



Машина, которая заметила закономерность

*Что такое искусственный интеллект, что он умеет и почему
он уверенно ошибается*



Технологии · 8–12 лет · 10 минут

Искусственный интеллект не думает и ничего не знает. Он ищет закономерности в огромных горах примеров и подбирает самое подходящее продолжение. Как его учат, почему он иногда врёт с невозмутимым лицом и кто на самом деле отвечает за решения.

Как компьютер научился узнавать котов

Представь задачу: научить компьютер отличать фотографию кота от фотографии собаки.

Первая мысль — объяснить правила. Усы, треугольные уши, вертикальный зрачок. Люди пробовали так делать лет тридцать подряд. Не вышло. Всегда находился кот со сложенными ушами, кот со спины, кот в темноте, кот, наполовину скрытый пакетом.

Тогда поступили иначе. Компьютеру **не стали ничего объяснять**. Ему показали миллион фотографий, у каждой написав: тут кот, тут собака. И велели самому найти, чем одни картинки отличаются от других.

Это сработало. Так и появилось то, что сегодня называют искусственным интеллектом.

Что значит «учится»

Слово «учится» тут звучит красиво, но внутри всё довольно прозаично.

Внутри модели миллионы чисел — их называют весами. Сначала они случайные. Модель смотрит на первую фотографию и наугад отвечает: «собака». Ей сообщают: неверно, это кот. Тогда числа чуть-чуть подкручиваются — так, чтобы в следующий раз ответ немного сдвинулся в сторону правильного.

Потом вторая фотография. Третья. Миллионная.

Каждая поправка крошечная. Но после миллионов поправок числа выстраиваются так, что модель почти всегда попадает в цель. Это похоже на то, как ты учишься кидать мяч в кольцо: никто не даёт тебе формулу броска, ты просто бросаешь, промахиваешься и подправляешь руку.

Важно вот что: **модель не получила правило «кот — это усы плюс уши»**. Она нашла свои собственные закономерности в пикселях, и человек обычно не может их пересказать словами. Она заметила, что так выглядели тысячи котов, — и всё.



Никто не объяснял ей, что такое кот, — она смотрела на миллион котов

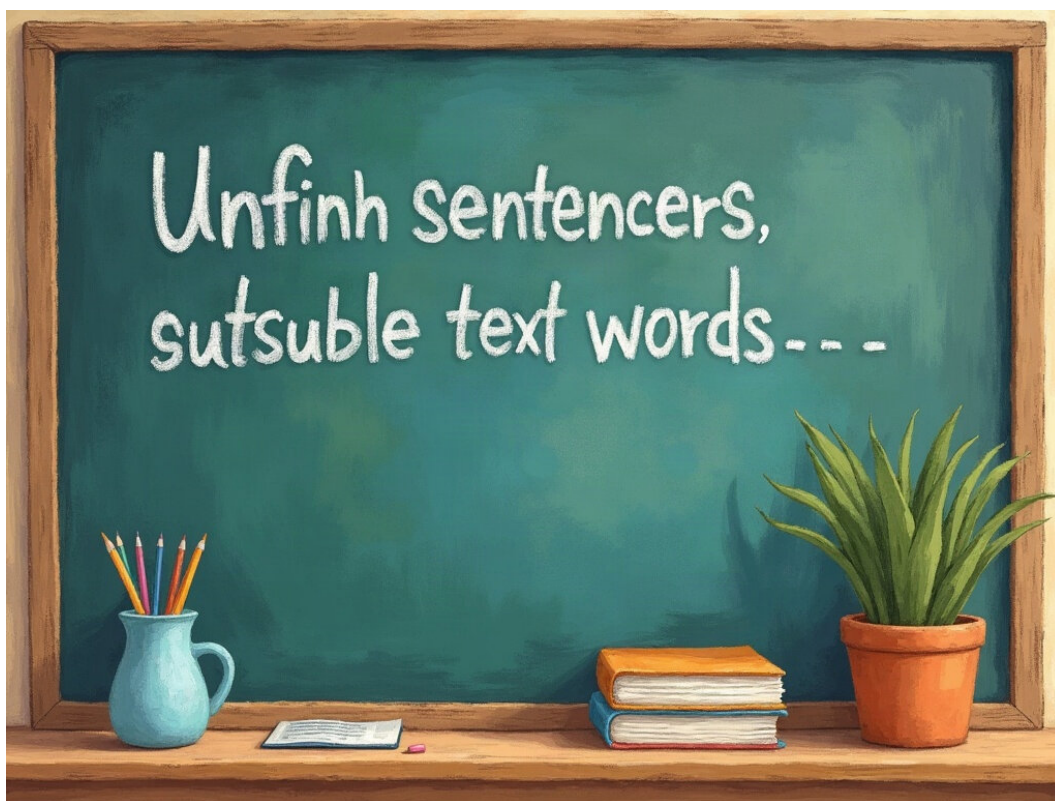
Нейросеть

Программа, устроенная как множество простых счётчиков, соединённых друг с другом слоями. Название пришло из биологии: устройство отдалённо напоминает нервные клетки мозга. Но сходство именно отдалённое — это вдохновение, а не копия. Настоящая нервная клетка устроена гораздо сложнее, и как работает мозг, учёные до сих пор знают плохо.

Как отвечает программа, которая пишет текст

Программы вроде тех, с которыми можно переписываться, устроены по тому же принципу, только учатся не на картинках, а на огромном количестве текстов: книгах, статьях, сайтах.

И задача у них ещё проще, чем кажется: **угадать, какое слово идёт следующим.**



Языковая модель просто угадывает, какое слово идёт дальше

«Кот сидел на...» — и модель прикидывает, что дальше вероятнее: «окне», «диване», «крыше». Выбирает одно, приписывает его и угадывает следующее. И так слово за словом.

Из этой простейшей задачи неожиданно получается очень много. Чтобы хорошо угадывать продолжение, приходится уловить и грамматику, и то, какие факты обычно стоят рядом друг с другом, и то, как

устроены рассуждения. Поэтому модель может ответить на вопрос, перевести текст или подсказать решение задачи.

Но стоит помнить, чем это остаётся внутри. Модель подбирает правдоподобное продолжение. Не «вспоминает правду» и не «понимает вопрос» — подбирает то, что чаще всего встречалось в похожих ситуациях.

Почему она уверенно ошибается

Отсюда главная особенность, о которой важно знать.

Модель может выдать полную чушь **точно таким же уверенным тоном**, как и правильный ответ. Она придумывает несуществующие книги, приписывает людям чужие слова, путает даты. Это называют галлюцинациями.

Причина понятна, если помнить про угадывание. Если модель не знает настоящего ответа, она всё равно должна выдать правдоподобное продолжение — и выдаёт то, что похоже на правду по форме. Она не умеет отличать «я знаю» от «я не знаю»: внутри и то и другое выглядит одинаково.

Есть и вторая беда. Модель учится на текстах и картинках, которые сделали люди, — вместе со всеми людскими ошибками и предвзятостями. Если в примерах была ерунда, модель спокойно её повторит и даже усилит.

Уверенность — не признак правоты

Уверенный тон ответа и его правильность — вещи почти не связанные. Модель формулирует выдумку так же гладко, как проверенный факт, и по одному лишь тексту отличить их невозможно. Поэтому любой важный ответ приходится проверять в источнике — это правило действует и для взрослых специалистов.

Что ИИ умеет, а что нет

Честный список выглядит так.

Получается хорошо:

- находить закономерности в огромных грудях данных, где человек утонет;
- узнавать речь, лица, дорожные знаки, буквы на старых рукописях;
- переводить с языка на язык, пересказывать длинный текст короче;
- подсказывать врачу подозрительные места на снимке;
- перебирать миллионы вариантов там, где вариантов слишком много для человека.

Не получается:

- понимать, что означают собственные ответы;
- отвечать за результат — машина ничем не рискует и ни о чём не жалеет;
- надёжно работать там, где похожих примеров не было;
- хотеть чего-либо, скучать, обижаться. Программа не выключена и не включена в смысле желаний — у неё их просто нет.

Вокруг ИИ много громких слов. С одной стороны обещают, что роботы вот-вот станут как люди. С другой — пугают, что они всех заменят. Обе картинки далеки от того, что есть на самом деле: перед нами полезный, довольно капризный инструмент, который иногда ошибается.

Кто принимает решения

Это, пожалуй, самое важное.

Программа может подсветить на снимке место, которое похоже на болезнь. Диагноз всё равно ставит врач. Программа может найти в куче документов подходящие — решение принимает человек. Программа может подсказать ответ на задачку — понять тему всё равно придётся тебе.

Так устроено не по доброте, а из осторожности. Раз модель способна ошибиться и не подать виду, за ней нужен тот, кто отвечает. В Европейском союзе это уже записано в законе: в важных вопросах — медицина, суд, приём на работу — человек не должен оказаться во власти решения, которое машина приняла сама. Похожие правила обсуждают и в других странах.



Последнее слово всё равно за человеком

Поймай модель на ошибке

Если у взрослых есть доступ к программе-помощнику, попробуй вместе такой опыт.

1. Спроси её про что-нибудь хорошо известное — например, попроси пересказать сюжет знаменитой сказки. Ответ, скорее всего, будет верным.
2. Теперь спроси о чём-то редком: например, о вашем небольшом городе, местной улице или книге малоизвестного автора.
3. Проверь ответ по надёжному источнику.
4. Обрати внимание не только на ошибки, но и на **тон**: заметил ли ты хоть по одному слову, что во втором случае модель не уверена?

Этот опыт стоит повторить с любым новым помощником, прежде чем ему доверять.

Главное в трёх строчках

- Искусственный интеллект не думает: он ищет закономерности в огромном количестве примеров и подбирает наиболее подходящее продолжение.
- Он может ошибаться уверенно и гладко, потому что не отличает знание от догадки, — поэтому важные ответы надо проверять.
- Решения всё равно принимают люди: у программы нет ни понимания, ни ответственности за последствия.

Вопросы для размышления

- Если модель училась только на текстах, написанных людьми, — может ли она узнать то, чего люди ещё не знают?
- Почему ошибку, сказанную уверенным тоном, заметить труднее, чем ошибку, сказанную неуверенно?
- Кто должен отвечать, если программа подсказала врачу неправильно: тот, кто её написал, тот, кто ею пользовался, или никто?