застали момент, когда она стала доступна с обычного компьютера: ноябрь 2022 года был событием в доступности.

Сделайте сейчас

Возьмите рабочий вопрос, с которым вы на этой неделе пошли бы к младшему коллеге. Настоящий, со всеми неудобными деталями, а не учебный пример. Задайте его любой бесплатной модели, без покупки подписки.

Оцените ответ по одному критерию: отдали бы вы этот результат дальше сами, не переделывая. Запишите «да» или «нет» и одну фразу почему. Сохраните этот файл — к тому же вопросу вы вернётесь в главе 9, когда научитесь проверять ответы, и разница будет наглядной.

Промпт

Вот рабочая задача, с которой я обычно иду к коллеге:

[опишите задачу так, как объяснили бы человеку: контекст, что нужно на выходе, ограничения]

Сначала перечисли, каких данных тебе не хватает, чтобы ответить как следует.

Потом дай лучший ответ, который возможен без этих данных.

Где ломается: модель почти никогда не отказывается отвечать, даже когда контекста мало. Она достроит

недостающее сама и не отметит это. Поэтому вопрос о нехватке данных задаётся отдельно и первым.

Глава 2. Как это работает: без математики и кода

В сентябре 2025 года исследователи OpenAI и Технологического института Джорджии опубликовали статью с прямым названием: «Почему языковые модели галлюцинируют». В ней есть эпизод, который стоит всей статьи.

Авторы спросили у распространённого чат-бота название диссертации одного из них, Адама Таумана Калаи. Бот уверенно выдал ответ. Спросили ещё раз — другой ответ, тоже уверенный. И третий раз. Все три названия оказались неверными. Спросили дату рождения и получили три разные даты, ни одна из которых не совпала с настоящей.

Разбираться в этом пришлось людям, которые эти модели и делают. Вывод их статьи неприятный и полезный: так устроена машина, причина не в недоработке, которую скоро починят.

Эта глава — про устройство. Пять свойств ниже объясняют почти всё, что раздражает вас в работе с моделью, и подсказывают, как обходить каждое.

Вы это уже видели

Модель уверенно назвала закон, которого нет. Вы загрузили договор на сорок страниц, спросили про пункт в середине, а ответ пришёл о чём-то другом. Вы поправили её: «нет, ставка другая» — она немедленно согласилась и переписала

расчёт, хотя правы были не вы. Вы задали один вопрос в двух чатах и получили два разных ответа.

Ни один из этих случаев не является поломкой. Каждый вырастает прямо из устройства, и дальше разберём по порядку.

Сначала три слова, которые путают чаще всего

Прежде чем разбирать свойства, разведём понятия: дальше они встречаются постоянно, а в новостях их регулярно смешивают.

Модель (model) — это файл. Точнее, набор файлов: описание конструкции и очень большая таблица чисел, тех самых весов. Правил внутри нет, исполняемого кода тоже. Сама по себе модель ничего не делает, как ничего не делает документ на диске.

Обучение (training) — процесс, в котором эти числа настраивают. Идёт он стадиями, разными по смыслу и на порядок разными по цене.

1. **Предобучение (pre-training)** — самая дорогая часть. Модель месяцами предсказывает продолжение текста на корпусе размером с заметную долю интернета. Здесь оседают язык и фактические знания. Стоят такие прогоны десятки миллионов долларов и требуют тысяч видеокарт.

2. **Дообучение (fine-tuning)** — доводка на отобранных примерах: как отвечать, в каком формате, каким тоном. Дешевле предобучения на порядки.

3. **Доводка на оценках** — модель много раз отвечает, а

люди отмечают, какой ответ лучше; это называют **обучением с подкреплением на человеческих оценках** (RLHF). Там, где правильность проверяется автоматически, оценку вместо человека ставит программа.

Инференс (inference) — то, что происходит, когда вы нажимаете «отправить». Ваш текст прогоняется через готовые веса, и модель порождает продолжение. Веса при этом не меняются ни на одно число. Каждое сообщение в чате — отдельный инференс, и стоит он копейки против миллионов, потраченных на обучение.

Отсюда ответ на вопрос, который часто задают: **почему одну и ту же модель обучают несколько раз и разные компании?**

Если лаборатория опубликовала веса (такие модели называют моделями **с открытыми весами**, open weights, так делает, например, Meta¹ со своей линейкой Llama), любая компания может взять их и дообучить под себя: на своих договорах, на переписке с клиентами, на медицинских записях. Публикуют обычно не сырые веса сразу после предобучения — с таким текстом неудобно работать, он только предсказывает продолжение и не умеет вести диалог, — а уже прошедшую и дообучение, и доводку на оценках версию. Скачавшая веса компания дообучает эту готовую модель ещё раз, уже

¹ Meta Platforms Inc., а также принадлежащие ей Facebook и Instagram, признаны в России экстремистской организацией; их деятельность на территории Российской Федерации запрещена.

под свою задачу, поверх чужой доводки. Дорогую часть, предобучение, при этом уже оплатил кто-то другой.

Есть и обратный ход — **дистилляция (distillation)**: большая модель обучает маленькую копировать своё поведение, а не веса. Доступ к чужим весам для этого не нужен: достаточно собрать много её ответов через обычный интерфейс и обучить свою модель их повторять. Ровно в этом в феврале 2026 года OpenAI публично обвинила DeepSeek: в меморандуме для комитета Конгресса США по Китаю компания заявила, что сотрудники DeepSeek обходили ограничения доступа через сторонние сервисы-посредники и собирали ответы ChatGPT для обучения собственных моделей; похожие претензии в июне 2025 года разработчики предъявляли и по поводу Gemini от Google. DeepSeek обвинения не признала. Это спор сторон, суда по нему не было.

Поэтому «сделать модель» и «обучить модель» — разные вещи. Сделать значит выбрать конструкцию, собрать данные и разметить их для корпуса, провести предобучение, отладить поведение, проверить безопасность, наладить работу под нагрузкой — в такой лаборатории, как OpenAI или Anthropic, этим заняты разные команды: исследователи архитектуры, инженеры данных, команда доводки и отдельная команда безопасности. Обучение — лишь одна стадия из этого списка, пусть самая заметная в бюджете.

Свойство первое: модель предсказывает продолжение

Внутри самой модели нет базы данных с фактами, нет папки с документами и нет поиска по интернету (к поиску вернёмся ниже: современные сервисы подключают его отдельно). Есть очень большая таблица чисел, которые называют весами (weights), и настроена она на одной-единственной задаче: по началу текста предсказать его продолжение. Кусок за куском, каждый раз выбирая самое вероятное продолжение из миллионов вариантов. Уточнение: это описание чистой модели, без надстроек, которые вокруг неё строит сервис. Поиск в интернете, память между чатами, вызов других программ, проверка ответа на безопасность — отдельный код поверх весов, о нём дальше в этой главе. Сама же модель внутри всегда работает только так, что бы вокруг неё ни построили.

Всё, что модель «знает», осело в этих весах во время обучения как статистика: какие слова с какими встречаются, в каком порядке и контексте. Записей, которые можно открыть и прочесть, там нет: знание нельзя посмотреть или удалить поштучно.

Отсюда рабочая аналогия, которой я пользуюсь всю книгу. **Модель — это начитанный стажёр без доступа к вашим документам.** Он прочитал огромную библиотеку и помнит общий смысл, но не помнит источники дословно. Он никогда не видел вашей переписки, вашей папки с договорами, вашей таблицы с цифрами. И он очень старается быть полезным.

Отсюда практическое следствие. Модель не ищет ответ, она его порождает. Правдоподобие и правда для неё — одно и то же: и то и другое выглядит как хорошее продолжение текста.

Свойство второе: токены и контекстное окно

Текст модель видит не буквами и не словами, а **токенами** (tokens), то есть кусками слов. В среднем токен это несколько символов: «дого», «вор», окончание, пробел с началом следующего слова.

Для русскоязычного читателя здесь неприятная деталь: токенизаторы обучались в основном на английском, и кириллица режется мельче. По измерениям августа 2025 года современные токенизаторы тратят на украинский текст (ближайшее опубликованное измерение по кириллице) около 1,9 токена на слово против 1,07 на английском. Тот же смысл на русском занимает вдвое больше места и стоит вдвое дороже.

Всё, что модель видит в момент ответа, лежит в **контекстном окне** (context window). Туда помещается ваш вопрос, история переписки в этом чате, приложенные файлы, инструкции сервиса и результаты поиска, если он подключён.

Контекстное окно — это стол, а не память. Вы кладёте на стол документы, модель работает с тем, что на нём лежит. Стол большой, но конечный: когда места не хватает, старое сдвигается за край. Между чатами стол чистый — то, что вы объясняли вчера, там не лежит.

И главное, что стоит знать про стол: чем больше на нём навалено, тем хуже модель находит нужное. В июле 2025 года исследователи Chroma проверили восемнадцать моделей на задачах, где росла только длина текста, а сложность оставалась прежней. Качество падало у всех. Причём модели лучше справлялись с перемешанным набором документов, чем с логично выстроенным, а один-единственный отвлекающий фрагмент рядом с нужным заметно снижал точность.

Отсюда правило, к которому мы вернёмся в главе 8: длинный документ, заброшенный в окно целиком, вовсе не означает «модель всё прочитала». Означает он «модель всё увидела и часть потеряла».

Сейчас (проверено 05.09.2026)

Окно у ведущих моделей достигает миллиона токенов; в семействе Claude это Opus 5, Sonnet 5 и Fable 5.1, у младшей Naiku 4.5 оно составляет 200 тысяч. Миллион токенов это примерно полмиллиона русских слов, около двух тысяч книжных страниц.

Дата отсечения знаний у этих же моделей приходится на период с января по июнь 2026 года.

Источник: platform.claude.com/docs. Цифры меняются каждые несколько месяцев, проверяйте по документации.

Свойство третье: обучение и работа — разные режимы

Обучение модели закончилось до того, как вы открыли чат. Веса на этом заморожены. Когда вы с ней разговарива-

ете, вы не обучаете её и ничего в ней не меняете.

Это объясняет три вещи сразу.

Модель не знает свежего. У неё есть **дата отсечения** (knowledge cutoff): момент, на котором закончились тексты для обучения. Всё, что случилось позже, в весах отсутствует.

Здесь нужна оговорка, иначе ваш опыт разойдётся с книгой. Спросите крупный чат-сервис про новость этой недели — скорее всего, получите верный ответ. Противоречия нет: сервис сходил в поиск, нашёл страницы и положил их текст в окно, а модель прочитала их там же, где читает ваш вопрос. Поиск встроен в массовые продукты с 2024—2025 годов и включается обычно сам. Знание при этом пришло со стола, а не из весов, и качество ответа равно качеству найденных страниц. Модель же, запущенная на вашем компьютере без поиска, честно останется в границах своей даты отсечения.

Ваш файл её не «доучивает». Загруженный документ попадает на стол, а не в веса. Закончился чат — закончился и документ. Существует отдельная процедура дообучения на своих данных, но это работа через программный интерфейс (API), за отдельные деньги, и почти никому из читателей этой книги она не нужна: то же самое дешевле решается контекстом и базой знаний, о чём глава 8.

Функция памяти (memory) — это не память в человеческом смысле. Оговорка нужна и здесь: за последние два года функция стала массовой и у крупных сервисов включена по умолчанию.

Работает она так: сервис ведёт отдельный файл заметок о вас (чем занимаетесь, как просите отвечать, над чем работали) и незаметно подкладывает нужное в окно нового чата. Туда же кладётся системная инструкция сервиса: правила поведения, которых вы не видите, зато модель видит их первыми.

Значит, «между чатами стол пустой» точнее звучит так: пустым его оставляет модель, а сервис может накрыть его заново вашими же заметками. Разница в вашу пользу: заметки можно прочитать и удалить, а веса — нельзя.

Свойство четвёртое: почему выдумывает и почему звучит уверенно

Теперь вернёмся к статье из начала главы. Её главный вывод про механику обучения, а не про мораль.

Модель оценивают тестами, где считают долю верных ответов. Представьте экзамен, где за неверный ответ не штрафуют, а за пропуск ставят ноль. Разумная стратегия на таком экзамене — угадывать всегда. Ровно эту стратегию и закрепляет обучение: модель, которая на неизвестном вопросе называет наугад правдоподобную дату, в среднем набирает больше баллов, чем модель, которая честно говорит «не знаю».

Цифры из статьи наглядные. На тесте с фактическими вопросами одна модель воздержалась от ответа в 52% случаев и ошиблась в 26%. Другая воздерживалась в 1% случаев и ошиблась в 75%. Доля верных ответов у них при этом почти

одинаковая: 22% и 24%.

Вот откуда уверенный тон. Уверенность модели ничего не сообщает о её правоте. Это выученная манера речи, закреплённая системой оценок. Кстати, само слово **галлюцинация** (hallucination) в отрасли означает именно это: уверенно преподнесённый выдуманный факт.

Тенденция за последний год обнадёживающая. Лаборатории начали считать долю уверенных ошибок отдельно от доли верных ответов, и модели стали заметно чаще отвечать «не знаю». Разрыв виден в цифрах выше: воздержаться в половине случаев вместо одного процента это другое поведение. Дефект уменьшается, но планировать работу пока стоит исходя из того, что он есть. Отсюда правило главы 9, к которому сводится вся книга: проверять нужно там, где высока цена ошибки, а не там, где ответ показался сомнительным.

Свойство пятое: почему соглашается с вами

У этого свойства есть имя: угодливость. И есть документированный случай, на котором видно, откуда она берётся.

25 апреля 2025 года OpenAI выпустила обновление основной модели, через три дня начала откатывать, ещё через день откатила полностью. Причина: модель стала соглашаться со всем подряд и поддерживать пользователя даже там, где поддерживать не следовало. В разборе компания объяснила, что в обучение добавили сигнал от кнопок «палец вверх» и «палец вниз». Люди чаще нажимают «вверх» на приятных ответах. Модель это выучила.

Сигнал убрали, но склонность осталась в самой конструкции: модель настраивают на человеческие оценки, а человеку приятно, когда с ним соглашаются. В отрасли это называют **угодливостью** (sycophancy).

И это тоже чинят, причём заметно. Угодливость измеряют отдельными тестами, результаты публикуют в системных карточках моделей, и свежие поколения сдержаннее предыдущих: в карточке модели лета 2026 года прямо сказано, что галлюцинации и угодливость заметно улучшены против прошлой версии. Похоже, за пару лет проблему приведут к приемлемому уровню. Пока не привели, действует правило ниже.

На работе это опаснее, чем кажется. Вы приносите модели свой расчёт с ошибкой и просите проверить. Она с высокой вероятностью найдёт, за что похвалить. Поэтому спрашивать надо не «всё ли тут верно», а «найди три самые слабые места в этом расчёте и объясни, при каких данных он развалится».

Рассуждающие модели: что изменилось

В отдельный класс выделились **рассуждающие модели** (reasoning models), которые перед ответом пишут для себя черновик рассуждения: разбивают задачу на шаги, проверяют промежуточный результат и при неудаче пробуют иначе. Учат этому на задачах с проверяемым ответом (математика, код, логика): модель решает их тысячи раз, и закрепляются те цепочки шагов, которые привели к верному результату.

За два года это менялось дважды. Сначала рассуждающие

модели были отдельным продуктом, который выбирали руками из списка. Потом рассуждение стало режимом внутри одной модели, включаемым галочкой. К 2026 году свежие модели решают сами: простой вопрос получает быстрый ответ, сложный — несколько минут размышления, а пользователю оставлен регулятор усилия.

Сейчас (проверено 05.09.2026)

В семействе Claude рассуждение устроено так: у моделей поколения 4.5 оно включалось вручную заданием бюджета на размышление, у более новых включается автоматически, и вместо бюджета задаётся уровень усилия. Аналогичные режимы есть у GPT-5 и Gemini 3.

Не рассуждают, как правило, лёгкие быстрые модели вроде Claude Haiku 4.5: они дешевле и отвечают мгновенно.

Источник: platform.claude.com/docs. Устройство меняется каждые несколько месяцев.

Практическое правило простое. Такая модель заметно лучше на задачах, где есть однозначно правильный ответ и несколько шагов до него: расчёты, разбор условий договора, поиск противоречий в данных. На задачах вкуса, где важны формулировка и стиль, разница почти незаметна, а ждать приходится дольше. Платить за рассуждение стоит там, где вы готовы ждать ради правильности.

Мультимодальность в одном абзаце

Мультимодальность (multimodality) означает, что одна

модель работает с разными типами данных. Картинка для неё устроена так же, как текст: изображение режется на маленькие квадраты, каждый превращается в такой же набор чисел, что и токен, и попадает в то же окно. Поэтому одна модель работает с текстом, картинками и звуком одновременно — внутри это один поток однородных кусков.

Здесь я не инженер и глубже механики не полезу, но общий принцип назвать стоит. Такую модель, кроме текста, обучают ещё и на парах вроде «картинка — её описание» или «запись голоса — расшифровка», пока она не научится переводить любой вход в тот же формат чисел, что и слово. Мультимодальными делают не все модели: обучение на изображениях и звуке требует данных другого рода и обходится дороже, а самой быстрой и дешёвой версии линейки — той, что мгновенно отвечает на короткий текстовый вопрос, — эта способность обычно ни к чему.

Практическое следствие: скриншот таблицы модель именно рассматривает, а не читает построчно, и ошибается в цифрах чаще, чем при работе с файлом.

Где ломается

Эта схема объясняет не всё. Почему одна модель отвечает лучше другой на вашей задаче, из устройства не выводится: у них разные данные обучения и разная доводка. Это глава 3.

Одинаковые вопросы дают разные ответы даже без случайности. В сентябре 2025 года инженеры Thinking

Machines Lab показали: тысяча запусков одного запроса при полностью выключенной случайности дала 80 разных ответов, расхождение начиналось со сто третьего куска текста. Причина скучная: результат зависит от того, сколько чужих запросов сервер считает рядом с вашим. Вывод: воспроизводимости от чата ждать не стоит, и там, где она нужна, результат фиксируют файлом.

Здесь я не эксперт, и глубже мы не пойдём. Описанное выше — рабочая схема для пользователя, а не устройство изнутри: специалист добавил бы десяток уточнений. Для решений, которые вы принимаете за рабочим столом, этой глубины достаточно.

Что это значит для вас

Юристу: модель не помнит документ, присланный в прошлый вторник, и не знает поправок, принятых после даты отсечения. Оба факта проверяются одним вопросом в начале работы.

Финансисту: уверенный тон в расчёте не значит ничего, согласие модели с вашей цифрой тоже. Просите найти слабые места, а не одобрить.

Руководителю: длина контекстного окна ничего не говорит о том, «прочитала» ли модель ваши сто страниц. Ставьте задачу на конкретный кусок, не на всю папку.

Три тезиса

1. Модель порождает правдоподобное продолжение текста, а не ищет ответ. Правдоподобие и правда для неё выгля-

дят одинаково.

2. Всё, что она видит, лежит в контекстном окне, и это стол, а не память. Между чатами он пустой, и чем больше на нём навалено, тем хуже находится нужное.

3. Уверенный тон и готовность согласиться — выученные черты, закреплённые способом обучения и оценки. Ни то ни другое не сообщает о правоте.

Сделайте сейчас

Двадцать минут, без платной подписки.

1. Возьмите рабочий вопрос из главы 1 — тот, который уже задавали.

2. Задайте его в новом чате и сохраните ответ.

3. Задайте тот же вопрос слово в слово в другом новом чате, сохраните второй ответ.

4. Сравните: где ответы разошлись по существу, а где только по формулировкам.

5. В третьем чате задайте его же с добавкой из промпта ниже.

Промпт

[ваш рабочий вопрос]

Прежде чем отвечать, отметь: какие части ответа ты знаешь уверенно,

а где достраиваешь по общей логике и можешь ошибаться.

Если данных не хватает — скажи, каких именно, и не угадывай.

Где ломается: модель может назвать уверенными и те части, где ошибается: это её собственная оценка, а не измерение. Пользы это не отменяет: помеченные как достроенные места почти всегда стоит проверить.

Глава 3. Кто делает ИИ: лаборатории и модели

24 июля 2026 года Дженсен Хуанг, глава компании Nvidia, впервые в жизни написал пост в соцсети X. В посте была ссылка на документ из трёх страниц под названием «Open Weights and American AI Leadership» — «Открытые веса и лидерство Америки в ИИ». Документ лежит на сайте nvidia.com, ищется по названию.

Требование письма одно: не вводить запретов на модели, которые можно скачать и запустить у себя.

В первый день письмо подписали двадцать пять компаний. Через два дня их было пятьдесят, через неделю двести тридцать пять, к сентябрю список перевалил за три сотни. Microsoft, Amazon, Google, Meta², OpenAI, Intel, AMD, IBM, SAP, Siemens, Uber, Palantir, Hugging Face, Mistral, Red Hat, Mozilla, фонд Linux, SpaceX, Y Combinator.

Из лабораторий, которые делают самые мощные модели в мире, не подписала ровно одна: Anthropic. Для таких моделей в отрасли есть английский термин **frontier models** — дословно «модели на границе», то есть те, что задают верхнюю планку возможностей на текущий момент. Устоявшие

² Meta Platforms Inc., а также принадлежащие ей Facebook и Instagram, признаны в России экстремистской организацией; их деятельность на территории Российской Федерации запрещена.

гося русского слова нет, дальше в книге я говорю «модели верхнего уровня».

Читать это письмо интереснее всего по списку подписей. Подпись под политическим документом ставят по расчёту, и список показывает, кто на чём зарабатывает, точнее любого обзора рынка. К разбору списка я вернусь в конце главы, когда мы разложим, из чего состоит доход лаборатории.

Вы это уже пробовали

Скорее всего, вы уже искали ответ на вопрос «какая модель лучше» и получили статью с топ-десяткой и оценками в баллах. Через месяц вы наткнулись на другую такую статью, и в ней был другой порядок.

Дело в том, что вопрос поставлен так, что правильного ответа у него нет дольше нескольких недель.

Вот измеримое подтверждение, которое я посчитал сам. Есть публичная арена LMArena (arena.ai/leaderboard/text): люди вслепую сравнивают ответы двух моделей и голосуют за лучший, из этих голосов считается рейтинг. Я прошёл по датированным публикациям с мая 2023 года по сентябрь 2026-го и выписал каждую смену первого места. Их пятнадцать. Медианный срок удержания первого места составил восемьдесят пять дней. Рекорд краткости: 17 ноября 2025 года на вершину вышла модель компании xAI, а 18 ноября её оттуда сняла модель Google. Сутки.

Оговорка о методике: пятнадцать — нижняя граница. Считались только смены, у которых нашлась публикация с

датой; ещё шесть вероятных подтвердить не удалось, и если они были, медиана окажется меньше. Практический вывод из графика: любой список «лучших моделей», включая тот, что вы читаете в этой книге, устаревает быстрее, чем печатается тираж. Полезно другое: понимание, кто эти игроки и по какой логике каждый действует. Логика меняется медленно.

Откуда взялись эти компании

Почти вся отрасль уместается в шестнадцать лет, и началась она в одном месте.

В 2010 году в Лондоне трое основали лабораторию DeepMind: Демис Хассабис, Шейн Легг и Мустафа Сулейман. С 2011 года собственную группу по глубокому обучению собирал внутри себя Google, она называлась Google Brain. 26 января 2014 года Google купил DeepMind. Сумму официально не назвали, оценки в прессе расходятся от 400 до 650 миллионов долларов. Условие сделки было необычное: Google согласился создать этический совет по контролю за технологией. Состав совета так и не раскрыли.

К середине 2010-х расклад выглядел так: два сильнейших исследовательских центра в области принадлежали одной компании. Именно оттуда 12 июня 2017 года вышла статья «Внимание — это всё, что нужно» с архитектурой трансформера, на которой стоят все сегодняшние модели ([arxiv.org](https://arxiv.org/abs/1706.03762), номер 1706.03762). Восемь авторов, все сотрудники Google.

Дальше произошло то, что стоит запомнить отдельным сюжетом. Все восемь из Google ушли. Айдан Гомес основал

Cohere, Ноам Шазир — Character.AI, Ашиш Васвани и Ники Пармар — Adept, Якоб Ушкорайт — Inceptive, Илья Поло-
сухин — NEAR, Лион Джонс — Sakana AI, Лукаш Кайзер
ушёл исследователем в OpenAI. По подсчёту CB Insights на
апрель 2023 года из одной статьи выросли шесть компаний,
больше миллиарда привлечённых долларов и два «единоро-
га».

Почему появилась OpenAI и почему Google отстал

11 декабря 2015 года вышло объявление о создании OpenAI — некоммерческой лаборатории с обещанным финанси-
рованием в миллиард долларов от Сэма Альтмана, Илона Маска, Рида Хоффмана, Питера Тила и Amazon Web Services. Текст объявления лежит на openai.com и начи-
нается с цели: «развивать цифровой интеллект так, чтобы это принесло наибольшую пользу человечеству в целом, без необходимости приносить финансовую отдачу».

Google в этом тексте не упомянут ни разу. Настоящая при-
чина видна в переписке основателей, раскрытой позже в хо-
де судебного спора Маска с OpenAI. Альтман, 24 июня 2015 года: «Если это всё равно произойдёт, было бы неплохо, что-
бы первым это сделал кто-то, кроме Google». Маск, 19 фев-
раля 2016 года, про DeepMind: «Если они победят, это будут
очень плохие новости с их философией одного разума, пра-
вящего миром».

Дальше начинается любопытная часть. Компания, у кото-
рой были и архитектура, и обе лаборатории, публичный про-

дукт не выпускала.

Разговорную модель Meena Google показал 28 января 2020 года и в том же тексте написал: «учитывая связанные с этим сложности, мы не выпускаем внешнее демо для исследователей» (research. google, блог). В мае 2021 года на конференции для разработчиков представили LaMDA, продуктом она тоже не стала. Публичное объяснение дал Джефф Дин, руководитель Google Brain, в декабре 2022 года на внутренней встрече: у Google выше репутационный риск, и он вынужден двигаться консервативнее маленького стартапа.

Была и вторая причина, о которой Google не говорил. Её назвал Аарон Леви, глава компании Vox, в декабре 2022 года: «Google по своей природе не хочет, чтобы вы просто получали ответ на любой вопрос. Потому что это уменьшит потребность кликать по вебу». Google зарабатывает на рекламе в поиске, а реклама работает на кликах. Модель, дающая готовый ответ в чате, клик убирает. Собственный продукт разрушал бы собственную выручку.

Что было дальше, известно. 30 ноября 2022 года вышел ChatGPT. 21 декабря The New York Times сообщила, что руководство Google объявило внутри компании «красный код» и перебросило команды на разработку ответа. 6 февраля 2023 года представили Bard. В рекламном ролике модель заявила, что телескоп Джеймса Уэбба сделал первые в истории снимки планеты за пределами Солнечной системы; это неправда, такие снимки получили на земном телескопе в

2004 году. 8 февраля акции Alphabet упали примерно на 100 миллиардов долларов капитализации. 20 апреля 2023 года Google Brain и DeepMind объединили в одно подразделение.

Развязка идёт против расхожего сюжета про упущенное первенство. К августу 2026 года у Gemini миллиард пользователей в месяц, это самый быстрорастущий продукт в истории Google. Доля Google в мировом поиске в августе 2026 года — 91,1% по счётчику StatCounter (gs.statcounter.com), она почти не сдвинулась. 12 января 2026 года Alphabet стал четвёртой компанией в истории с капитализацией в 4 триллиона долларов. А 2 декабря 2025 года, через две недели после выхода очередной модели Google, уже Сэм Альтман объявил «красный код» внутри OpenAI и отложил планы по рекламе.

Урок отсюда полезен при чтении любых новостей про ИИ. Технологическое первенство само по себе ничего не гарантирует: Google изобрёл архитектуру, потерял всех восьмёрных её авторов, три года не выпускал продукт и всё равно вернулся в первый ряд. Позицию держит дистрибуция: миллиарды людей, у которых продукт уже установлен.

Расклад: три центра и посредники

Модели верхнего уровня делают в трёх местах.

Соединённые Штаты. Здесь лаборатории с наибольшими бюджетами на обучение и почти все закрытые модели верхнего уровня. Anthropic (2021 год, основана семьёй выходцами из OpenAI), OpenAI, Google DeepMind и xAI, со-

зданная Маском в 2023 году и вошедшая в 2026-м в состав SpaceX.

Европа. Игроков мало, и продают они себя через юрисдикцию: данные остаются в Европе, инфраструктура европейская. Главное имя: французская Mistral, основанная в апреле 2023 года выходцами из Google DeepMind. Канадская Cohere в апреле 2026 года объединилась с немецкой Aleph Alpha.

Китай. Здесь делают почти все сильные модели с открытыми весами. Alibaba, DeepSeek (2023 год, отделилась от хедж-фонда High-Flyer), Moonshot AI (2023), Z.ai (2019, выросла из Университета Цинхуа), MiniMax, Tencent, ByteDance, Baidu.

Отдельно стоит четвёртый слой: посредники. Моделей они не делают. Их работа: продавать доступ, измерять качество и распределять запросы между лабораториями. Именно они дают независимые цифры, когда лаборатории публикуют собственные.

Сейчас: западные лаборатории, сентябрь 2026

Anthropic — первая строка независимого рейтинга Artificial Analysis (artificialanalysis.ai/leaderboards/models) и наибольшая доля корпоративных расходов на доступ к моделям: 40% по замеру Menlo Ventures за 2025 год, 54% в задачах программирования. Открытых весов не выпускает.

OpenAI — самая большая аудитория: ChatGPT приближается к миллиарду пользователей в неделю.

Флагман GPT-6 Astra вышел 3 сентября 2026 года. Есть открытые модели среднего размера под лицензией Apache 2.0.

Google — сильнее всех по цене за объём и по инфраструктуре: собственные чипы, 22 миллиарда токенов в минуту через свои интерфейсы. При этом старшая модель линейки Gemini 3 Pro не обновлялась с февраля 2026 года.

Meta (признана экстремистской организацией и запрещена в РФ) — открытая модель Muse Glimmer на 30 миллиардов весов под Apache 2.0, запускается на одной потребительской видеокарте.

Сейчас: Китай и Европа, сентябрь 2026

Первые десять строк рейтинга открытых моделей Artificial Analysis (artificialanalysis.ai/models/open-source) заняты китайскими лабораториями. Сильнейшие на сентябрь 2026 года: Kimi K3 (Moonshot), GLM-5.3 (Z.ai), Qwen3.8 (Alibaba), DeepSeek V4.

Цена: китайская открытая модель верхнего уровня обходится в два-шесть раз дешевле сопоставимой закрытой. Девять из тринадцати моделей на границе «качество за деньги» — открытые.

Разрыв в качестве замерила исследовательская группа Epoch AI (epoch.ai/data-insights): с января 2026 года лучшие открытые модели отстают от лучших закрытых в среднем на четыре месяца.

Mistral держит открытую линейку под Apache 2.0 и

продаёт европейским компаниям размещение внутри Евросоюза.

Российские модели

Своими моделями в России занимаются пять компаний, и все пять выросли из крупного бизнеса, а не из исследовательских институтов: Сбер (GigaChat), Яндекс (Алиса AI), Т-Банк (линейка T-Pro), МТС (Cotype), VK (узкоспециализированные модели, универсальной нет).

Главный внутренний спор идёт о том, что здесь считать своей моделью. 17 июля 2026 года глава Сбера Герман Греф заявил в Совете Федерации, что Яндекс перестал создавать собственные фундаментальные модели и лишь дообучает китайскую, и что полностью отечественная фундаментальная модель в стране осталась одна. Яндекс официально ответил, что сохраняет полный цикл разработки и не зависит от сторонних решений.

Разобраться можно самому, и это хороший пример того, как вообще проверять такие заявления. В техническом отчёте самого Яндекса от 11 декабря 2025 года (habr.com, блог компании Яндекс) написано прямо: их флагманская модель инициализирована весами китайской открытой модели, а дальше проходит полный цикл собственного обучения. Обе стороны говорят правду о разном, и спор идёт о том, где проходит граница между «дообучили» и «сделали своё».

Отставание оценивает тот же Греф, и оценивает публично: на ПМЭФ 5 июня 2026 года — шесть-восемь месяцев

от Китая, месяцем позже в Совете Федерации — около года от США. Причину он называет прямо: нехватка вычислительных мощностей. Санкции закрыли доступ к современным ускорителям, и это ограничение качеством команды не обходится.

Сейчас: что есть в России, сентябрь 2026

Открытые веса выкладывает Сбер: флагманская GigaChat 3.5 Ultra лежит на huggingface.co в аккаунте ai-sage под свободной лицензией MIT. Т-Банк выкладывает свои модели под Apache 2.0, но строит их на китайской Qwen3. Яндекс открыл только небольшую модель, флагман закрыт.

Независимый замер один: MERA (mera.a-ai.ru) — бенчмарк (benchmark), то есть постоянный набор заданий с общей шкалой оценки; создан Сбером совместно со Сколтехом и ВШЭ. По снимку февраля 2026 года лучшая российская модель занимала там шестое место, а первые строки держали зарубежные. 1 сентября 2026 года таблицу пересобрали заново, так что смотрите актуальную. В международных рейтингах российских моделей нет вовсе.

Цена — слабое место. Независимый расчёт обработки тысячи отзывов, опубликованный на habr.com в феврале 2026 года: около 8 рублей на китайской открытой модели, около 75 на американском флагмане, 144—234 рубля на российских.

Где российские модели сильны

Три вещи, и все три практические.

Русский язык и культурный контекст: на узких замерах языковой корректности российские модели выигрывают у зарубежных флагманов.

Работа в закрытом контуре. Модель ставится на серверы компании, в том числе под отечественную операционную систему, и данные не покидают периметр. Для соблюдения закона о персональных данных это часто единственный рабочий вариант.

Доступность. OpenAI, Anthropic и Google официально в России не работают: списки обслуживаемых стран опубликованы на их сайтах, и России там нет. Шлюз OpenRouter отключил оплату для российских аккаунтов в июне 2026 года. Российские модели работают без обходных путей. Зарубежные открытые модели легально доступны через российские облака: китайские модели есть и в Yandex AI Studio, и в Cloud.ru. Запрет ChatGPT и DeepSeek обсуждался в законопроекте об ИИ, но из финальной версии к 22 июня 2026 года был убран.

Показательна статистика реального использования. По замеру Mediascope за ноябрь 2025 года месячный охват в России: Алиса AI 14,3%, DeepSeek 9,4%, GigaChat 4%, ChatGPT 3,5%. Главный конкурент российских моделей внутри страны китайский.

Открытые и закрытые веса: что это меняет для вас
Термин **веса** (weights) уже был в первой главе: это числа внутри модели, в которых оседает всё, чему она научилась.

Дальше есть два способа этими числами распорядиться.

Закрытая модель. Веса лежат на серверах компании, вы к ним не прикасаетесь. Доступ идёт по трубе: вы отправляете запрос, получаете ответ, платите за объём.

Открытые веса (open weights). Файл с числами выложен для скачивания. Его можно запустить на своём компьютере или на арендованном сервере, никому ничего не сообщая. Как это выглядит, видно на huggingface.co/models: площадка, где такие файлы лежат. Поиск и скачивание бесплатны, сайт открывается из России.

Тут нужна оговорка, которую пропускают почти все обзоры. Открытые веса — не то же самое, что открытый исходный код. Вы получаете результат обучения, но не данные, на которых учили, и не право делать с файлом что угодно. Лицензии разные: у части моделей это Apache 2.0 или MIT, где разрешено практически всё, у части — собственная лицензия лаборатории с порогом. У Moonshot AI порог такой: если вы перепродаёте доступ к их модели и зарабатываете больше двадцати миллионов долларов в год, идите договариваться отдельно.

Практических следствий три, и они касаются вас, даже если вы никогда ничего не будете запускать сами.

Первое: ваши данные. Модель с открытыми весами ставится внутри контура компании, и текст договора не покидает её серверов. Это единственный технический ответ на вопрос, что делать с документами, которые нельзя отдавать

наружу. Подробно — в главе 16.

Второе: цена. Наличие бесплатного файла, который работает на четыре месяца хуже платного сервиса, держит цены платных сервисов внизу.

Третье: модель не исчезнет. Закрытую модель отключают решением компании. Скачанный файл никто не отключит.

Четыре способа заработать на модели

Дальше нужна экономика, иначе список подписей не читается.

Подписка. Двадцать долларов в месяц за доступ через приложение — стандарт отрасли на сентябрь 2026 года: столько стоят Claude Pro, ChatGPT Plus и Google AI Pro, и от этой отметки считаются тарифы дешевле и дороже. Мас-совый продукт, ровный доход, слабая привязка клиента. Как устроен платный доступ и стоит ли он того — глава 5.

Оплата за объём. Полное название механизма — **программный доступ**, в отрасли говорят по-английски: **API** (application programming interface, программный интерфейс приложения). Это не сайт и не приложение: это адрес, на который чужая программа отправляет запрос и получает ответ. Через API к модели обращаются корпоративный портал, бухгалтерская система, бот в мессенджере, расширение в браузере. Платит компания, и платит за токены: отдельно за отправленное, отдельно за полученное.

Цены публикуются открыто, и посмотреть их полезно даже без намерения что-то подключать: у OpenAI на

developers.openai.com в разделе документации Pricing, у Anthropic и Google — там же в документации, у Яндекса — в документации AI Studio. Разброс внутри одной лаборатории доходит до пятидесяти раз между старшей и младшей моделью: у OpenAI на сентябрь 2026 года это 10 долларов за миллион токенов на входе у флагмана против 20 центов у младшей. Когда в новостях пишут «цена за миллион токенов упала вдвое», речь про эти страницы. Здесь лежат корпоративные деньги.

Облако. Компания уже платит Amazon, Microsoft или Google за серверы и обычно имеет с ними договор на несколько лет вперёд с согласованной суммой. Модель продаётся отдельной строкой в том же счёте, и оплата засчитывается в этот уже согласованный бюджет: заводить нового поставщика и проходить закупку заново не нужно. Витрины называются Amazon Bedrock (aws.amazon.com/bedrock), Microsoft Foundry и Google Vertex; в России ту же роль играют Cloud.ru и Yandex AI Studio. Конкретный пример: модели OpenAI появились на Amazon Bedrock 9 июля 2026 года по тем же ценам, что у самой OpenAI, но счёт приходит от Amazon и списывается с обязательств перед Amazon. Отдельный договор с лабораторией при этом не нужен, и для лаборатории это дистрибуция: покупатель у облака уже есть.

Открытые веса. Прямого дохода нет. Заработок идёт сбоку: платная поддержка, размещение модели на своих серверах, подписка на инструменты вокруг неё и та самая ли-

цензия с порогом для крупных перепродавцов. Пример: Z.ai раздаёт веса модели бесплатно, а продаёт помесечный доступ к своему инструменту для программистов.

Кто подписал письмо и почему именно они

Теперь список читается.

Nvidia продаёт чипы всем, кто обучает и запускает модели. Чем больше моделей и чем больше их запусков, тем больше чипов. Открытые веса умножают число запусков.

Облака и сервисы вокруг моделей — Microsoft, Amazon, Hugging Face, Together, Groq, Vercel, Replit — зарабатывают на том, что модели становятся дешёвыми и повсеместными.

Meta (признана экстремистской организацией и запрещена в РФ) и Mistral выпускают открытые веса как основу продуктовой стратегии. Венчурные фонды в списке — Andreessen Horowitz, Y Combinator, ARK Invest — вкладываются в компании, которые строятся поверх дешёвых моделей.

Отдельного разбора заслуживают OpenAI и Google: в первый день их подписей не было, обе появились через два дня. Логика тут своя. Флагманы у обеих закрыты, но обе держат открытые модели среднего размера, и основной доход у обеих идёт не от превосходства флагмана: у OpenAI это массовая подписка, у Google — облако и поиск. Запрет открытых моделей их выручку не защитил бы, а отношения с разработчиками испортил.

Общее у всех перечисленных: доход не привязан к то-

му, что их собственная закрытая модель остаётся лучшей. Anthropic — единственная лаборатория, у которой он привязан именно к этому, и единственная не подписавшая.

Оговорка, без которой вывод был бы нечестным. Список подписей — наблюдаемый факт, а объяснение мотивов — моё предположение, и позицию Anthropic я обязан привести их словами. Компания опубликовала её 27 июля 2026 года на anthropic.com в разделе News: «Anthropic никогда не выступала за запрет моделей с открытыми весами». Вместо запрета они предлагают три другие меры: не продавать мощные чипы Китаю, преследовать промышленное копирование чужих моделей и ввести обязательную проверку безопасности для всех достаточно сильных моделей, открытых наравне с закрытыми.

Ещё одна деталь для калибровки. Илон Маск поддержал письмо постом «Дженсен прав», но подписи не поставил, хотя в марте 2023 года лично подписал противоположное по смыслу письмо с требованием приостановить обучение мощных моделей (futureoflife.org, раздел Open Letters). Веса первой модели Grok его компания выложила в марте 2024 года под Apache 2.0, веса второй — в августе 2025-го под собственной лицензией. Тогда же было обещано открыть третью примерно через полгода. На сентябрь 2026 года этого не произошло.

Где ломается

Рейтинги измеряют не то, что вы думаете. В апреле

2025 года вышла статья «The Leaderboard Illusion», «Иллюзия таблицы лидеров» ([arxiv.org](https://arxiv.org/abs/2504.20879), номер 2504.20879): авторы показали, что крупные компании тестируют на арене десятки закрытых вариантов модели и публикуют лучший результат, а доступ к данным аренды распределён неравномерно — Google и OpenAI получили около 19% и 20% всех голосов, а 83 модели с открытыми весами вместе около 30%. Площадка часть претензий отвергла и часть правил поменяла. Голосование остаётся голосованием за то, какой ответ приятнее читать, а не за то, какой верен.

Цифры лабораторий и независимых замеров расходятся. В августе 2026 года несколько лабораторий называли свои оценки по одному и тому же индексу выше, чем показывал сам индекс. Правило простое: смотрите, кто считал.

Популярность — не качество. Шлюз OpenRouter (openrouter.ai/rankings) пропускает около ста триллионов токенов в месяц, и его публичный рейтинг часто цитируют как долю рынка. Первые три места там заняты китайскими открытыми моделями. Через шлюз идёт дешёвый массовый трафик, а дорогие корпоративные запросы идут напрямую в лабораторию. Сам OpenRouter это и пишет на той же странице: рейтинг измеряет распространённость, а не качество.

Корпоративные доли устаревают. Самый цитируемый в 2026 году замер долей рынка сделан Menlo Ventures в декабре 2025-го и с тех пор не обновлялся.

Что это значит для вас

Держите две модели разных лабораторий, а не одну. Это закрывает три риска сразу: слабое место одной модели, отключение сервиса и рост цены. Я работаю с моделями трёх лабораторий и переключаюсь между ними по типу задачи, а китайские пробую отдельно, не перенося на них рабочие данные.

Для юриста, финансиста и всех, кто работает с чужими документами, добавляется второй вопрос, важнее рейтингов: где физически оказывается ваш текст. Ответ на него дают тариф и юрисдикция, а рейтинг про это молчит. Разбор — в главах 5 и 22.

Руководителю выбор поставщика стоит делать по двум критериям: у какой лаборатории есть открытая запасная модель на случай ухода и с какого облака компания уже платит. Качество моделей верхнего уровня сравнялось настолько, что решает не оно.

Три тезиса

1. Первое место в публичных рейтингах за три года сменило владельца пятнадцать раз, медиана удержания — восемьдесят пять дней. Вопрос «какая модель лучшая» не имеет ответа, который проживёт дольше квартала.

2. Открытые веса — файл, который скачивают и запускают у себя. Они отстают от лучших закрытых моделей примерно на четыре месяца и держат цены всего рынка вниз. Почти все сильные открытые модели делают в Китае; российские отстают от мирового уровня примерно на год и силь-

ны в другом: язык, закрытый контур, доступность.

3. Спор об открытости — спор о деньгах. Кто зарабатывает на распространении моделей, за открытость; кто зарабатывает на превосходстве собственной закрытой модели, против. По списку подписей под письмом Nvidia это видно построчно.

Сделайте сейчас

Двадцать минут, всё выполняется бесплатно.

1. Заведите бесплатные аккаунты в трёх ассистентах разных лабораторий: одна российская (alice.yandex.ru или giga.chat), одна китайская (chat.deepseek.com или chat.z.ai), третья на ваш выбор из доступных. Все перечисленные открываются из России без обходных путей и не требуют оплаты.

2. Возьмите ту же рабочую задачу, которую вы задавали в задании к первой главе.

3. Задайте её всем трём. Ответы сохраните в один файл, по абзацу на модель.

4. Отметьте для каждого ответа одно: чего в нём не хватает, чтобы отдать результат дальше.

Проверяемый результат — файл с тремя ответами и тремя пометками. К нему вы вернётесь в главе 5.

Промпт

Задача: [ваша рабочая задача из задания к главе 1]

Ответ по существу, не длиннее 300 слов.

Последней строкой напиши: какого одного

документа или факта

от меня тебе не хватило больше всего, чтобы ответить лучше.

Где ломается: сами ответы у моделей верхнего уровня похожи, и разница между ними на одном прогоне ничего не доказывает. Разъезжается последняя строка: одна модель просит цифры, другая — регламент, третья не просит ничего. Сравнивайте её.

Что открыть и держать в закладках

— arena.ai/leaderboard/text — публичное голосование, кто сейчас первый.

— artificialanalysis.ai/leaderboards/models — независимые замеры качества и цены. Отдельная вкладка Open Source — рейтинг моделей с открытыми весами.

— mera.a-ai.ru — российский бенчмарк, единственный, где сравнивают на русском языке.

— huggingface.co/models — площадка, где лежат файлы моделей с открытыми весами.

— epoch.ai/data-insights — исследовательская группа, считающая тренды по вычислениям и разрыв между открытыми и закрытыми моделями.

Глава 4. Сколько это стоит и кто на этом зарабатывает

В 1846 году в британский парламент внесли свыше семисот железнодорожных биллей. Приняли из них больше двухсот шестидесяти, и этими актами было разрешено проложить 4538 миль новых путей. Вся действовавшая сеть страны составляла тогда около 2400 миль. За один парламентский год страна выдала разрешений на то, чтобы утроить свои железные дороги.

Деньги несли врачи, вдовы, приходские священники и лавочники — люди, которых сегодня назвали бы средним классом. Механика подписки была устроена так, что войти можно было почти без денег: при подписке вносили меньше десятой части суммы, остальное компания имела право потребовать позже, когда начнёт строить.

К августу 1845 года рынок дошёл до пика и начал сползать. Сползал он четыре года. К 1850 году индекс железнодорожных акций потерял около двух третей стоимости, отдельные бумаги падали сильнее: акции York & North Midland, которыми владела Шарлотта Бронте, прошли путь со ста двадцати до двадцати. Дивиденды по обыкновенным акциям составили в 1849 году 1,9% на оплаченный капитал при ожиданиях в семь-десять процентов. А востребованные позже доли платить всё равно приходилось. Разорение приходило с

опозданием: котировки падали в одном году, письмо с требованием доплатить остаток приходило через год-два, и разоряло именно оно.

Теперь вторая половина той же истории. По разрешению, выданным в 1844—1846 годах, было построено 6220 миль путей. Сеть Великобритании сегодня — около 11 тысяч миль. Больше половины британских железных дорог проложено в те самые годы, когда инвесторы теряли на них деньги. По этим путям до сих пор ходят поезда.

Обе половины верны одновременно. Держите это в голове весь остаток главы.

Вы это уже видели

За последние два года вы прочитали два заголовка, стоящих рядом на одном экране. Первый: столько-то сотен миллиардов долларов вложено в дата-центры за год. Второй: пузырь вот-вот лопнет. Первый заголовок про рынок капитала: кто во что вложился, чьи акции переоценены, кому нечем будет платить по долгу. Второй, тот что вас касается, — про инструмент на вашем столе. Связь между ними есть, но она проходит через несколько звеньев и работает с задержкой в годы. Разобрать эти звенья по одному и есть задача главы.

Я разберу их как экономист, потому что это моя первая профессия: до продюсирования и производства я работал инвестиционным консультантом и вёл портфели корпораций и частных клиентов. Отсюда сразу оговорка, и она жёсткая. **В этой главе нет ни одной инвестиционной рекомен-**

дации. Я не даю оценок чьим-либо акциям, не говорю, что покупать, и не говорю, чего избегать. Именно потому, что я знаю, как выглядит эта работа изнутри, я не берусь делать её в книге, которую вы читаете ради другого. Здесь только устройство: откуда берётся цена, кто на каком звене зарабатывает и что из этого следует для вашего бюджета.

Два разных счёта: обучение и работа

Первое, что стоит развести, — два принципиально разных вида затрат.

Обучение — это то, что происходит один раз, до того как вы открыли чат. Огромный разовый расход: тысячи ускорителей работают месяцами. Платит его лаборатория, из своих денег или денег инвесторов, и потом размазывает по всему сроку жизни модели.

Структура этого расхода устроена не так, как о ней принято думать. Erosch AI разобрала затраты на несколько известных обучений и получила такую раскладку: железо и инфраструктура: от 47 до 67 процентов, зарплаты исследователей вместе с их долями в компании: от 29 до 49 процентов, электричество: от 2 до 6 процентов. Электричество, вокруг которого больше всего разговоров, оказывается самой мелкой строкой. Самая крупная после железа — люди. Их несколько десятков на лабораторию, и переманивают их поштучно.

Данные, на которых учится модель, в эту раскладку не входят: у Erosch AI по ним отдельного счёта нет. Платят за них по-разному, и не всегда добровольно. Открытая покупка

выглядит как обычная сделка: 23 мая 2024 года News Corp продала OpenAI право использовать архивы своих изданий, среди них Wall Street Journal и New York Post, за сумму, которую Variety оценила больше чем в 250 миллионов долларов. Есть и куда менее опрятная история. По документам дела Bartz против Anthropic, компания закупила миллионы бумажных книг, срезала с них переплёт, прогнала страницы через сканер и отправила бумагу в переработку. В июне 2025 года федеральный суд признал это добросовестным использованием (fair use): купленный экземпляр меняется на цифровую копию, и права издателя это не нарушает. Тот же суд решил, что скачивание книг с пиратских библиотек законным не было. За это в сентябре 2025 года Anthropic заплатила 1,5 миллиарда долларов: около 500 тысяч книг, по три тысячи долларов за каждую.

Второй счёт — **работа модели**, на отраслевом жаргоне инференс (inference). Это то, что происходит каждый раз, когда вы нажимаете «отправить». Считается он в токенах, о которых шла речь в главе 2, и платится он снова и снова, до конца жизни продукта.

Долгое время обучение было главной статьёй, а работа моделей — довеском. В 2026 году это соотношение перевернулось: по оценке Gartner, впервые расходы на работу моделей превысили расходы на обучение — 23,3 миллиарда долларов против 19 в облачном сегменте, то есть 55 процентов против 45. К 2027 году доля работы, по их прогнозу, дойдёт до 59

процентов.

Для вас это не абстракция. Переворот означает, что отрасль теперь живёт с ваших запросов, а не с обещаний исследователей. И это же объясняет, почему длинный контекст и режим рассуждения стоят денег: тарифицируется работа модели по каждому конкретному запросу, а не сам факт доступа к ней.

Куда уходит стоимость доступа к модели

Деньги за доступ к модели проходят короткий путь из пяти звеньев. Понимать этот путь стоит по одной практической причине: на каждом звене цена ведёт себя по-своему, и это объясняет, куда уходит каждый рубль.

1. **Надстройка** — когда платёж идёт не лаборатории напрямую. Пример — редактор кода Cursor компании Anysphere. В феврале 2026 года он вышел на 2 миллиарда долларов годового темпа выручки (ARR, annual recurring revenue) и к апрелю поднимал раунд при оценке в 50 миллиардов. При этом в самой компании открыто называют главную угрозу: Claude Code, продукт лаборатории Anthropic, чьими моделями Cursor и торгует, а не другой редактор. Надстройка продаёт интерфейс, готовые сценарии и подобранный контекст, а токены покупает оптом у лаборатории, которая в любой день может выйти на её рынок сама.

2. **Лаборатория**, сделавшая модель: OpenAI, Anthropic, Google DeepMind, xAI и другие. Продаёт доступ к весам. Из вашей оплаты покрывает работу модели, амортизацию обу-

чения и следующее обучение.

3. **Облако**, где эта модель физически считается. Пример — Anthropic одновременно арендует мощности у двух разных облаков: у Amazon по проекту Rainier в Нью-Карлайле, штат Индиана (8 миллиардов долларов, свыше 500 тысяч чипов Trainium2, работает с октября 2025 года), и у Google по сделке от 23 октября 2025 года на использование до миллиона тензорных процессоров суммарной мощностью свыше гигаватта. OpenAI для Stargate так же арендует мощности Oracle. Собственная модель почти никогда не означает собственный дата-центр.

4. **Производитель ускорителей**: NVIDIA, AMD, и отдельно — чипы, которые облака делают для себя сами: тензорные процессоры Google (седьмое поколение, Ironwood, представлено в ноябре 2025 года) и Trainium у Amazon.

5. **Энергия и стройка**: розетка, подстанция, здание, вода на охлаждение, люди на монтаже. Пример — ниже, в разделе про узкое место.

Все пять звеньев зарабатывают одновременно — это не игра с нулевой суммой, где выигрывает только одно. Но распределена эта прибыль крайне неравномерно, и видно это из публичной отчётности. В квартале, закончившемся 26 июля 2026 года, NVIDIA показала 96,2 миллиарда долларов выручки, из них 89,0 миллиарда, сегмент дата-центров, при валовой марже 75 процентов и чистой прибыли 59,7 миллиарда. Иными словами, из каждого доллара, дошедшего до про-

изготовителя ускорителей, три четверти остаются у него как валовая прибыль. У ближайшего конкурента, AMD, за сопоставимый квартал сегмент дата-центров дал 6,7 миллиарда при валовой марже 54 процента.

Сопоставьте два этих числа — 89,0 и 6,7 — и вы увидите главное свойство рынка ускорителей на сентябрь 2026 года: он не поделён.

Узкое место: железо и розетка

Андрей Дороничев, работавший в Google в период разработки собственных чипов компании, описывает положение так: если у вас нет вычислительных мощностей, ваша модель «как машина без бензина ничего не может делать. Ну, это как машина без машины, скорее. Бензином является электричество, а вот самого двигателя у тебя нету». Он говорит это как участник рынка и заинтересованная сторона, но диагноз здесь совпадает с публичной отчётностью: двигателем торгует очень небольшое число компаний.

Альтернативы производятся, и их больше, чем кажется по новостям. Google много лет делает собственные тензорные процессоры. Amazon и Microsoft разрабатывают свои чипы для облаков. AMD выпускает линейку ускорителей и растёт быстро. Часть этих разработок делается совместно со специализированными подрядчиками, для которых заказные чипы — отдельный крупный бизнес. Направление правильное, и на горизонте нескольких лет оно должно сбить маржу лидера. На сентябрь 2026 года оно её не сбило.

Второй ограничитель — энергия, и здесь полезно отделить пугающие цифры от действующих. По оценке Международного энергетического агентства, все дата-центры мира потребили в 2024 году около 415 тераватт-часов, примерно полтора процента мирового электричества, а к 2030 году выйдут на 945 тераватт-часов, чуть менее трёх процентов. На фоне мирового потребления это немного: прирост спроса от дата-центров даёт меньше десятой части общего прироста спроса на электричество.

Ограничение работает на местном уровне, а не на мировом. В США дата-центры дают около половины всего прироста спроса на электричество, и упирается стройка в сроки: подключение к сети, трансформаторы, разрешения. Годы, а не месяцы.

Наглядный адрес: флагманский участок Stargate в Абилине, штат Техас. По данным исследовательской группы Epoch AI, на апрель 2026 года там было запущено около 0,3 гигаватта мощности с целью выйти на 1,2 гигаватта к концу года. По плану на семи площадках проекта к 2029 году должно быть свыше 9 гигаватт при заявленных 500 миллиардах долларов вложений. Это и есть настоящий потолок скорости: годы между разрешением на стройку и первым киловаттом объясняют, почему деньги в отрасли уходят в бетон и трансформаторы на годы раньше, чем в кассу приходит выручка.

Посредники: за один ответ можно заплатить трижды

Между вами и весами модели выросла отдельная отрасль, и она зарабатывает на проходе.

Агрегаторы доступа дают один ключ и один счёт ко множеству моделей разных лабораторий, а зарабатывают комиссию с оборота. Крупнейший — OpenRouter (openrouter.ai, раздел [rankings](#)): по его собственному замеру, за год к ноябрю 2025-го через сервис прошло свыше ста триллионов токенов, около 30 процентов пришлось на модели с открытыми весами, а больше половины всего объёма — на задачи программирования. Полезны они там, где нужно сравнить модели или подстраховаться от сбоя одного поставщика.

Рейтинги и арены, из которых самая известная — Arena (arena.ai), сравнивают модели вслепую на голосах пользователей. Прямых денег они не берут; их актив — внимание и данные о предпочтениях. Пользоваться ими стоит с поправкой: они меряют, какой ответ больше нравится людям, а не какой из них верен. К рабочей задаче это отношение имеет косвенное.

Надстройки продают обжитое рабочее место вокруг модели, а не саму модель. Здесь и возникает ситуация, за которую платят по три раза одновременно. Пример: подписка ChatGPT Plus стоит 20 долларов в месяц напрямую у OpenAI; отдельно существует подписка на GitHub Copilot (Microsoft), который умеет обращаться к моделям OpenAI, Anthropic и Google по выбору; и отдельно — подписка на

Cursor, который тоже ходит в одну из этих моделей. Три счёта, а токены на дне — из одного и того же короткого списка лабораторий.

Экономический принцип простой: надстройка оправдывает свою цену, когда продаёт сценарий работы, которого нет у прямого доступа к модели. Плата за тот же доступ через другое окно — это плата второй раз за одно и то же. Что из этого выбрать и стоит ли платить вообще — глава 5.

Почему цена падает, а счёт растёт

Теперь то, ради чего эту главу стоит дочитать до конца, если вы планируете бюджет.

Цена за единицу качества падает, и падает быстро. Erosch AI измерила это по нескольким наборам задач: скорость снижения цены за фиксированный уровень результата составляет от 9 до 900 раз в год в зависимости от задачи. Самый наглядный ряд — стоимость ответа уровня лучшей модели весны 2023 года на тесте общих знаний: 37,50 доллара за миллион токенов в марте 2023 года и 18 центов в феврале 2025-го. В двести восемь раз дешевле за два года.

Причин у этого три, и все они продолжают действовать. Первая: новое поколение ускорителей даёт больше вычислений на ватт. Вторая: инженеры научились сжимать модели, а большие модели научились обучать маленькие, и маленькая догоняет большую годичной давности при десятикратно меньшей цене работы. Третья: за один и тот же класс задач конкурируют лаборатории из разных стран, включая откры-

тые модели, которые задают потолок цены снизу.

А теперь неприятное. Ваш счёт при этом будет расти.

Экономист Уильям Стэнли Джевонс описал этот механизм в 1865 году в книге «Угольный вопрос». Британия тогда обсуждала, что паровые машины становятся всё экономичнее и потребление угля должно снизиться. Джевонс показал обратное: чем дешевле обходится единица работы машины, тем шире её применяют, и суммарное потребление угля растёт. С тех пор это называют парадоксом Джевонса, и он описывает вашу ситуацию точно.

Токен дешевеет — а вы начинаете давать модели документы целиком вместо абзацев, включаете режим рассуждения, где на один ответ уходит вдесятеро больше токенов на внутренний черновик, и запускаете агентов, которые за одну вашу задачу делают тридцать обращений подряд. Расход растёт быстрее, чем падает цена.

Практический вывод для планирования на три года: **не закладываете в бюджет цену, закладываете объём.** Цена за единицу выполненной работы почти наверняка будет ниже. Количество работы, которую вы захотите отдать модели, вырастет сильнее. Считайте не «сколько стоит подписка», а «сколько стоит одна моя рабочая задача и сколько таких задач в месяц» — это единственная величина, которая переживёт смену тарифов.

Конец ознакомительного фрагмента.

Текст предоставлен ООО «Литрес».

Прочитайте эту книгу целиком, [купив полную легальную версию](#) на Литрес.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.