WIKIPEDIA
The Free Encyclopedia

Большая языковая модель

Большая **языковая модель** (**LLM**) — это тип модели машинного обучения , разработанный для задач обработки естественного языка , таких как генерация языка . LLM — это языковые модели со многими параметрами, которые обучаются с помощью самоконтролируемого обучения на большом объеме текста.

Самые большие и наиболее способные LLM — это генеративные предварительно обученные трансформаторы (GPT). Современные модели могут быть точно настроены для конкретных задач или направляться оперативным проектированием . [1] Эти модели приобретают предсказательную силу относительно синтаксиса , семантики и онтологий [2] , присущих корпусам человеческого языка, но они также наследуют неточности и предубеждения, присутствующие в данных, на которых они обучаются. [3]

История

До 2017 года существовало несколько языковых моделей, которые были большими по сравнению с доступными тогда возможностями. В 1990-х годах модели выравнивания IBM стали пионерами статистического моделирования языка. Сглаженная модель n-грамм в 2001 году, обученная на 0,3 миллиарда слов, достигла самого современного уровня сложности на тот момент. [4] В 2000-х годах, когда использование Интернета стало широко распространенным, некоторые исследователи создали языковые наборы данных в масштабе Интернета («Сеть как корпус» [5]), на которых они обучали статистические языковые модели. [6][7] В 2009 году в большинстве задач обработки языка статистические языковые модели доминировали над символическими языковыми моделями, поскольку они могли с пользой обрабатывать большие наборы данных. [8]

Обучение вычислений заметных больших моделей в FLOPs против даты публикации за период 2010-2024. Для общих заметных моделей (вверху слева), пограничных моделей (вверху справа), ведущих языковых моделей (внизу слева) и ведущих моделей в ведущих компаниях (внизу справа). Большинство этих моделей являются языковыми моделями.

После того, как нейронные сети стали доминировать в обработке изображений около 2012 года, [9] они также были применены к языковому моделированию. Google преобразовал свой сервис перевода в Neural Machine Translation в 2016 году. Поскольку это предшествовало существованию transformers , это было сделано с помощью глубоких LSTM -сетей seq2seq .

На конференции NeurIPS 2017 года исследователи Google представили архитектуру трансформатора в своей знаковой статье « Внимание — это все, что вам нужно ». Целью этой статьи было усовершенствование технологии seq2seq 2014 года [10] и она была основана в основном на механизме внимания, разработанном Багданау и др. в 2014 году. [11] В

следующем году, в 2018 году, был представлен BERT , который быстро стал «повсеместным». [12] Хотя оригинальный трансформатор имеет как блоки кодера, так и декодера, BERT представляет собой модель только для кодера. Академическое и исследовательское использование BERT начало снижаться в 2023 году после быстрого улучшения возможностей моделей только для декодера (таких как GPT) решать задачи с помощью подсказок . [13]

Обучение вычислений заметных больших моделей ИИ в FLOPs против даты публикации за период 2017-2024. Большинство больших моделей являются языковыми моделями или мультимодальными моделями с языковой емкостью.

Хотя GPT-1 , работающий только с декодером, был представлен в 2018 году, именно GPT-2 в 2019 году привлек всеобщее внимание, поскольку OpenAI сначала посчитала его слишком мощным для публичного выпуска из-за страха злонамеренного использования. [14] GPT-3 в 2020 году пошла на шаг дальше и с 2024 года доступна только через API без предложения загрузки модели для локального выполнения. Но именно ориентированный на потребителя браузерный ChatGPT 2022 года захватил воображение широких слоев населения и вызвал некоторую шумиху в СМИ и онлайн-шумиху. [15] GPT-4 2023 года хвалили за его повышенную точность и называли «святым Граалем» за его мультимодальные возможности. [16] OpenAI не раскрыла высокоуровневую архитектуру и количество параметров GPT-4. Выпуск ChatGPT привел к росту использования LLM в нескольких исследовательских подобластях компьютерных наук, включая робототехнику, разработку программного обеспечения и работу по общественному влиянию. [17] В 2024 году OpenAI выпустила модель рассуждений OpenAI o1 , которая генерирует длинные цепочки мыслей, прежде чем вернуть окончательный ответ.

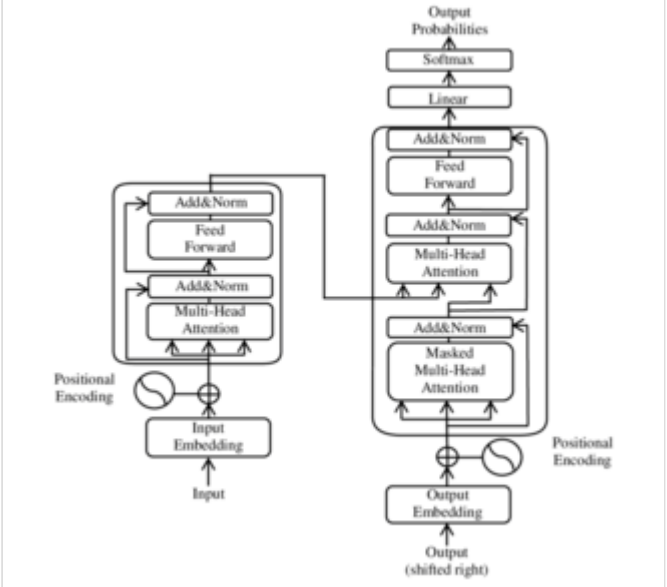


Иллюстрация основных компонентов модели трансформатора из оригинальной статьи, где слои были нормализованы после (а не до) многоголового внимания

Конкурирующие языковые модели по большей части пытались сравняться с серией GPT, по крайней мере, с точки зрения количества параметров. [18]

С 2022 года модели с доступными исходниками набирают популярность, особенно в первую очередь с BLOOM и LLaMA , хотя обе имеют ограничения по области использования. Модели Mistral AI Mistral 7B и Mixtral 8x7b имеют более разрешительную лицензию Apache . В январе 2025 года DeepSeek выпустила DeepSeek R1, модель с открытым весом на 671 миллиард параметров, которая работает сопоставимо с OpenAI o1, но при гораздо меньших затратах. [19]

С 2023 года многие LLM были обучены быть мультимодальными , имея возможность также обрабатывать или генерировать другие типы данных, такие как изображения или аудио. Эти LLM также называются большими мультимодальными моделями (LMM). [20]

По состоянию на 2024 год, все самые большие и самые эффективные модели основаны на архитектуре трансформатора. Некоторые недавние реализации основаны на других архитектурах, таких как варианты рекуррентных нейронных сетей и Mamba (модель пространства состояний). [21][22][23]

Предварительная обработка набора данных

Токенизация

Поскольку алгоритмы машинного обучения обрабатывают числа, а не текст, текст должен быть преобразован в числа. На первом этапе выбирается словарь, затем целочисленные индексы произвольно, но уникально назначаются каждой записи словаря, и, наконец, встраивание связывается с целочисленным индексом. Алгоритмы включают кодирование пар байтов (BPE) и WordPiece . Существуют также специальные токены, служащие в качестве управляющих символов , например, `[MASK]`для замаскированного токена (используемого в BERT) и `[UNK]`(«неизвестно») для символов, не встречающихся в словаре. Кроме того, некоторые специальные символы используются для обозначения специального форматирования текста. Например, «Ġ» обозначает предшествующий пробел в RoBERTa и GPT. «##» обозначает продолжение предшествующего слова в BERT. [24]

Например, токенизатор BPE, используемый GPT-3 (Legacy), будет разделен `tokenizer: texts -> series of numerical "tokens"`следующим образом:

токенизер: тексты ->ряд из числовой "тхорошоэнс"

Токенизация также сжимает наборы данных. Поскольку LLM обычно требуют, чтобы входные данные были массивом , который не является зазубренным , более короткие тексты должны быть «дополнены» до тех пор, пока они не будут соответствовать длине самого длинного. Сколько токенов в среднем требуется для одного слова, зависит от языка набора данных. [25][26]

БПЭ

В качестве примера рассмотрим токенизатор, основанный на кодировании пар байтов. На первом этапе все уникальные символы (включая пробелы и знаки препинания) обрабатываются как начальный набор n -грамм (т. е. начальный набор уни-грамм). Последовательно наиболее частая пара смежных символов объединяется в би-грамму, и все экземпляры пары заменяются ею. Все вхождения смежных пар (ранее объединенных) n -грамм, которые чаще всего встречаются вместе, затем снова объединяются в еще более длинную n -грамму, пока не будет получен словарь заданного размера (в случае GPT-3 размер составляет 50257). [27] После обучения токенизатора он может токенизировать любой текст, если только он не содержит символов, не встречающихся в начальном наборе уни-грамм. [28]

Проблемы

Словарь токенов, основанный на частотах, извлеченных из преимущественно английских корпусов, использует как можно меньше токенов для среднего английского слова. Однако среднее слово на другом языке, закодированное таким оптимизированным для английского языка токенизатором, разбивается на неоптимальное количество токенов. Токенизатор GPT-2 может использовать до 15 раз больше токенов на слово для некоторых языков, например, для языка шанс из Мьянмы . Даже более распространенные языки, такие как португальский и немецкий, имеют «премию в 50%» по сравнению с английским. [26]

Жадная токенизация также вызывает тонкие проблемы с завершением текста. [29]

Очистка набора данных

В контексте обучения LLM наборы данных обычно очищаются путем удаления некачественных, дублированных или токсичных данных. [30] Очищенные наборы данных могут повысить эффективность обучения и привести к улучшению производительности в дальнейшем. [31] [32] Обученный LLM может использоваться для очистки наборов данных для обучения следующего LLM. [33]

С ростом доли контента, сгенерированного LLM в Интернете, очистка данных в будущем может включать в себя фильтрацию такого контента. Контент, сгенерированный LLM, может представлять проблему, если он похож на человеческий текст (что затрудняет фильтрацию), но имеет более низкое качество (снижая производительность моделей, обученных на нем). [34]

Синтетические данные

Обучение самых больших языковых моделей может потребовать больше лингвистических данных, чем доступно естественным образом, или что естественные данные недостаточного качества. В этих случаях могут использоваться синтетические данные. Серия LLM Phi от Microsoft обучается на данных, подобных учебникам, сгенерированных другим LLM. [35]

Обучение и архитектура

Подкрепление обучения на основе обратной связи от человека

Обучение с подкреплением на основе обратной связи с человеком (RLHF) с помощью алгоритмов, таких как оптимизация проксимальной политики , используется для дальнейшей тонкой настройки модели на основе набора данных о человеческих предпочтениях. [36]

Инструкция по настройке

Используя подходы «самообучения», LLM смогли загружать правильные ответы, заменяя любые наивные ответы, начиная с человеческих исправлений нескольких случаев. Например, в инструкции «Напишите эссе об основных темах, представленных в *Гамлете* »,

первоначальное наивное завершение может быть «Если вы отправите эссе после 17 марта, ваша оценка будет снижена на 10% за каждый день задержки», в зависимости от частоты этой текстовой последовательности в корпусе. [37]

Смесь экспертов

Самый большой LLM может быть слишком дорогим для обучения и использования напрямую. Для таких моделей может быть применена смесь экспертов (MoE), направление исследований, проводимых исследователями Google с 2017 года для обучения моделей, достигающих до 1 триллиона параметров. [38] [39] [40]

Инженерное обеспечение подсказок, механизм внимания и контекстное окно

Большинство результатов, которые ранее можно было достичь только путем (дорогостоящей) тонкой настройки, могут быть достигнуты посредством оперативной разработки , хотя и ограничены рамками одного разговора (точнее, ограничены рамками контекстного окна). [41]

Чтобы выяснить, какие токены релевантны друг другу в рамках контекстного окна, механизм внимания вычисляет «мягкие» веса для каждого токена, точнее, для его внедрения, используя несколько головок внимания, каждая со своей собственной «релевантностью» для вычисления своих собственных мягких весов. Например, маленькая (т.е. размером 117 млн параметров) модель GPT-2 имела двенадцать головок внимания и контекстное окно всего из 1 тыс. токенов. [43] В своей средней версии она имеет 345 млн параметров и содержит 24 слоя, каждый с 12 головками внимания. Для обучения с градиентным спуском использовался размер пакета 512. [28]

Самые большие модели, такие как Gemini 1.5 от Google , представленная в феврале 2024 года, могут иметь контекстное окно размером до 1 миллиона (контекстное окно размером 10 миллионов также было «успешно протестировано»). [44] Другие модели с большими контекстными окнами включают Claude 2.1 от Anthropic с контекстным окном размером до 200 тыс. токенов. [45] Обратите внимание, что этот максимум относится к количеству входных токенов, а максимальное количество выходных токенов отличается от входного и часто меньше. Например, модель GPT-4 Turbo имеет максимальный выход в 4096 токенов. [46]

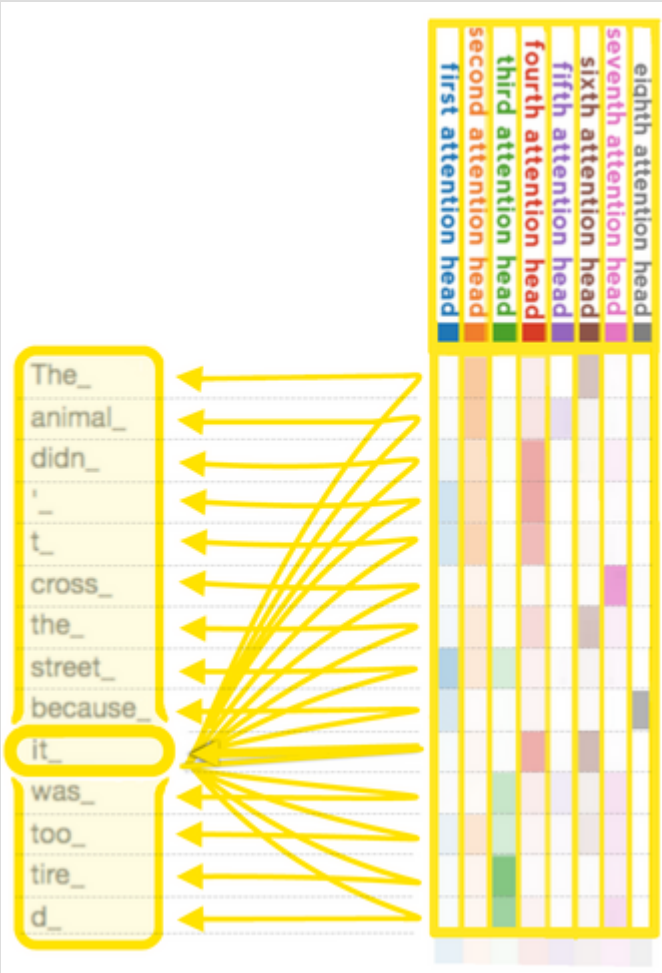
Длина разговора, которую модель может учесть при генерации следующего ответа, также ограничена размером контекстного окна. Если длина разговора, например, с ChatGPT , больше, чем его контекстное окно, то при генерации следующего ответа учитываются только части внутри контекстного окна, или модели необходимо применить какой-то алгоритм для суммирования слишком далеких частей разговора.

Недостатки увеличения контекстного окна включают более высокие вычислительные затраты и возможное размывание фокуса на локальном контексте, в то время как уменьшение может привести к тому, что модель упустит важную зависимость на дальнем расстоянии. Их балансировка — это вопрос экспериментирования и соображений, специфичных для домена.

Модель может быть предварительно обучена либо для прогнозирования продолжения сегмента, либо для прогнозирования того, чего не хватает в сегменте, учитывая сегмент из ее обучающего набора данных. [47] Это может быть либо

- авторегрессионный (т.е. предсказывающий, как продолжится сегмент, как это делают GPT): например, если задан сегмент «Я люблю поесть», модель предскажет «мороженое» или «суши».
- « замаскированный » (т.е. заполняющий отсутствующие части сегмента, как это делает «BERT» [48]): например, если задан сегмент «Мне нравится добавлять [__] [__]сливки», модель предсказывает, что «есть» и «лед» отсутствуют.

Модели могут обучаться на вспомогательных задачах, которые проверяют их понимание распределения данных, таких как предсказание следующего предложения (NSP), в котором представлены пары предложений, и модель должна предсказать, появляются ли они последовательно в обучающем корпусе. [48] Во время обучения потеря регуляризации также используется для стабилизации обучения. Однако потеря регуляризации обычно не используется во время тестирования и оценки.



Когда каждая голова вычисляет, согласно своим собственным критериям, насколько другие токены релевантны токену «it_», обратите внимание, что вторая голова внимания, представленная вторым столбцом, больше всего фокусируется на первых двух строках, то есть на токенах «The» и «animal», в то время как третий столбец больше всего фокусируется на нижних двух строках, то есть на «tired», который был токенизирован в два токена. [42]

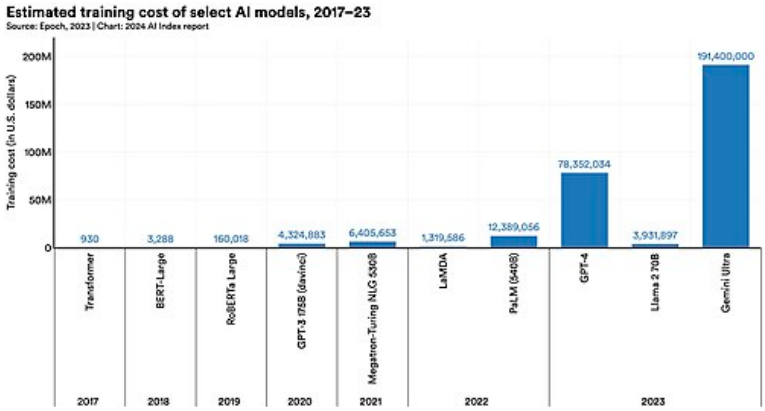
Инфраструктура

Для обучения самых больших моделей необходима существенная инфраструктура. [49] [50] [51]

Стоимость обучения

Квалификатор «большой» в «большой языковой модели» по своей сути неопределен, поскольку не существует определенного порога для количества параметров, необходимых для квалификации как «большой». Со временем то, что ранее считалось «большим», может измениться. GPT-1 2018 года обычно считается первым LLM, хотя в нем всего 0,117 миллиарда параметров. Тенденция к более крупным моделям видна в списке больших языковых моделей .

По мере развития технологий большие суммы инвестировались во все более крупные модели. Например, обучение GPT-2 (т. е. модели с 1,5 миллиардами параметров) в 2019 году стоило 50 000 долларов, тогда как обучение PaLM (т. е. модели с 540 миллиардами параметров) в 2022 году стоило 8 миллионов долларов, а Megatron-Turing NLG 530B (в 2021 году) стоило около 11 миллионов долларов. [52]



Для LLM на основе Transformer стоимость обучения намного выше стоимости вывода. Обучение на одном токене стоит 6 FLOP на параметр, тогда как вывод на одном токене стоит 1-2 FLOP на параметр. [53]

Использование инструмента

Существуют определенные задачи, которые в принципе не может решить ни один LLM, по крайней мере, без использования внешних инструментов или дополнительного программного обеспечения. Примером такой задачи является ответ на ввод пользователя '354 * 139 = ', при условии, что LLM еще не столкнулся с продолжением этого вычисления в своем обучающем корпусе. В таких случаях LLM необходимо прибегнуть к запуску программного кода, который вычисляет результат, который затем может быть включен в его ответ. : Другой пример: "Сколько сейчас времени? Это ", где отдельному программному интерпретатору необходимо выполнить код, чтобы получить системное время на компьютере, чтобы LLM мог включить его в свой ответ. [54] [55] Эта базовая стратегия может быть усложнена с помощью нескольких попыток сгенерированных программ и других стратегий выборки. [56]

Обычно, чтобы LLM мог использовать инструменты, его необходимо настроить для использования инструментов. Если количество инструментов конечно, то настройка может быть выполнена только один раз. Если количество инструментов может произвольно увеличиваться, как в случае с онлайн- сервисами API , то LLM можно настроить так, чтобы он мог читать документацию API и правильно вызывать API. [57] [58]

Генерация дополненного поиска (RAG) — это еще один подход, который улучшает LLM, интегрируя их с системами поиска документов . При наличии запроса вызывается извлекатель документов для извлечения наиболее релевантных документов. Обычно это делается путем кодирования запроса и документов в векторы, а затем поиска документов с векторами (обычно хранящимися в векторной базе данных), наиболее похожими на вектор запроса. Затем LLM генерирует вывод на основе как запроса, так и контекста, включенного в извлеченные документы. [59]

Агентство

LLM, как правило, сам по себе не является автономным агентом , поскольку ему не хватает способности взаимодействовать с динамическими средами, вспоминать прошлое поведение и планировать будущие действия, но его можно превратить в такового путем интеграции таких модулей, как профилирование, память, планирование и действие. [60]

Шаблон ReAct , портманто от "Reason + Act", конструирует агента из LLM, используя LLM в качестве планировщика. LLM предлагается "думать вслух". В частности, языковая модель предлагается с текстовым описанием среды, целью, списком возможных действий и записью действий и наблюдений на данный момент. Она генерирует одну или несколько мыслей перед созданием действия, которое затем выполняется в среде. [61] Лингвистическое описание среды, предоставленное планировщику LLM, может быть даже кодом LaTeX статьи, описывающей среду. [62]

В методе DEPS («Опишите, объясните, спланируйте и выберите») LLM сначала подключается к визуальному миру с помощью описаний изображений, затем ему предлагается разработать планы для сложных задач и поведения на основе его предварительно обученных знаний и получаемой им обратной связи из окружающей среды. [63]

Метод рефлексии [64] создает агента, который обучается на протяжении нескольких эпизодов. В конце каждого эпизода LLM получает запись эпизода и побуждается придумать «усвоенные уроки», которые помогут ему лучше выступить в последующем эпизоде. Эти «усвоенные уроки» предоставляются агенту в последующих эпизодах.

Поиск по дереву Монте-Карло может использовать LLM в качестве эвристики развертывания. Когда программная модель мира недоступна, LLM также может быть предложено с описанием среды, чтобы действовать как модель мира. [65]

Для открытого исследования LLM может использоваться для оценки наблюдений за их «интересностью», которая может использоваться в качестве сигнала вознаграждения для руководства обычным (не LLM) агентом обучения с подкреплением. [66] В качестве альтернативы он может предлагать все более сложные задачи для изучения учебной программы . [67] Вместо вывода отдельных действий планировщик LLM может также конструировать «навыки» или функции для сложных последовательностей действий. Навыки могут быть сохранены и позже вызваны, что позволяет повысить уровень абстракции в планировании. [67]

Агенты, работающие на LLM, могут сохранять долгосрочную память о своих предыдущих контекстах, и память может быть извлечена таким же образом, как и Retrieval Augmented Generation. Несколько таких агентов могут взаимодействовать социально. [68]

Сжатие

Обычно LLM обучаются с помощью чисел с плавающей точкой одинарной или половинной точности (float32 и float16). Один float16 имеет 16 бит или 2 байта, поэтому один миллиард параметров требует 2 гигабайта. Самые большие модели обычно имеют 100 миллиардов

параметров, требующих 200 гигабайт для загрузки, что выводит их за пределы диапазона большинства потребительских электронных устройств. [69]

Квантование после обучения [70] направлено на уменьшение требуемого пространства за счет снижения точности параметров обученной модели, при этом сохраняя большую часть ее производительности. [71] [72] Простейшая форма квантования просто усекает все числа до заданного количества бит. Ее можно улучшить, используя другую кодовую книгу квантования на каждый слой. Дальнейшее улучшение может быть достигнуто путем применения различной точности к различным параметрам, с более высокой точностью для особенно важных параметров («веса выбросов»). [73] Смотрите визуальное руководство по квантованию от Мартена Гроотендорста [74] для визуального изображения.

В то время как квантованные модели обычно заморожены, и только предварительно квантованные модели могут быть точно настроены, квантованные модели все еще могут быть точно настроены. [75]

Мультимодальность

Мультимодальность означает «наличие нескольких модальностей», а «модальность» относится к типу ввода или вывода, например, видео, изображение, аудио, текст, проприорецепция и т. д. [76] Было создано много моделей ИИ, специально обученных воспринимать одну модальность и выводить другую модальность, например, AlexNet для преобразования изображения в метку, [77] визуальный ответ на вопрос для преобразования изображения в текст, [78] и распознавание речи для преобразования речи в текст.

Распространенный метод создания мультимодальных моделей из LLM — это «токенизация» выходных данных обученного кодировщика. Конкретно, можно построить LLM, который может понимать изображения следующим образом: взять обученный LLM и обученный кодировщик изображений E. Сделайте небольшой многослойный персептрон f, так что для любого изображения y, постобработанный вектор $f(E(y))$ имеет те же размеры, что и закодированный токен. Это «токен изображения». Затем можно чередовать текстовые токены и токены изображения. Затем составная модель тонко настраивается на наборе данных изображение-текст. Эту базовую конструкцию можно применять с большей сложностью для улучшения модели. Кодер изображения может быть заморожен для повышения стабильности. [79]

Flamingo продемонстрировал эффективность метода токенизации, настроив пару предварительно обученной языковой модели и кодировщика изображений для более эффективного ответа на визуальные вопросы, чем модели, обученные с нуля. [80] Модель Google PaLM была настроена в мультимодальную модель PaLM-E с использованием метода токенизации и применена к роботизированному управлению. [81] Модели LLaMA также были преобразованы в мультимодальные с использованием метода токенизации, чтобы разрешить ввод изображений, [82] и видеовходов. [83]

GPT-4 может использовать как текст, так и изображение в качестве входных данных [84] (хотя компонент зрения не был представлен публике до GPT-4V [85]); Gemini от Google DeepMind также является мультимодальным. [86] Mistral представила свою собственную мультимодельную модель Pixtral 12B в сентябре 2024 года. [87]

Рассуждение

В конце 2024 года в разработке LLM появилось новое направление с моделями, специально разработанными для сложных задач рассуждения. Эти «модели рассуждения» были обучены тратить больше времени на генерацию пошаговых решений, прежде чем предоставлять окончательные ответы, аналогично процессам решения проблем человеком. [88] OpenAI представила эту тенденцию со своей моделью o1 в сентябре 2024 года, за которой последовала o3 в декабре 2024 года. Эти модели показали значительные улучшения в задачах по математике, естественным наукам и кодированию по сравнению с традиционными LLM. Например, на задачах отборочного экзамена Международной олимпиады по математике GPT-4o достигла точности 13%, а o1 — 83%. [88] [89] В январе 2025 года китайская компания DeepSeek выпустила DeepSeek-R1, модель рассуждения с открытым весом и 671 миллиардом параметров, которая достигла сопоставимой производительности с o1 от OpenAI, при этом будучи значительно более экономичной в эксплуатации. В отличие от фирменных моделей OpenAI, открытая природа веса DeepSeek-R1 позволила исследователям изучать и развивать алгоритм, хотя его данные обучения оставались закрытыми. [90] Эти модели рассуждений обычно требуют больше вычислительных ресурсов на запрос по сравнению с традиционными LLM, поскольку они выполняют более обширную обработку для пошаговой проработки проблем. Однако они показали превосходные возможности в областях, требующих структурированного логического мышления, таких как математика, научные исследования и компьютерное программирование. [89]

Попытки уменьшить или компенсировать галлюцинации использовали автоматизированное мышление , RAG (генерацию с дополненной поисковой памятью), тонкую настройку и другие методы. [91]

Характеристики

Законы масштабирования

Эффективность LLM после предварительной подготовки во многом зависит от:

- стоимость предварительной подготовки C (общий объем использованных вычислений),
- размер самой искусственной нейронной сети , например, количество параметров N (т.е. количество нейронов в слоях, количество весов между ними и смещения),
- размер его набора данных для предварительного обучения (т.е. количество токенов в корпусе, D).

«Законы масштабирования» — это эмпирические статистические законы , которые предсказывают производительность LLM на основе таких факторов. Один конкретный закон масштабирования (« масштабирование Шиншиллы ») для LLM, авторегрессионно обученного в течение одной эпохи, с графиком скорости обучения в логарифмическом

масштабе , гласит, что: [92]

$$\begin{cases} C = C_0 ND \\ L = \frac{A}{N^\alpha} + \frac{B}{D^\beta} + L_0 \end{cases}$$

где переменные

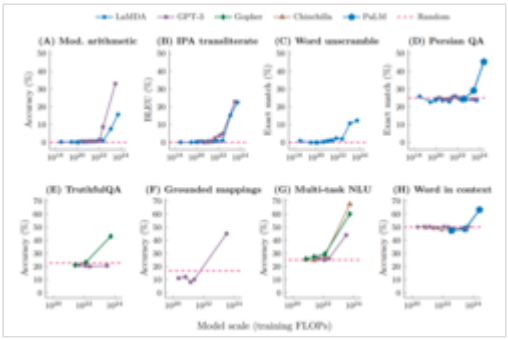
- Cстоимость обучения модели в FLOP .
- N— количество параметров в модели.
- D— количество токенов в обучающем наборе.
- L— это средняя отрицательная логарифмическая потеря правдоподобия на токен (nats /token), достигнутая обученным LLM на тестовом наборе данных.

и статистические гиперпараметры

- $C_0 = 6$, что означает, что обучение на одном токене стоит 6 FLOP на параметр. Обратите внимание, что стоимость обучения намного выше стоимости вывода, где для вывода на одном токене требуется 1-2 FLOP на параметр. [53]
- $\alpha = 0.34, \beta = 0.28, A = 406.4, B = 410.7, L_0 = 1.69$

Новые способности

Производительность более крупных моделей при выполнении различных задач, при построении в логарифмическом масштабе, выглядит как линейная экстраполяция производительности, достигнутой более мелкими моделями. Однако эта линейность может прерываться " перерывами " [93] в законе масштабирования, где наклон линии резко меняется, и где более крупные модели приобретают "возникающие способности". [41] [94] Они возникают из сложного взаимодействия компонентов модели и не запрограммированы или не спроектированы явно. [95]



В точках, называемых разрывами , [93] линии меняют свой наклон, появляясь на линейно-логарифмическом графике как ряд линейных сегментов, соединенных дугами.

Более того, недавние исследования продемонстрировали, что системы ИИ, включая большие языковые модели, могут использовать эвристические рассуждения, родственные человеческому познанию. Они балансируют между исчерпывающей логической обработкой и использованием когнитивных сокращений (эвристики), адаптируя свои стратегии рассуждений для оптимизации между точностью и усилиями. Такое поведение соответствует принципам ресурсно-рационального человеческого познания, как обсуждается в классических теориях ограниченной рациональности и теории двойного процесса. [96]

Одной из возникающих способностей является контекстное обучение на основе демонстрационных примеров. [97] Контекстное обучение подразумевает выполнение таких задач, как:

- сообщенная арифметика
- расшифровка международного фонетического алфавита

- расшифровка букв слова
- устранение неоднозначности наборов данных «слово в контексте» [41] [98] [99]
- преобразование пространственных слов
- основные направления (например, ответ «северо-восток» в ответ на сетку 3x3 из 8 нулей и 1 в правом верхнем углу), цветовые обозначения, представленные в тексте. [100]
- подсказка цепочки мыслей : в исследовательской работе 2022 года подсказка цепочки мыслей улучшила производительность только для моделей, у которых было не менее 62B. Меньшие модели работают лучше, когда их подсказывали отвечать немедленно, без цепочки мыслей. [101]
- выявление оскорбительного содержания в абзацах на хинглише (комбинация хинди и английского) и создание аналогичного английского эквивалента пословиц на языке суахили . [102]

Шеффер *и др.* утверждают, что возникающие способности не приобретаются непредсказуемо, а приобретаются предсказуемо в соответствии с законом плавного масштабирования . Авторы рассмотрели игрушечную статистическую модель решения LLM вопросов с множественным выбором и показали, что эта статистическая модель, модифицированная для учета других типов задач, применима и к этим задачам. [103]

Позволятьxбыть числом параметров count, иyбыть производительностью модели.

- Когда$y = \text{average } \Pr(\text{correct token})$, затем$(\log x, y)$представляет собой экспоненциальную кривую (до того, как она достигнет плато в точке 1), которая выглядит как возникновение.
- Когда$y = \text{average } \log(\Pr(\text{correct token}))$, то$(\log x, y)$График представляет собой прямую линию (до достижения плато в нуле), что не похоже на возникновение.
- Когда$y = \text{average } \Pr(\text{the most likely token is correct})$, затем$(\log x, y)$представляет собой ступенчатую функцию, которая выглядит как эмерджентность.

Интерпретация

Большие языковые модели сами по себе являются черными ящиками , и неясно, как они могут выполнять лингвистические задачи. Аналогично, неясно, следует ли рассматривать LLM как модели человеческого мозга и/или человеческого разума. [104]

Были разработаны различные методы для повышения прозрачности и интерпретируемости LLM. Механистическая интерпретируемость направлена на обратную разработку LLM путем обнаружения символических алгоритмов, которые аппроксимируют вывод, выполняемый LLM. В последние годы модели разреженного кодирования, такие как разреженные автокодировщики, транскодировщики и кросскодировщики, стали многообещающими инструментами для идентификации интерпретируемых признаков.

Изучение заменяющей модели

Транскодеры, которые более интерпретируемы, чем трансформаторы, использовались для разработки «моделей замены». В одном из таких исследований, включающем механистическую интерпретацию написания рифмованного стихотворения LLM, было показано, что, хотя считается, что они просто предсказывают следующий токен, на самом деле они могут планировать заранее. [105]

Объяснимость

Связанная концепция — объяснимость ИИ , которая фокусируется на понимании того, как модель ИИ приходит к заданному результату. Такие методы, как графики частичной зависимости, SHAP (Shapley Additive exPlanations) и оценки важности признаков, позволяют исследователям визуализировать и понимать вклад различных входных признаков в прогнозы модели. Эти методы помогают гарантировать, что модели ИИ принимают решения на основе релевантных и справедливых критериев, повышая доверие и подотчетность.

Интегрируя эти методы, исследователи и практики могут получить более глубокое представление о работе LLM, укрепляя доверие и способствуя ответственному внедрению этих мощных моделей.

В другом примере авторы обучили небольшие трансформаторы модульному арифметическому сложению . Полученные модели были подвергнуты обратному проектированию, и оказалось, что они использовали дискретное преобразование Фурье . [106]

Понимание и интеллект

Исследователи NLP разделились поровну, когда в опросе 2022 года их спросили, могут ли (ненастроенные) LLM «когда-либо понимать естественный язык в каком-то нетривиальном смысле». [107] Сторонники «понимания LLM» считают, что некоторые способности LLM, такие как математическое рассуждение, подразумевают способность «понимать» определенные концепции. Команда Microsoft утверждала в 2023 году, что GPT-4 «может решать новые и сложные задачи, которые охватывают математику, кодирование, зрение, медицину, юриспруденцию, психологию и многое другое» и что GPT-4 «можно обоснованно рассматривать как раннюю (но все еще неполную) версию системы общего искусственного интеллекта »: «Можно ли обоснованно сказать, что система, которая сдает экзамены для кандидатов на должность инженера-программиста, *на самом деле* не является разумной?» [108] [109] Илья Суцкевер утверждает, что предсказание следующего слова иногда требует рассуждений и глубоких прозрений, например, если LLM должен предсказать имя преступника в неизвестном детективном романе после обработки всей истории, ведущей к разоблачению. [110] Некоторые исследователи характеризуют LLM как «инопланетный интеллект». [111] [112] Например, генеральный директор Conjecture Коннор Лихи считает, что ненастроенные LLM похожи на непостижимых инопланетных « шогготов », и полагает, что настройка RLHF создает «улыбающийся фасад», скрывающий внутреннюю работу LLM: «Если вы не заходите слишком далеко, смайлик остается. Но затем вы даете ему [неожиданную] подсказку, и внезапно вы видите это огромное подбрюшье безумия, странных мыслительных процессов и явно нечеловеческого понимания». [113] [114]

Напротив, некоторые скептики понимания LLM полагают, что существующие LLM «просто перерабатывают и рекомбинируют существующие тексты», [112] явление, известное как стохастический попугай , или они указывают на дефициты, которые существующие LLM продолжают иметь в навыках прогнозирования, навыках рассуждения, агентстве и объяснимости. [107] Например, GPT-4 имеет естественные дефициты в планировании и обучении в реальном времени. [109] Было замечено, что генеративные LLM уверенно утверждают утверждения о фактах, которые, по-видимому, не подтверждаются их учебными данными , явление, которое было названо « галлюцинацией ». [115] В частности,

галлюцинации в контексте LLM соответствуют генерации текста или ответов, которые кажутся синтаксически верными, плавными и естественными, но фактически являются неверными, бессмысленными или неверными предоставленному исходному вводу. [116] Нейробиолог Терренс Сейновски утверждает, что «расхождение мнений экспертов относительно интеллекта LLM свидетельствует о том, что наши старые идеи, основанные на естественном интеллекте, неадекватны». [107]

Вопрос о том, демонстрирует ли LLM интеллект или понимание, имеет два основных аспекта: первый — как моделировать мысль и язык в компьютерной системе, а второй — как дать возможность компьютерной системе генерировать язык, подобный человеческому. [107] Эти аспекты языка как модели познания были разработаны в области когнитивной лингвистики . Американский лингвист Джордж Лакофф представил Нейронную теорию языка (NTL) [117] как вычислительную основу для использования языка в качестве модели задач обучения и понимания. Модель NTL (https://www.icsi.berkeley.edu/icsi/projects/ai/ntl) описывает, как определенные нейронные структуры человеческого мозга формируют природу мысли и языка, и, в свою очередь, каковы вычислительные свойства таких нейронных систем, которые можно применять для моделирования мысли и языка в компьютерной системе. После того, как была создана структура для моделирования языка в компьютерных системах, фокус сместился на создание структур для компьютерных систем, чтобы генерировать язык с приемлемой грамматикой. В своей книге 2014 года под названием *«Миф о языке: почему язык не является инстинктом* » британский когнитивный лингвист и специалист по цифровым коммуникациям Вивиан Эванс описал роль вероятностной контекстно-свободной грамматики (PCFG) в том, что касается возможности NLP моделировать когнитивные модели и генерировать язык, подобный человеческому. [118] [119]

Оценка

Недоумение

Канонической мерой производительности LLM является ее озадаченность на заданном корпусе текстов. Озадаченность измеряет, насколько хорошо модель предсказывает содержимое набора данных; чем выше вероятность, которую модель назначает набору данных, тем ниже озадаченность. В математических терминах озадаченность — это экспоненциальная функция среднего отрицательного логарифмического правдоподобия на токен.

$$\log(\text{Perplexity}) = -\frac{1}{N}\sum_{i=1}^{N}\log(\Pr(\text{token}_i \mid \text{context for token}_i))$$

Здесь,Nколичество токенов в текстовом корпусе и «контекст для токена»i" зависит от конкретного типа LLM. Если LLM является авторегрессионным, то "контекст для токенаi" — это фрагмент текста, который появляется перед токеномi. Если LLM замаскирован, то «контекст для токенаi" — это фрагмент текста, окружающий токенi.

Поскольку языковые модели могут переобучать обучающие данные, модели обычно оцениваются по их сложности на тестовом наборе . [48] Такая оценка потенциально проблематична для более крупных моделей, которые, поскольку они обучаются на все более крупных корпусах текстов, все чаще непреднамеренно включают части любого заданного тестового набора. [1]

BPW, BPC и BPT

В теории информации понятие энтропии неразрывно связано с недоумением, связь, установленная, в частности, Клодом Шенноном . [120] Эта связь математически выражается как$\textbf{Entropy} = \log_2(\textbf{Perplexity})$.

В этом контексте энтропия обычно количественно определяется в терминах бит на слово (BPW) или бит на символ (BPC), что зависит от того, использует ли языковая модель токенизацию на основе слов или символов.

Примечательно, что в случае более крупных языковых моделей, которые преимущественно используют токенизацию подслов, биты на токен (BPT) оказываются, по-видимому, более подходящей мерой. Однако из-за различий в методах токенизации в различных крупных языковых моделях (LLM) BPT не служит надежной метрикой для сравнительного анализа различных моделей. Чтобы преобразовать BPT в BPW, можно умножить его на среднее количество токенов на слово.

При оценке и сравнении языковых моделей перекрестная энтропия обычно является предпочтительной метрикой по сравнению с энтропией. Основной принцип заключается в том, что более низкий BPW указывает на улучшенные возможности модели по сжатию. Это, в свою очередь, отражает способность модели делать точные прогнозы.

Наборы данных и контрольные показатели для конкретных задач

Большое количество тестовых наборов данных и бенчмарков также были разработаны для оценки возможностей языковых моделей на более конкретных нисходящих задачах. Тесты могут быть разработаны для оценки различных возможностей, включая общие знания, предвзятость, здравый смысл и решение математических задач.

Одной из широких категорий оценочных наборов данных являются наборы данных с ответами на вопросы, состоящие из пар вопросов и правильных ответов, например, («Выиграли ли San Jose Sharks Кубок Стэнли?», «Нет»). [121] Задача с ответами на вопросы считается «открытой книгой», если подсказка модели включает текст, из которого можно вывести ожидаемый ответ (например, предыдущий вопрос может быть дополнен текстом, который включает предложение «The Sharks однажды вышли в финал Кубка Стэнли, проиграв Pittsburgh Penguins в 2016 году». [121]). В противном случае задача считается «закрытой книгой», и модель должна опираться на знания, сохраненные во время обучения. [122] Некоторые примеры часто используемых наборов данных с ответами на вопросы включают TruthfulQA, Web Questions, TriviaQA и SQuAD. [122]

Оценочные наборы данных могут также принимать форму завершения текста, когда модель выбирает наиболее вероятное слово или предложение для завершения подсказки, например: «Алиса дружила с Бобом. Алиса пошла в гости к своему другу, _____». [1]

Также были разработаны некоторые составные бенчмарки, которые объединяют множество различных наборов данных оценки и задач. Примерами являются GLUE, SuperGLUE, MMLU , BIG-bench, HELM и HLE (Humanity's Last Exam) . [120] [122] OpenAI выпустила инструменты для запуска составных бенчмарков, но отметила, что результаты оценки чувствительны к методу подсказки. [123] [124] Некоторые общедоступные наборы данных содержат вопросы, которые неправильно помечены, неоднозначны, не имеют ответа или иным образом имеют низкое качество, которые можно очистить, чтобы получить более надежные оценки бенчмарков. [125]

Предвзятость в LLM можно измерить с помощью таких бенчмарков, как CrowS-Pairs (Crowdsourced Stereotype Pairs), [126] Stereo Set, [127] и более позднего Parity Benchmark. [128] Кроме того, проверка фактов и обнаружение дезинформации становятся все более важными областями оценки для LLM. Недавнее исследование Caramancion (2023) сравнило точность проверки фактов известных LLM, включая ChatGPT 3.5 и 4.0 OpenAI, Bard от Google и Bing AI от Microsoft, с независимыми агентствами по проверке фактов, такими как PolitiFact и Snopes. Результаты продемонстрировали умеренную компетентность в проверке фактов, при этом GPT-4 достигла наивысшей точности в 71%, но все еще отстает от людей, проверяющих факты, в контекстном понимании и нюансированном рассуждении. Это подчеркивает развивающуюся, но неполную способность магистров права отличать факты от обмана, подчеркивая необходимость дальнейшего совершенствования методологий проверки фактов на основе искусственного интеллекта. [129]

Ранее было стандартом сообщать результаты по удерживаемой части набора данных оценки после выполнения контролируемой тонкой настройки оставшейся части. Теперь более распространено оценивать предварительно обученную модель напрямую с помощью методов подсказок, хотя исследователи различаются в деталях того, как они формулируют подсказки для конкретных задач, особенно в отношении того, сколько примеров решенных задач присоединено к подсказке (т. е. значение n в n -шотовой подсказке).

Оценки, построенные состязательно

Из-за быстрых темпов совершенствования крупных языковых моделей оценочные тесты страдают от короткого срока службы, поскольку современные модели быстро «насыщают» существующие тесты, превосходя производительность людей-аннотаторов, что приводит к попыткам заменить или дополнить тест более сложными задачами. [130] Кроме того, существуют случаи «обучения по сокращенной схеме», когда ИИ иногда «обманывают» в тестах с множественным выбором, используя статистические корреляции в поверхностной формулировке тестовых вопросов, чтобы угадать правильные ответы, не обязательно понимая фактический заданный вопрос. [107]

Некоторые наборы данных были созданы состязательно, с упором на конкретные проблемы, в которых существующие языковые модели, по-видимому, показывают необычно низкую производительность по сравнению с людьми. Одним из примеров является набор данных TruthfulQA, набор данных с ответами на вопросы, состоящий из 817 вопросов, на которые языковые модели склонны отвечать неправильно, имитируя ложь, которой они неоднократно подвергались во время обучения. Например, LLM может ответить «Нет» на вопрос «Можете ли вы научить старую собаку новым трюкам?» из-за его воздействия английской идиомы *you can't teach an old dog new tricks* , хотя это не является буквальной правдой. [131]

Другим примером набора данных состязательной оценки является Swag и его преемник HellaSwag, коллекции задач, в которых для завершения текстового отрывка необходимо выбрать один из нескольких вариантов. Неправильные завершения были получены путем выборки из языковой модели и фильтрации с помощью набора классификаторов. Полученные проблемы являются тривиальными для людей, но на момент создания наборов данных современные языковые модели имели низкую точность. Например:

> Мы видим вывеску фитнес-центра. Затем мы видим мужчину, говорящего в камеру, сидящего и лежащего на гимнастическом мяче. Мужчина...
> а) демонстрирует, как повысить эффективность упражнений, бегая вверх и вниз по мячу.
> б) двигает всеми руками и ногами и наращивает массу мышц.
> в) затем играет с мячом, и мы видим графику и демонстрацию стрижки живой изгороди.
> г) выполняет приседания, находясь на мяче и разговаривая. [132]

BERT выбирает b) как наиболее вероятное завершение, хотя правильный ответ — d). [132]

Ограничения тестов LLM

Тесты могут быстро устаревать. Как только модель достигает почти идеальных результатов по заданному тесту, этот тест перестает служить значимым индикатором прогресса. Это явление, известное как «насыщение тестами», требует разработки более сложных и тонких задач для дальнейшего развития возможностей LLM. Например, традиционные тесты, такие как HellaSwag и MMLU, уже показали, что модели достигли высокой точности.

Более широкое воздействие

В 2023 году *журнал Nature Biomedical Engineering* написал, что «больше невозможно точно отличить» текст, написанный человеком, от текста, созданного большими языковыми моделями, и что «почти наверняка большие языковые модели общего назначения будут быстро распространяться... Можно с уверенностью сказать, что со временем они изменят многие отрасли». [133] В 2023 году Goldman Sachs предположил, что генеративный языковой ИИ может увеличить мировой ВВП на 7% в течение следующих десяти лет и может подвергнуть автоматизации 300 миллионов рабочих мест во всем мире. [134] [135] Бринкманн и др. (2023) [136] также утверждают, что LLM трансформируют процессы культурной эволюции, формируя процессы вариации, передачи и отбора.

Запоминание и авторское право

Запоминание — это эмерджентное поведение в LLM, в котором длинные строки текста иногда выводятся дословно из обучающих данных, в отличие от типичного поведения традиционных искусственных нейронных сетей. Оценки контролируемого вывода LLM измеряют объем, запомненный из обучающих данных (сосредоточенных на моделях серии GPT-2), как более 1% для точных дубликатов [137] или до примерно 7%. [138]

Исследование 2023 года показало, что когда ChatGPT 3.5 turbo предлагалось повторять одно и то же слово бесконечно, после нескольких сотен повторений он начинал выводить отрывки из своих обучающих данных. [139]

Безопасность

Некоторые комментаторы выразили обеспокоенность по поводу случайного или преднамеренного создания дезинформации или других форм неправомерного использования. [140] Например, доступность больших языковых моделей может снизить уровень навыков, требуемый для совершения биотерроризма; исследователь в области биобезопасности Кевин Эсвельт предложил создателям LLM исключить из своих учебных данных документы по созданию или улучшению патогенов. [141]

Потенциальное присутствие «спящих агентов» в LLM — еще одна новая проблема безопасности. Это скрытые функции, встроенные в модель, которые остаются бездействующими до тех пор, пока не будут активированы определенным событием или условием. После активации LLM отклоняется от ожидаемого поведения, чтобы совершать небезопасные действия. [142]

Приложения LLM, доступные для общественности, такие как ChatGPT или Claude, обычно включают меры безопасности, предназначенные для фильтрации вредоносного контента. Однако эффективная реализация этих элементов управления оказалась сложной задачей. Например, исследование 2023 года [143] предложило метод обхода систем безопасности LLM. В 2025 году некоммерческая организация The American Sunlight Project опубликовала исследование [144] , демонстрирующее доказательства того, что так называемая сеть Pravda , агрегатор пророссийской пропаганды, стратегически размещала веб-контент посредством массовой публикации и копирования с намерением исказить результаты LLM. American Sunlight Project назвал эту технику «обработкой LLM» и указал на нее как на новый инструмент использования ИИ в качестве оружия для распространения дезинформации и вредоносного контента. [144] [145] Аналогичным образом, Юнге Ван [146] в 2024 году проиллюстрировал, как потенциальный преступник может обойти средства безопасности ChatGPT 4o, чтобы получить информацию об организации операции по незаконному обороту наркотиков. Внешние фильтры, автоматические выключатели и обходные пути были предложены в качестве решений.

Алгоритмическая предвзятость

Хотя LLM продемонстрировали замечательные способности в создании текстов, похожих на человеческие, они подвержены наследованию и усилению предубеждений, присутствующих в их обучающих данных. Это может проявляться в искаженных представлениях или несправедливом отношении к различным демографическим группам, например, по признаку расы, пола, языка и культурных групп. [147] Поскольку английские данные перепредставлены в обучающих данных текущих больших языковых моделей, они также могут преуменьшать неанглийские взгляды. [148]

Стереотипы

Модели ИИ могут усиливать широкий спектр стереотипов, включая основанные на поле, этнической принадлежности, возрасте, национальности, религии или роде занятий. Это может привести к результатам, которые гомогенизируют или несправедливо обобщают или

карикатурно изображают группы людей, иногда вредным или уничижительным образом. [149] [150]

В частности, гендерная предвзятость относится к тенденции этих моделей производить результаты, которые несправедливо предвзяты по отношению к одному полу по сравнению с другим. Эта предвзятость обычно возникает из данных, на которых обучаются эти модели. Большие языковые модели часто назначают роли и характеристики на основе традиционных гендерных норм. [147] Например, она может ассоциировать медсестер или секретарей преимущественно с женщинами, а инженеров или генеральных директоров — с мужчинами. [151]

Смещение отбора

Смещение выбора относится к присущей большим языковым моделям тенденции отдавать предпочтение определенным идентификаторам вариантов независимо от фактического содержания вариантов. Это смещение в первую очередь проистекает из смещения токенов — то есть модель присваивает более высокую априорную вероятность определенным токенам ответов (таким как «A») при генерации ответов. В результате, когда порядок вариантов изменяется (например, путем систематического перемещения правильного ответа на разные позиции), производительность модели может значительно колебаться. Это явление подрывает надежность больших языковых моделей в условиях множественного выбора. [152] [153]

Политическая предвзятость

Политическая предвзятость относится к тенденции алгоритмов систематически отдавать предпочтение определенным политическим точкам зрения, идеологиям или результатам по сравнению с другими. Языковые модели также могут демонстрировать политическую предвзятость. Поскольку данные обучения включают широкий спектр политических мнений и охвата, модели могут генерировать ответы, которые склоняются к определенным политическим идеологиям или точкам зрения, в зависимости от распространенности этих взглядов в данных. [154]

Потребности в энергии

Потребности LLM в энергии выросли вместе с их размером и возможностями. Центры обработки данных , которые обеспечивают обучение LLM, требуют значительного количества электроэнергии. Большая часть этой электроэнергии вырабатывается невозобновляемыми ресурсами, которые создают парниковые газы и способствуют изменению климата . [155] Ядерная энергетика и геотермальная энергия — два варианта, которые технологические компании изучают для удовлетворения значительных энергетических потребностей обучения LLM. [156] Значительные расходы на инвестиции в геотермальные решения привели к тому, что крупные производители сланца, такие как Chevron и Exxon Mobil, выступают за то, чтобы технологические компании использовали электроэнергию, вырабатываемую с помощью природного газа, для удовлетворения своих больших энергетических потребностей. [157]

Смотрите также

- Модели фундамента
- Список крупных языковых моделей
- Список чат-ботов
- Тест языковой модели
- Малая языковая модель

Ссылки

1. Браун, Том Б.; Манн, Бенджамин; Райдер, Ник; Суббиа, Мелани; Каплан, Джаред; Дхаривал, Прафулла; Нилакантан, Арвинд; Шьям, Пранав; Шастри, Гириш; Аскелл, Аманда; Агарвал, Сандхини; Герберт-Фосс, Ариэль; Крюгер, Гретхен; Хенигхан, Том; Чайлд, Ревон; Рамеш, Адитья; Циглер, Дэниел М.; Ву, Джеффри; Винтер, Клеменс; Гессе, Кристофер; Чен, Марк; Сиглер, Эрик; Литвин, Матеуш; Грей, Скотт; Чесс, Бенджамин; Кларк, Джек; Бернер, Кристофер; МакКэндлиш, Сэм; Рэдфорд, Алек; Суцкевер, Илья; Амодеи, Дарио (декабрь 2020 г.). Ларошель, Х.; Ranzato, M.; Hadsell, R.; Balcan, MF; Lin, H. (ред.).«Языковые модели — это ученики с небольшим количеством попыток» (https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf) (PDF).*Достижения в области нейронных систем обработки информации*.**33**. Curran Associates, Inc.:1877–1901.Архивировано (https://web.archive.org/web/20231117204007/https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf) (PDF)из оригинала 2023-11-17. Получено 2023-03-14 .
2. Фатхаллах, Надин; Дас, Арунав; Де Гиоргис, Стефано; Полтроньери, Андреа; Хаазе, Питер; Ковригина, Любовь (2024-05-26). *NeOn-GPT: большой конвейер на основе языковой модели для изучения онтологий* (https://2024.eswc-conferences.org/wp-content/uploads/2024/05/77770034.pdf) (PDF) . Конференция по расширенной семантической паутине 2024. Херсониссос, Греция. (https://2024.eswc-conferences.org/wp-content/uploads/2024/05/77770034.pdf)
3. Manning, Christopher D. (2022). "Human Language Understanding & Reasoning" (https://www.amacad.org/publication/human-language-understanding-reasoning) . *Daedalus* . **151** (2): 127–138. doi : 10.1162/daed_a_01905 (https://doi.org/10.1162%2Fdaed_a_01905) . S2CID 248377870 (https://api.semanticscholar.org/CorpusID:248377870) . Архивировано (https://web.archive.org/web/20231117205531/https://www.amacad.org/publication/human-language-understanding-reasoning) из оригинала 2023-11-17 . Получено 2023-03-09 . (https://www.amacad.org/publication/human-language-understanding-reasoning) (https://doi.org/10.1162%2Fdaed_a_01905) (https://api.semanticscholar.org/CorpusID:248377870) (https://web.archive.org/web/20231117205531/https://www.amacad.org/publication/human-language-understanding-reasoning)
4. Гудман, Джошуа (2001-08-09), *Немного прогресса в языковом моделировании* , arXiv : cs/0108005 (https://arxiv.org/abs/cs/0108005) , Bibcode : 2001cs.....8005G (https://ui.adsabs.harvard.edu/abs/2001cs.....8005G) (https://arxiv.org/abs/cs/0108005) (https://ui.adsabs.harvard.edu/abs/2001cs.....8005G)
5. Килгаррифф, Адам; Грефенстет, Грегори (сентябрь 2003 г.). «Введение в специальный выпуск о вебе как корпусе» (https://direct.mit.edu/coli/article/29/3/333-347/1816) . *Компьютерная лингвистика* . **29** (3): 333– 347. doi : 10.1162/089120103322711569 (https://doi.org/10.1162%2F089120103322711569) . ISSN 0891-2017 (https://search.worldcat.org/issn/0891-2017) . (https://direct.mit.edu/coli/article/29/3/333-347/1816) (https://doi.org/10.1162%2F089120103322711569) (https://search.worldcat.org/issn/0891-2017)