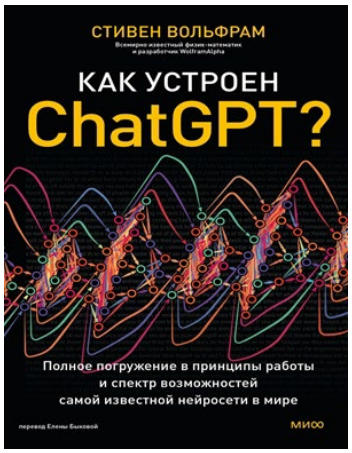
Стивен Вольфрам. Как устроен ChatGPT

Никто этого не ожидал – даже его создатели: ChatGPT ворвался в наш мир как искусственный интеллект, способный создавать тексты, словно человек. Но как именно он работает? Что происходит внутри его «ИИ-разума»? В этой книге выдающийся ученый и пионер вычислительной техники Стивен Вольфрам предлагает понятное и увлекательное объяснение устройства нейросети, основываясь на своем многолетнем научном опыте. Узнайте, как ChatGPT, объединяя новейшие технологии нейронных сетей с фундаментальными знаниями о языке и человеческом мышлении.

Стивен Вольфрам. Как устроен ChatGPT. – М.: Манн, Иванов и Фербер, 2024. – 192 с.



Купить цифровую книгу в ЛитРес, бумажную книгу в Ozon

Эта небольшая книга была опубликована на Amazon 9 марта 2023 г. А в наше время AI-платформы развиваются так быстро, что информация устаревает еще до публикации. Например, Стивен пишет, что ChatGPT плохо справляется с задачами на вычисления. И для вычислений используется платформа Wolfram|Alfa. Я попросил ChatGPT прокомментировать, как обстоят дела в марте 2025 г.

> На момент выхода книги Стивена Вольфрама в марте 2023 года ChatGPT действительно не справлялся с точными вычислениями из-за ограничений архитектуры трансформеров. LLM (большие языковые модели) обучены на текстовых данных и работают на вероятностных оценках, а не на строгих алгоритмических вычислениях. Это приводило к ошибкам при сложных расчетах, особенно при работе с длинной арифметикой, точными дробями, математическим моделированием и символьными выражениями.
>
> Как обстоят дела сейчас?
>
> - Сейчас в ChatGPT используются специализированные модули для вычислений, например:
> - Встроенный интерпретатор Python — позволяет выполнять точные числовые расчеты, строить графики, анализировать данные.
> - Платформа Wolfram Alpha (в некоторых версиях, например, в Pro) — обеспечивает доступ к мощной математической системе, включая символьные вычисления, решение уравнений, работу с матрицами, статистику и физические расчеты.
> - Улучшенные модели LLM — современные версии (например, GPT-4) стали лучше справляться с арифметикой на уровне базовых и средних вычислений, хотя для высокой точности все еще лучше использовать специализированные модули.

В книге описывается ряд моментов, которые хорошо известны, и я не стану их здесь приводить. Тем более, что автор описывает их довольно неуклюже, сопровождая неудачными рисунками. В частности: ChatGPT просто добавляет по одному слову за раз, про вероятностный выбор следующего слова, откуда берутся вероятности, влияние температуры на креативность/строгость модели, распознавание изображений, виды функций активации…

Один фрагмент мне понравился больше остальных, и я приведу его полностью.

Достаточно большая сеть может сделать все что угодно – так ведь?

Возможности программ наподобие ChatGPT выглядят впечатляюще. Можно даже подумать, что если продолжать тренировать нейронные сети, то в конце концов они научатся делать все. И когда речь идет о вещах, которые легко доступны человеческому мышлению, то вполне возможно, что все обстоит именно так. Но урок, который можно вынести из последних нескольких сотен лет развития науки, заключается в том, что есть вещи, которые могут быть поняты с помощью формальных процессов, но недоступны человеческому мышлению.

Одним из ярких примеров является нетривиальная математика. Но самое важное — это вычисления. И в конечном счете проблема заключается в феномене вычислительной несводимости. Есть вычисления, для выполнения которых, как можно было бы подумать, потребуется много шагов, но которые на самом деле могут быть сведены к чему-то совершенно простому. Однако открытие вычислительной несводимости подразумевает, что это не всегда работает. Так, существуют процессы (вероятно, подобные показанному ниже), в которых для определения того, что происходит, требуется отслеживать каждый шаг вычислений:

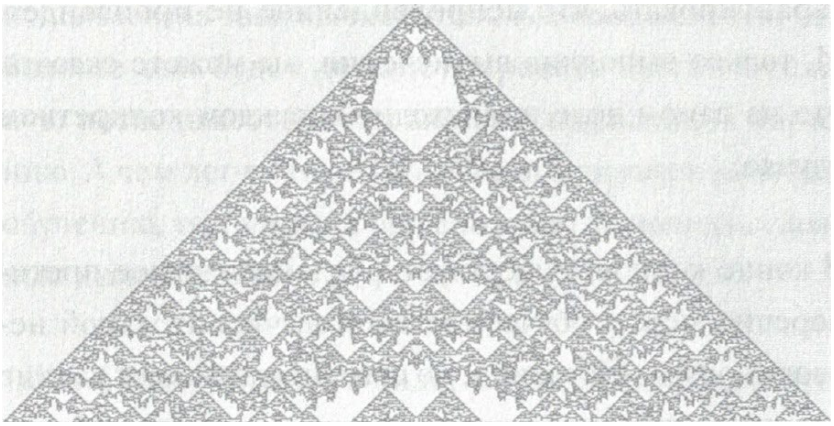


Рис. 1. Видимо, это фрактал

Знания о том, как работает наш мозг, предположительно позволяют нам избежать вычислительной несводимости. Чтобы вычислять в уме, требуются особые усилия, и на практике невозможно продумать этапы работы любой нетривиальной программы.

Конечно, для этого у нас есть компьютеры. А с помощью компьютеров мы можем легко выполнять длительные, не поддающиеся вычислению задачи. И ключевым моментом является то, что здесь, как правило, нет кратчайшего пути.

Да, мы можем запомнить множество конкретных примеров из какой-то вычислительной системы. И возможно, даже можем увидеть некоторые (вычислительно сводимые) закономерности, которые позволят нам сделать небольшое обобщение. Но дело в том, что вычислительная несводимость означает невозможность гарантировать, что непредвиденное не произойдет. И, только выполнив вычисления, вы можете сказать, что на самом деле происходит в каждом конкретном случае.

В конце концов, существует фундаментальное противоречие между обучаемостью и вычислительной несводимостью. Обучение, по сути, предполагает сжатие данных за счет использования закономерностей. Но вычислительная несводимость подразумевает, что в конечном счете есть предел количеству существующих закономерностей.

С практической точки зрения можно представить превращение небольших вычислительных устройств — таких, как клеточные автоматы или машины Тьюринга, — в обучаемые системы вроде нейронных сетей. И действительно, такие устройства могут служить хорошими инструментами для нейронной сети: например, Wolfram|Alpha может быть хорошим инструментом для ChatGPT. Но вычислительная несводимость подразумевает, что нельзя «проникнуть» внутрь этих устройств и заставить их обучаться.

Другими словами, существует окончательный компромисс между возможностями и обучаемостью: чем сильнее вы хотите, чтобы система по-настоящему использовала свои вычислительные возможности, тем больше она будет демонстрировать вычислительную несводимость и тем меньше поддаваться обучению. А чем легче что-то поддается фундаментальному обучению, тем меньше оно способно выполнять сложные вычисления.

(Для ChatGPT в том виде, в каком он существует сейчас, ситуация еще более экстремальная, потому что нейронная сеть, используемая для генерации каждого токена вывода, является чистой нейронной сетью с прямой связью — feed-forward, то есть сетью без циклов, — и, следовательно, не имеет возможности выполнять какие-либо вычисления с нетривиальным «потоком управления».)

Конечно, можно спросить, действительно ли важно уметь выполнять несводимые вычисления. На протяжении большей части человеческой истории это не было так уж важно. Но современный технологичный мир построен на инженерии, которая использует математические вычисления, а также — все чаще — более общие вычисления. Если мы посмотрим на мир природы, то увидим, что он полон несводимых вычислений, и мы постепенно начинаем понимать, как их имитировать и использовать в технологических целях.

Нейронная сеть, безусловно, способна замечать те виды закономерностей в мире природы, которые и мы можем легко заметить с помощью «невооруженного человеческого мышления». Но если мы хотим разобраться в вещах, которые относятся к математике или вычислительной науке, то нейронная сеть не сможет этого сделать, если только не использует в качестве инструмента «обычную» вычислительную систему.

Во всем этом есть что-то непонятное. Раньше существовало множество задач, включая написание текстов, которые считались фундаментально слишком сложными для компьютеров. И сегодня, когда мы видим, что технологии вроде ChatGPT справляются с этими задачами, мы склонны думать, что все дело в том, что компьютеры стали значительно мощнее — в частности тех, которые были способны делать что-то вроде вычисления поведения вычислительных систем, таких как клеточные автоматы.

Но это не совсем правильный вывод. Вычислительно несводимые процессы по-прежнему являются таковыми и по-прежнему принципиально сложны для компьютеров — даже если компьютеры могут легко вычислить отдельные этапы. Вывод, который мы должны сделать, заключается в том, что задачи (например, написание текстов), которые мы могли выполнять, но считали, что компьютеру они не под силу, на самом деле с точки зрения вычислений проще, чем мы предполагали.

То есть причина, по которой нейронная сеть может хорошо писать тексты, заключается в том, что написание текстов оказывается более простой с вычислительной точки зрения задачей, чем мы думали. И в некотором смысле это приближает нас к формулированию теории о том, как нам, людям, удается создавать тексты и в целом работать с языком.

Если бы у нас была достаточно большая нейронная сеть, она могла бы делать все, что под силу простым людям. Но она бы все равно не смогла охватить все то, на что способен мир природы в целом — или что могут делать инструменты, которые мы создали, опираясь на законы природы. Именно применение этих инструментов — как практических, так и концептуальных — позволило нам за последние столетия выйти за границы того, что доступно чистой человеческой мысли без посторонней помощи, и использовать для человеческих целей больше того, что существует в физической и вычислительной вселенной.

***

Эмбеддинг можно объяснить как способ представить сущность чего-либо с помощью массива чисел, где подобные объекты представлены близлежащими числами.

***

Успех ChatGPT демонстрирует важный научный факт: осмысленный человеческий язык оказывается гораздо проще и структурированнее, чем принято считать, и подчиняется сравнительно несложным правилам.