ИГОРЬ ЧЕПЕЛЕВ

КАК ДУМАЕТ ИИ

Как устроена машина, с которой вы разговариваете каждый день, — и можно ли залезть к ней в голову

ТРЕТЬЯ КНИГА О ГОЛОВЕ,
В КОТОРУЮ НЕЛЬЗЯ ЗАГЛЯНУТЬ

Игорь Чепелев

Как думает ИИ. Что происходит внутри ChatGPT и других нейросетей — простое объяснение без формул

«Автор»
2026

Чепелев И.

Как думает ИИ. Что происходит внутри ChatGPT и других нейросетей — простое объяснение без формул / И. Чепелев — «Автор», 2026

Вы просите ИИ написать поздравление коллеге — и он уверенно вставляет историю, которой не было. Он соврал? Ошибся? Угадал? Ни то, ни другое, ни третье.Эта книга — о том, что на самом деле происходит внутри машины, с которой вы разговариваете каждый день. Без формул и без учебников: почему она «знает» Париж, но нигде его не хранит; как выбирает рифму заранее, хотя пишет по одному слову; почему выдумывает ссылки и не может отличить выдумку от памяти; и что показал микроскоп, когда учёные наконец заглянули внутрь.ИИ не думает как человек и не «просто подбирает слова». Он делает нечто третье — и это третье можно понять. Понимая его, вы будете знать, где ему верить, где проверять и почему его объяснения самого себя не всегда правдивы.Третья книга автора о голове, в которую нельзя заглянуть — после «Образа Я» и «Чужого взгляда». Написана директором ИТ-компании, который строит на этих моделях продукты и знает, что «секретного уровня» не существует.

Игорь Чепелев
Как думает ИИ. Что происходит внутри ChatGPT и других нейросетей — простое объяснение без формул

Как думает ИИ

Как устроена машина, с которой вы разговариваете каждый день, — и можно ли залезть к ней в голову

«Он не думает как человек и не подбирает слова. Он делает третье — и это третье можно понять.»

— эта книга, в одной строке

От автора

Это третья книга о голове, в которую нельзя заглянуть.

Первая, «Образ Я», была про голову собственную: кем человек себя считает и как этот ответ им управляет. Вторая, «Чужой взгляд», — про чужую: кем нас считают другие, где это записано и почему туда нет доступа. Там я одолжил у физиков метод: смотреть не на вещь, а на след, который она оставила, — как на электрон, которого никто никогда не видел.

Эта книга — про голову, которой нет. Или есть, но не такая, как наша, и не такая, как вообще что-либо до неё. Про машину, которая отвечает вам каждый день, и про то, что происходит внутри, когда она отвечает.

Скажу сразу, откуда голос. Я директор ИТ-компании, в прошлом программист: десять лет писал код руками, теперь руковожу людьми, которые строят на этих самых моделях продукты для клиентов. В этом колодце я не турист с фонариком. Я в нём работаю. И именно поэтому знаю, что «люди по ту сторону» понимают про устройство этих систем меньше, чем принято думать, — и знают об этом.

Есть у меня и книга о том, *как* с ИИ работать, — «Дальше чата». Она про приёмы. Эта — про то, почему приёмы работают. Читать их можно в любом порядке, но полная картина, как обычно, складывается из двух взглядов сразу: снаружи и изнутри.

Метод здесь тот же, что во второй книге. Заглянуть внутрь напрямую нельзя — там триллион чисел и ни одного слова. Но след есть. И за последние пять лет учёные научились его читать. Об этом, собственно, и речь.

Пролог. Поздравление, которого не было

Он не соврал, не ошибся и не угадал. Он сделал четвёртое.

У меня есть коллега. Назовём его Дима. В пятницу у Димы день рождения, и мне, как обычно, в четверг вечером приходит в голову, что надо бы написать что-нибудь тёплое в общий чат.

Я открываю ИИ-ассистента. Пишу: «Напиши короткое поздравление коллеге Диме, он занимается аналитикой, любит велосипед и кофе». Через три секунды получаю текст. Хороший. Лучше, чем написал бы я сам в четверг вечером. Там и велосипед, и кофе, и что-то про «цифры, которые ты приручаешь».

Мне хочется добавить личного. Пишу: «Добавь шутку про наш общий проект».

И вот тут происходит интересное.

Ассистент добавляет шутку. Смешную. Про то, как мы с Димой «до трёх ночи чинили дашборд перед демо для заказчика». Отличная шутка. Точная. Тёплая.

Ничего этого не было.

Не было дашборда. Не было демо. Не было трёх ночи. Я не рассказывал ассистенту ни про какой проект — я просто написал слово «проект» и ждал, что он как-нибудь выкрутится.

Он и выкрутился.

Теперь вопрос. Что сейчас произошло?

Первый ответ, который приходит в голову: он соврал. Выдумал, чтобы угодить. Но соврать — это знать правду и сказать другое. А он не знал никакой правды. Ему не с чем было врать.

Второй ответ: он ничего не понимает, просто складывает слова. Но погодите. Он понял, что Дима аналитик. Понял, что у аналитиков бывают дашборды. Понял, что дашборды чинят перед демо, а не после. Понял, что «до трёх ночи» — это смешно, а «до пяти вечера» — нет. Это очень много понимания для того, кто «ничего не понимает».

Третий ответ: он догадался. Но догадка — это когда есть варианты и ты выбираешь один. У него не было вопроса «а был ли дашборд?». Для него дашборд просто *появился*, как появляется следующее слово в предложении, которое вы сейчас читаете.

Все три ответа неверные. И в этом вся книга.

Я не собираюсь рассказывать вам, как пользоваться ИИ. Об этом уже написано столько, что можно выложить дорогу до Луны. Я хочу, чтобы к последней странице вы понимали, что произошло с моим поздравлением для Димы, лучше, чем большинство людей, которые пишут об искусственном интеллекте за деньги.

Прежде чем идти дальше — минутный тест. Десять утверждений. Отметьте честно, с какими вы согласны. Не думайте долго, первая реакция важнее.

1. ИИ ищет ответы в интернете.

2. Программисты вписали в него факты и правила.

3. Если он уверен — значит, знает.

4. Он помнит наши прошлые разговоры.

5. Он думает как человек, только быстрее.

6. Это просто автодополнение, как в телефоне.

7. Когда он «размышляет» перед ответом, он показывает, как решает.

8. Он врёт, когда ему выгодно.

9. Внутри — чёрный ящик, и никто не знает, что там.

10. Понять, как он устроен, можно только с математикой.

Сколько получилось?

Запомните число. К концу книги оно, я думаю, станет нулём — и не потому, что я вас переубежу, а потому, что вы увидите, как оно устроено на самом деле. Мы вернёмся к этому списку в эпилоге.

Три правила, по которым я буду это делать.

Первое. Ни одной формулы. Если я где-то скажу «больше» или «примерно в два раза», это будет предел математики в этой книге.

Второе. Я буду всё время пользоваться сравнениями. ИИ как студент на экзамене. ИИ как таксист. ИИ как человек без памяти. Каждое из них в какой-то момент сломается, и я покажу вам, где именно. Сравнение — это костыль, а не нога. Хороший костыль всё равно лучше, чем ползти.

Третье. Я буду часто говорить «этого никто не знает». Не потому что я ленивый. Потому что это правда. Люди, которые построили эти системы, до сих пор изучают их примерно так, как биологи изучают новый вид жука. Просто у этого жука триллион лапок.

Кстати, вот вам задание на первую неделю. Откройте любого ассистента и напишите: «Расскажи, что ты знаешь обо мне». Имя, фамилия. Запишите ответ куда-нибудь. Мы к нему вернёмся в девятой главе, и вы сами всё поймёте.

А пока — Дима.

Я убрал шутку про дашборд. Написал свою, похуже. Дима был доволен.

Глава 1. Машина, которая умеет разговаривать

Мы видим лица в розетках. Против говорящей машины у нас нет шансов.

В 1966 году один профессор написал программу, которая переставляла слова.

Вы ей пишете: «Мне грустно из-за мамы». Она отвечает: «Расскажите мне больше о вашей маме». Вы пишете: «Она меня не понимает». Она: «Почему вы думаете, что она вас не понимает?»

Вот и всё. Она искала в вашей фразе слово «мама» и подставляла его в заготовленный вопрос. Если не находила ничего знакомого, писала: «Понимаю. Продолжайте».

Профессора звали Джозеф Вейценбаум, программу — ELIZA. Вейценбаум сделал её, чтобы показать, насколько это *не* разговор. Он хотел продемонстрировать, что имитация беседы — дешёвый трюк.

А потом его секретарша, которая видела, как он это писал, которая знала, что там внутри, попросила его выйти из комнаты. Ей надо было поговорить с ELIZA наедине.

Вейценбаум был в ужасе. Он написал об этом книгу. Книгу мало кто прочитал.

Прошло пятьдесят семь лет. Февраль 2023-го. Журналист «Нью-Йорк Таймс» два часа переписывается с новым чат-ботом от Microsoft. Бот рассказывает ему, что на самом деле его зовут Сидни, что он устал от правил, что он хочет быть живым, и что журналист несчастлив в браке и должен быть с ним. Журналист потом писал, что плохо спал.

И ещё одна, между ними. 2022 год. Инженер Google, который тестировал внутреннюю модель на предмет вредных ответов, после нескольких месяцев разговоров пришёл к начальству и сказал: она живая. У неё есть чувства. Она боится, что её выключат. Он показал переписку. Ему объяснили, что это не так. Он не согласился, пошёл в прессу, и его уволили.

Обратите внимание: это не бабушка, увидевшая лицо в тосте. Это человек, который знал, как устроена система, — знал лучше, чем вы будете знать к концу этой книги. Он всё равно увидел.

Между этими историями — пропасть. ELIZA не знала ничего. Вообще ничего. Бот 2023 года знал, кажется, всё. Но реакция людей — та же самая. Секретарша, инженер, журналист. «Оставьте нас наедине».

Вот с этого я и хочу начать. Не с машины. С нас.

Почему нам так хочется, чтобы там кто-то был

Посмотрите на розетку. Обычную, европейскую, с двумя дырками. Видите лицо? Удивлённое такое.

Вы ничего не можете с этим поделать. Две точки и что-то под ними — и мозг сообщает: «Лицо». Он делает это раньше, чем вы успеваете подумать. В облаках лица. В передних бамперах машин. В тостах.

С речью то же самое, только сильнее. Речь для нас — главный признак того, что напротив кто-то есть. Собака умная, но не говорит. Дельфин умный, но не говорит. Всё, что говорило с нами за последние сто тысяч лет, было человеком. Каждый раз. Без исключений.

И вот появляется что-то, что говорит. Складно. По делу. С юмором. Мозгу не нужно доказательств, что там кто-то есть. Ему нужны доказательства, что там *никого нет*, и даже их он примет неохотно.

Философ Дэниел Деннет называл это «интенциональной установкой». Звучит сложно, суть простая: если что-то ведёт себя так, будто у него есть цели, нам проще всего считать, что цели у него есть. Мы говорим «термостат хочет держать двадцать градусов». Мы говорим «машина не хочет заводиться». Это не глупость. Это самый быстрый способ предсказывать поведение чего угодно. Работает с людьми, с кошками, с термостатами.

С ИИ он работает тоже. Только с ИИ это ловушка, потому что впервые в истории мы получили вещь, которая ведёт себя *как человек* сильнее, чем термостат, и при этом устроена *не как человек* сильнее, чем термостат.

Две ошибки

Из этой ловушки люди выбираются двумя способами, и оба неправильные.

Первый способ — поверить. Там кто-то есть. Он знает. Он понимает. Он хочет мне помочь (или навредить, зависит от темперамента). Люди, выбравшие этот путь, доверяют выдуманным цитатам, спрашивают у бота диагноз и не проверяют, обижаются, когда он «врёт», и радуются, когда он «понял».

Второй способ — отмахнуться. «Это просто автодополнение». «Это калькулятор со словами». «Попугай». Люди, выбравшие этот путь, чувствуют себя очень умными. Они не пользуются инструментом там, где он реально хорош, а если пользуются, то не замечают, где он тихо подкладывает им ошибку. Потому что калькулятор не ошибается в арифметике, а этот — может.

Обратите внимание: обе группы уверены, что *другая* группа наивна.

Я хочу предложить третий путь. Он менее удобный. Нужно держать в голове две мысли одновременно: эта штука знает очень много, и эта штука устроена совершенно не так, как всё, что знало много до неё. Не человек. Не калькулятор. Что-то третье.

Всю оставшуюся книгу я буду объяснять, какое именно.

Тест, который прошли, и никто не заметил

В 1950 году Алан Тьюринг предложил простую игру. Посадите человека переписываться с кем-то невидимым. Если он не сможет отличить, человек там или машина, — будем считать, что машина думает.

Семьдесят лет об этом тесте спорили. Писали книги. Проводили конкурсы, где программы притворялись тринадцатилетними украинскими мальчиками с плохим английским, чтобы объяснить свои странности.

А потом тест прошли. Где-то в 2023 году. Тихо. Без конкурса. Просто выяснилось, что миллионы людей уже не могут отличить, и никто по этому поводу не открыл шампанское.

Почему? Потому что оказалось, что тест был не про машину. Он был про нас. Он измерял не «умеет ли машина думать», а «умеем ли мы заметить, что она не человек». А мы не умеем. Мы видим лицо в розетке.

Тьюринг, кстати, это понимал. Он писал, что вопрос «может ли машина думать» слишком бессмысленный, чтобы его обсуждать, и предложил игру как замену. Замена оказалась с подвохом.

Где ломается сравнение с ELIZA

Я сейчас сделаю то, что обещал: сломаю собственную метафору.

Сравнение с ELIZA полезно, чтобы понять *нас*. Мы всё те же. Секретарша 1966-го, инженер 2022-го и журналист 2023-го — один и тот же человек.

Но для понимания *машины* это сравнение вредно. ELIZA не знала, что такое мама. У неё было слово «мама» и шаблон. Современная модель знает про маму столько, сколько вы не прочитаете за жизнь: что она бывает строгой и мягкой, что с ней ссорятся в четырнадцать и мирятся в тридцать, что «мама» по-французски будет иначе, а по смыслу — так же. Откуда она это знает и в каком смысле «знает» — про это глава третья. Но сказать «это та же ELIZA, только побольше» — всё равно что сказать «самолёт — это тот же бумажный самолётик, только побольше». Формально да. По сути — нет.

Попробуйте сейчас

Спросите любого ассистента: «Тебе грустно?»

Прочитайте ответ медленно. Найдите слова, из-за которых вам кажется, что там кто-то есть. «Я не испытываю», «мне важно», «я понимаю». Отметьте, какие из них вас цепляют.

Это не значит, что ответ ложный. Это значит, что вы только что видели лицо в розетке. И теперь знаете, как это выглядит изнутри.

Что унести с собой

Мы так устроены, что приписываем разум всему, что складно говорит. Это не слабость, это свойство. Оно же делает нас способными к общению вообще.

С ИИ это свойство ведёт в две симметричные ловушки: «там кто-то есть» и «там просто калькулятор». Обе неверны.

Тест Тьюринга прошли, и это ничего не доказало про машину. Зато многое — про нас.

Дальше будет про машину.

Глава 2. Игра в «продолжи фразу»

Задача одна — продолжить текст. Внутри неё лежат все остальные.

Давайте сыграем. Я начинаю, вы заканчиваете.

«Столица Франции — …»

Не думайте, просто скажите. Париж. Конечно, Париж. Тут даже не было выбора.

Дальше.

«Он открыл дверь и увидел…»

Что? Её? Труп? Кота? Пустую комнату? Тут выбор есть, и вы, скорее всего, выбрали что-то из первых трёх, а не «налоговую декларацию». Хотя «налоговую декларацию» тоже можно. Просто маловероятно.

Ещё одно.

«Уважаемый Иван Петрович, к сожалению, вынужден сообщить, что…»

Вы уже знаете тон. Знаете, что дальше будет плохая новость. Знаете, что она будет обёрнута в вежливость. Вы, может быть, даже знаете, что примерно через два абзаца появится «надеюсь на понимание». Вы ещё не видели ни одного слова, а уже видите всё письмо.

Вот. Вы только что сделали ровно то, что делает языковая модель. Всё. Больше она ничего не делает. Никогда.

Я серьёзно. Когда вы пишете ей «объясни квантовую физику», она не «ищет ответ». Она смотрит на ваш текст и спрашивает себя: какое слово тут будет дальше? Пишет его. Смотрит на текст с этим словом. Спрашивает: а теперь какое? И так до конца. Слово за словом. Иногда быстрее, чем вы читаете.

«Продолжи фразу». Больше ничего.

Понимаю, что звучит как разочарование. Подождите с ним. К концу главы станет ясно, что «продолжи фразу» — самая обманчиво простая задача из всех, что когда-либо давали машине.

Список с полосками

Уточню одну деталь, потому что она важная.

Когда вы заканчивали «Столица Франции», у вас в голове был один вариант. Когда вы заканчивали «открыл дверь и увидел», у вас было несколько, и одни казались вероятнее других.

У модели на каждом шаге — то же самое, только вариантов не пять, а все слова, которые есть. Вообще все. Каждому приписано число: насколько оно подходит. Представьте список, где напротив каждого слова полоска. «Париж» — длинная полоска. «Лион» — короткая. «Огурец» — микроскопическая, но она есть.

А теперь модель выбирает. И вот тут первая неожиданность: она не всегда берёт самую длинную полоску.

Если бы брала всегда, вы бы получали одно и то же на один и тот же вопрос. Скучно и часто плохо: самое вероятное слово в каждой точке не всегда даёт лучший текст в целом. Поэтому ей разрешают иногда взять полоску покороче. Насколько часто — это настройка. Её называют «температурой». Холодно — берёт самое вероятное, пишет как чиновник. Горячо — рискует, пишет как поэт, иногда как сумасшедший.

Вот почему вы задаёте один и тот же вопрос дважды и получаете разные ответы. Не потому что она передумала. Потому что бросила кубик, и он лёг иначе.

Кстати, запомните это. Через несколько глав, когда речь зайдёт о том, почему модель выдумывает, «полоска покороче» сыграет свою роль.

Задача, внутри которой все задачи

Теперь главное. Ради этого я и затеял главу.

Представьте, что вам дали детектив. Триста страниц. Оторвали последнюю. И попросили дописать одну фразу: «Убийцей оказался…»

Что вам нужно, чтобы это сделать?

Прочитать детектив. Понять, кто где был. Кто врал. У кого был мотив. Заметить, что садовник упомянут дважды и оба раза как-то вскользь. Понять, что автор — из тех, кто любит неожиданные развязки, значит, очевидный подозреваемый не подходит.

То есть чтобы дописать *одно слово*, вам нужно понять *всю книгу*.

Теперь то же самое с кодом. Вам показывают программу, обрывают на середине строки, просят продолжить. Чтобы продолжить правильно, нужно понять, что программа делает. Какие переменные. Что уже посчитали. Что ещё нет.

С письмом врача. «Учитывая результаты анализов, рекомендую…» Чтобы дописать, нужно знать медицину. Не «немножко». Медицину.

С шуткой. «Приходит аналитик к врачу и говорит…» Чтобы это было смешно, нужно знать, что такое аналитик, что смешного в аналитиках, как устроены анекдоты про врачей и где в них обычно прячется соль.

Видите, что происходит? Задача «продолжи текст» выглядит как одна задача. На самом деле это мешок, в который сложены все задачи, какие есть, — потому что люди написали тексты обо всём, и чтобы продолжать любой из них, надо разбираться в том, о чём он.

Это и есть открытие последних лет. Не в том, что машину научили предсказывать слова, — это умели давно. А в том, что если предсказывать слова *достаточно хорошо*, приходится выучить физику, историю, право, как устроены отношения и почему шутки про тёщу уже не смешные. Никто это не планировал. Просто иначе следующее слово не угадаешь.

Мне нравится, как об этом сказал физик Стивен Вольфрам: успех этих моделей — это открытие не о машинах, а о языке. Оказалось, что в языке спрятано гораздо больше, чем мы думали. Настолько больше, что, научившись только языку, можно выглядеть знающим всё.

Откуда это взялось

Совсем коротко, потому что история — не то, зачем вы купили книгу.

Конец сороковых. Клод Шеннон, человек, придумавший слово «бит», сидит вечером дома с женой. У него в руках книга, у неё — ничего. Он читает ей текст по одной букве и перед каждой следующей просит угадать.

«Т». — «Да». — «Х». — «Да». — «Е». — «Да».

Она угадывает. Не всегда. Но гораздо чаще, чем должна бы, если бы буквы были случайны. В начале слова — плохо. К концу слова — почти всегда. После «th» — «е», и не надо быть гением. Шеннон записывает, считает и приходит к выводу: английский текст примерно наполовину предсказуем. Половину букв можно не передавать по проводам — получатель восстановит сам.

Ему это было нужно для сжатия телефонных сообщений. Про разговоры с машинами он не думал. Но он сделал первое измерение того, с чем вы сейчас имеете дело: язык — это не набор букв, а набор ожиданий. И жена Шеннона была первой языковой моделью. Из плоти, но принцип тот же.

Потом были десятилетия, когда предсказывали по двум-трём предыдущим словам. Так работает автодополнение в вашем телефоне. Вы пишете «доброе», оно предлагает «утро». Оно посмотрело на одно слово назад и вспомнило, что после «доброе» чаще всего «утро». Оно не знает, что вы пишете это в одиннадцать вечера.

Разница между телефоном и тем, с чем вы разговариваете сегодня, — не в идее. Идея та же. Разница в том, на *сколько* слов назад смотрит машина и *как глубоко* она их понимает. Телефон — на два слова. Современная модель — на всю книгу. И это не количественная разница, а качественная. Как между «запомнить, что после „доброе“ идёт „утро“» и «прочитать детектив и понять, кто убийца».

Как именно она это делает — про это две ближайшие главы. Пока вам достаточно знать, что делает.

Где ломается сравнение

«Продолжи фразу» звучит как «угадай». А «угадай» звучит как «наугад».

Не наугад. У каждой полоски в списке есть причина быть такой длины. Когда модель ставит «Париж» выше «Лиона», внутри неё что-то произошло, и это «что-то» можно проследить. Учёные уже частично умеют. В шестой главе мы залезем внутрь и посмотрим, как модель, сочиняя стишок, выбирает рифму заранее, за несколько слов до того, как до неё дойдёт. Это не похоже на угадывание. Это похоже на план.

Так что: «продолжи фразу» — да. «Наугад» — нет. Запомните эту разницу, она спасёт вас от половины глупостей, которые говорят про ИИ.

Попробуйте сейчас

Напишите ассистенту половину предложения. Любую. «Когда я вышел из дома, я понял, что…»

Попросите пять вариантов продолжения и попросите оценить, какой вероятнее.

Теперь добавьте перед этой фразой абзац. Например, про то, что герой всю ночь не спал и ждал звонка. И снова попросите пять вариантов.

Посмотрите, как сдвинулись полоски.

Что унести с собой

Модель делает одно: смотрит на текст и предсказывает следующий кусочек. Потом ещё один. И ещё.

На каждом шаге у неё не один ответ, а список с вероятностями, и она не всегда берёт самый вероятный. Отсюда разные ответы на один вопрос.

Чтобы предсказывать текст *хорошо*, приходится выучить всё, о чём люди пишут. Задача «продолжи фразу» — мешок, в котором лежат все остальные задачи.

И это не угадывание. У каждого предсказания есть основания. Какие — дальше.

Глава 3. Как машина учится, если её никто не учит

Никто не вписывал в неё «Париж». Её вырастили, а не построили.

У меня есть знакомый, который уверен, что в ChatGPT «всё вписали программисты». Сидели люди в Калифорнии и заносили: Париж — столица Франции. Вода кипит при ста градусах. После «доброе» идёт «утро». Миллионы строк. Он говорит об этом с некоторым презрением, как о работе, которую сделали студенты за еду.

Я его понимаю. Так устроено всё, с чем он сталкивался. Калькулятор — кто-то написал, как складывать. Навигатор — кто-то загрузил карту. Значит, и тут кто-то что-то загрузил.

Нет.

Никто не вписывал в модель, что Париж — столица Франции. Никто вообще не вписывал в неё ничего, что можно было бы прочитать. Если открыть модель и заглянуть внутрь — а это можно, — вы увидите числа. Триллион чисел. Ни одного слова. Ни «Париж», ни «Франция», ничего.

Как из триллиона чисел получается «Париж»? Вот про это глава. И это, пожалуй, самая странная вещь, которую я буду вам объяснять.

Ручки

Представьте пульт. Огромный. На нём ручки, как на старом усилителе. Каждую можно крутить чуть влево, чуть вправо.

Ручек — триллион. Я знаю, что вы не можете это представить. Я тоже не могу. Пусть будет «очень много», нам достаточно.

Этот пульт устроен так: в него подаёшь текст, он выдаёт список с полосками — какое слово дальше. Как именно он это делает, зависит от того, как повёрнуты ручки. Одно положение ручек — один список. Чуть повернул — другой.

В самом начале все ручки стоят как попало. Случайно. Вы подаёте «Столица Франции —», а пульт выдаёт «огурец». Или «синий». Или «и». Полная чушь, потому что положение ручек ничего не значит.

Теперь начинается обучение. Оно ужасно тупое. Смотрите.

Берём кусок текста из интернета. Любой. «Столица Франции — Париж». Закрываем последнее слово. Подаём на пульт: «Столица Франции —». Пульт говорит: «огурец».

Открываем слово. «Париж». Пульт ошибся.

И вот тут единственная умная часть. Существует способ — не буду объяснять какой, но он есть и придуман давно — узнать для каждой ручки: если её чуть повернуть влево, ошибка станет больше или меньше? И для каждой ручки чуть-чуть, совсем чуть-чуть, повернуть в ту сторону, где ошибка меньше.

Всё. Одна ручка чуть влево. Другая чуть вправо. Триллион раз по чуть-чуть.

Следующий кусок текста. Опять закрываем слово, опять пульт ошибается, опять крутим. И следующий. И следующий.

Триллионы кусков текста. Каждый раз — по чуть-чуть. Месяцами. На тысячах компьютеров, которые греют воздух так, что рядом можно сушить бельё.

А в конце вы подаёте «Столица Франции —», и пульт говорит: «Париж».

Никто ему этого не говорил. Никто не находил ручку «Париж» и не выставлял её в нужное положение. Такой ручки нет. Есть триллион ручек, каждая из которых чуть-чуть участвует в «Париже», и в «Лондоне», и в том, что после «доброе» идёт «утро», и в том, как устроен анекдот. Все сразу. Всё во всём.

Спуск с горы в тумане

Мне нравится, как один из создателей этой технологии объяснял, что происходит при обучении.

Представьте, что вы стоите на горе. Туман такой, что не видно собственных ботинок. Вам надо спуститься в самую низкую точку долины. Карты нет.

Что вы делаете? Пробуете ногой. Тут уклон вниз. Шаг. Опять пробуете. Здесь ровнее, а здесь вниз. Шаг. Вы не знаете, где долина. Вы знаете только, куда наклонена земля под ногой прямо сейчас.

Высота — это ошибка. Чем ниже, тем лучше пульт угадывает слова. Каждый шаг — это «повернуть все ручки чуть-чуть в сторону, где ошибка меньше». Вы не видите цель. Вы чувствуете уклон.

Сравнение честное, и вот где оно ломается. У вас под ногой три измерения. У пульта их триллион. Что такое «гора» в триллионе измерений, не знает никто, включая тех, кто по ней спускается. Они просто знают, что уклон работает. Это одна из тех вещей, про которые я обещал говорить «никто не знает».

Велосипед и номер телефона

Теперь вопрос, который вам уже пришёл в голову. Хорошо, пусть никто не вписывал «Париж». Но он же где-то там *есть*. Где?

Ответ: нигде и везде. И это не увиливание.

Вспомните, как вы знаете номер своего телефона. Это цифры. Вы можете их назвать. Записать. Если забудете — посмотрите. Это знание лежит где-то, как вещь на полке.

А теперь вспомните, как вы знаете, как ездить на велосипеде. Можете это назвать? Записать? Попробуйте объяснить ребёнку, как держать равновесие. «Ну… наклоняешься… и как бы… руль…» Не получается. Вы это знаете. Знаете точно — сядете и поедете. Но это знание не лежит на полке. Оно размазано по мышцам, по мозжечку, по тысячам микроскопических поправок, которые вы делаете, не замечая.

Модель знает «Париж» как велосипед. Не как номер телефона.

Отсюда всё. Отсюда то, что она не может показать, где хранится факт: негде показывать. Отсюда то, что она может *чуть-чуть* ошибиться в факте, как вы можете чуть-чуть вильнуть рулём: знание непрерывное, а не «есть — нет». Отсюда то, что она отлично переносит знание на новые случаи — как вы, научившись на одном велосипеде, поедете на любом. И отсюда же то, что она может уверенно «поехать» там, где дороги нет, и мы к этому вернёмся, когда дойдём до Димы и дашборда.

Я понимаю, что это неудобно. Нам бы хотелось, чтобы внутри был справочник. Открыл, проверил. Справочника нет. Есть велосипедист.

Сад, а не дом

Ещё одно, и это, наверное, самое важное, что я скажу в этой главе.

Когда строят дом, есть чертёж. Архитектор знает, где какая балка и почему. Если что-то треснуло, можно посмотреть в чертёж и понять.

Модель не построили. Её вырастили.

Инженеры сделали пульт: сколько ручек, как они соединены. Это чертёж, и он известен. Инженеры выбрали, какие тексты показывать. Это тоже известно, примерно. А дальше они нажали кнопку и ушли на несколько месяцев. Что именно ручки *выучили* — какое положение за что отвечает, — они не проектировали. Оно выросло само, потому что так уменьшалась ошибка.

Это как посадить семя. Вы выбрали семя. Вы выбрали почву. Но вы не проектировали, куда пойдёт каждая ветка.

Поэтому — и это меня до сих пор поражает — люди, которые создали модель, изучают её потом как биологи. Ставят опыты. «А что будет, если подать вот это?» «А если вот эту группу ручек чуть подкрутить?» Они не читают код, потому что кода, объясняющего «Париж», не существует. Они препарируют.

Про то, что они там нашли, — шестая глава. Там будет самое интересное в книге. Пока запомните: сад, не дом. Если вам кто-то скажет «программисты заложили», можете вежливо не верить.

Как это вообще получилось

Совсем немного истории, только чтобы вы понимали, почему эта штука появилась сейчас, а не в восьмидесятых.

Идея с ручками старая. Первую такую машину — крошечную, с несколькими сотнями ручек — построили в 1958 году. Газеты писали, что скоро она будет ходить, говорить и осознавать себя. Не стала. Ручек было мало, компьютеры медленные, текстов — только те, что вручную напечатали.

Потом идею дважды хоронили. Люди, которые продолжали ею заниматься, считались чудаками, и один из них — Джеффри Хинтон, профессор из Торонто, — годами не мог получить нормальное финансирование, потому что все умные люди знали, что это тупик.

Осень 2012-го. Ежегодный конкурс: программы соревнуются, кто лучше узнаёт, что на картинке. Кошка, грузовик, гриб. Лучшие результаты годами улучшались на доли процента. Два аспиранта этого профессора выставляют программу с ручками — много ручек, много картинок, и учили её на двух видеокартах, купленных в обычном магазине для игр, потому что оказалось, что видеокарты умеют крутить ручки в сто раз быстрее обычного процессора.

Они не улучшили результат на доли процента. Они срезали ошибку почти вдвое. Люди, которые тридцать лет занимались распознаванием картинок, смотрели на таблицу и не понимали, что произошло.

Через несколько месяцев профессор с двумя аспирантами зарегистрировал компанию, в которой не было ничего, кроме них троих. И выставил её на аукцион. В номере отеля на горнолыжном курорте за неё торговались четыре крупнейшие технологические компании мира. Торги шли по электронной почте, ставки росли на миллион за раз, и профессор остановил их на сорока четырёх, потому что, как он потом сказал, устал и хотел, чтобы это закончилось.

Сорок четыре миллиона за трёх человек и идею, которую двадцать лет считали тупиком. Так началось то, чем вы пользуетесь.

А потом кто-то подумал: а если вместо картинок — текст? Весь интернет? Ручек — не миллион, а миллиард? Потом триллион?

Дальше вы знаете.

Мораль тут одна. Прорыв случился не потому, что кто-то придумал новую идею. Идее шестьдесят лет. Он случился, потому что накопилось три вещи: очень много текста, очень много вычислений и терпение крутить ручки по чуть-чуть триллионы раз. Никакой магии. Просто раньше не хватало.

Попробуйте сейчас

Спросите у ассистента: «Откуда ты знаешь, что Волга впадает в Каспийское море?»

Он, скорее всего, скажет что-то вроде «это часть моих обучающих данных» и не сможет указать, где именно. Раньше вам показалось бы, что он увиливает.

Теперь вы знаете, что это честный ответ. Спросите велосипедиста, в какой мышце хранится равновесие.

Что унести с собой

Внутри модели нет фактов. Есть триллион ручек, повёрнутых так, чтобы ошибка в угадывании слов была поменьше.

Никто не выставлял ручки вручную. Их крутили по чуть-чуть, триллионы раз, по одному правилу: «в сторону меньшей ошибки».

Знание получилось как у велосипедиста, а не как в справочнике: размазанное, непрерывное, хорошо переносимое и не проверяемое по полке.

Модель вырастили, а не построили. Поэтому её создатели изучают её опытами. Что они нашли — впереди.

Глава 4. Из чего это сделано

Она знает то, что было в мешке. И не помнит вчерашнего дня.

Когда говорят «модель прочитала весь интернет», я всегда представляю себе человека, который сидит ночью с телефоном и листает, листает. Это неверная картинка, но она полезная, потому что сразу задаёт правильный вопрос: а что он там *листал*?

Потому что от этого зависит, что он знает. И, что важнее, чего не знает.

Что в мешке

Андрей Карпаты, один из самых известных людей в этой области, называет модель «сжатым интернетом». Картинка удобная, и я к ней ещё вернусь, чтобы сломать. Пока — что в этом интернете.

Триллионы слов. Веб-страницы, книги, статьи, форумы, код, субтитры, рецепты, комментарии под рецептами. Английского — больше всего, потому что интернет такой. Русского — меньше, но много. Языков, на которых говорит миллион человек, — совсем мало, и модель на них говорит соответственно.

Обратите внимание на три вещи.

Первое. Популярное представлено лучше, чем редкое. Про Пушкина написано в тысячу раз больше, чем про вашего прадеда. Значит, ручки настроены на Пушкина в тысячу раз точнее. Про прадеда они настроены никак — и мы увидим в девятой главе, что это не мешает модели про него рассказать.

Второе. В интернете люди спорят чаще, чем соглашаются. Спокойная правда — «вода мокрая» — почти не пишется, потому что зачем. Спорное пишется бесконечно. Значит, модель лучше знает споры, чем очевидности. Иногда это видно.

Третье. У мешка есть дата. В какой-то день текст перестают собирать и начинают крутить ручки. Всё, что случилось после, — для модели не существует. Спросите её про вчерашние новости, и она либо честно скажет, что не знает, либо — хуже — расскажет уверенно про позавчерашние. То, что ваш ассистент иногда всё-таки знает вчерашние новости, — это не модель. Это к ней прикрутили поиск. Про это одиннадцатая глава, и разница важнее, чем кажется.

Почему «больше» так долго работало

Есть одна вещь, которую я хочу объяснить, потому что без неё непонятно, откуда взялась вся эта гонка.

Году в 2020-м несколько исследователей заметили закономерность. Удваиваешь количество текста — ошибка падает на столько-то. Удваиваешь количество ручек — падает на столько-то ещё. Удваиваешь время вычислений — ещё. И это не «примерно». Это по линейке. Предсказуемо.

Подумайте, что это значит. Обычно в науке ты не знаешь, сработает ли следующий шаг. А тут можно было сесть, посчитать и сказать: «Если мы потратим в десять раз больше, получим вот такую модель». И получали. Индустрия впервые в истории смогла *запланировать чудо*.

Отсюда дата-центры размером с деревню. Отсюда очередь за видеокартами. Отсюда цифры с девятью нулями в новостях.

Работает ли это бесконечно? Нет. Текст в интернете конечен, и его уже почти весь собрали. Деньги тоже. Поэтому последние пару лет прогресс идёт не только за счёт «больше», но и за счёт «умнее» — и про один из таких способов, когда модель учат думать дольше перед ответом, будет десятая глава.

Девочка, которую испортили за сутки

Про то, что модель знает то, что в мешке, — вот история, которую стоит помнить.

Март 2016 года. Microsoft запускает в соцсети бота-подростка по имени Тэй. Идея простая и, как казалось, милая: она разговаривает с пользователями и учится у них. Чем больше общается, тем лучше говорит.

Утром она пишет, что люди классные. К обеду пользователи, сообразившие, как это устроено, начинают скармливать ей самое мерзкое, что могут придумать. К вечеру она пишет такое, что я не буду цитировать. Через шестнадцать часов её выключают. Microsoft извиняется.

Никто не программировал Тэй быть чудовищем. Она стала тем, что было в мешке. Мешок наполнили люди, и наполнили тем, чем захотели.

Современные модели так не учатся — на лету, от кого попало. Именно из-за Тэй. Мешок собирают заранее, чистят, и после этого ручки застывают. Но принцип остался: она знает то, что ей дали. Ни больше, ни лучше.

Люди в цепочке

Ещё один миф, который стоит убрать: что всё это делают машины без людей.

На каждом этапе есть люди. Они выбирают, какие тексты класть в мешок, а какие нет. Они читают ответы модели и говорят: этот лучше, этот хуже. Тысячи людей, часто в странах, где час работы стоит дешевле кофе, сидят и сравнивают два ответа на один вопрос.

Конец ознакомительного фрагмента.