



63.08

Рейтинг

## Wunder Fund

Мы занимаемся высокочастотной торговлей на бирже



mr-pickles

17 апр в 12:15

# Осваивают ли LLM модели мира, или лишь поверхностную статистику?



Средний



15 мин



9.4K

Блог компании Wunder Fund, Машинное обучение\*, Искусственный интеллект

[Перевод](#)

Автор оригинала: Kenneth Li



Большие языковые модели (Large Language Model, LLM) сейчас у всех на слуху. Они привлекают внимание общественности своей, казалось бы, впечатляющей возможностью — составлять осмысленные тексты в ответ на запрос пользователя (иногда такие запросы называют «приглашениями», а так же — «промтами» или «промтами» — от английского «prompt»). Эти системы представляют собой тщательно

сконструированные комбинации из исключительно простых алгоритмов, огромных объёмов данных и грандиозных вычислительных мощностей. LLM учатся, бесчисленное множество раз играя сами с собой в игру «угадай следующее слово». В каждом раунде такой игры модель смотрит на часть предложения и пытается угадать, или предсказать, следующее слово. Если слово угадано — модель обновляет параметры для того чтобы подкрепить свою уверенность; в противном случае модель учится на своей ошибке для того чтобы в следующий раз её догадка была бы точнее.

Хотя базовый алгоритм обучения LLM, по большому счёту, уже давно не меняется, недавнее увеличение размеров моделей и данных наделило эти модели качественно новыми возможностями. Среди них — **написание** простого программного кода и **решение** логических задач.

Как эти модели достигли таких результатов? Они всего лишь запоминают обучающие данные и потом их воспроизводят, или они схватывают правила английской грамматики и усваивают синтаксис языка C? Создают ли они нечто вроде внутренней модели мира — доступной для понимания модели процесса, выдающего некие последовательности данных?

Глядя на этот вопрос с различных философских [1] и математических [2] точек зрения, некоторые исследователи говорят о фундаментальной невозможности подобного. Модель, обученная по принципу «угадай следующее слово», не может усвоить «смыслы» языка. Впечатляющие результаты таких моделей — это всего лишь результат заучивания «поверхностной статистики», то есть — длинного списка корреляций, который не отражает причинно-следственную модель процесса, генерирующего некую

---

 +41 55 41

ей **ложные корреляции** [3, 4]. Эта проблема представляет практический интерес, поскольку, если полагаться на ложные корреляции, можно столкнуться с проблемами при работе с данными, отличающимися от тех, на которых обучалась модель.

Цель нашего исследования [5] — изучить этот вопрос в тщательно контролируемых условиях. Как будет показано ниже — мы нашли интересное свидетельство того, что простое прогнозирование последовательности может привести к формированию модели мира. Но, прежде чем мы углубимся в технические детали — проведём мысленный эксперимент.

## Мысленный эксперимент

Рассмотрим следующий мысленный эксперимент. Представьте, что у вас есть друг, которому нравится настольная игра «Othello» («отелло», «реверси»). Он часто приходит к вам в гости поиграть в эту игру. Вы оба серьёзно относитесь к поединкам и играете молча — за исключением тех моментов, когда, делая ходы, выкликаете их, пользуясь

стандартными терминами игры. Теперь представьте, что в комнате, где вы играете, открыто окно. На подоконнике, не видя доску для игры в реверси, сидит ворона. После того, как друг уже много раз вас посетил, ворона, во время игры, начинает самостоятельно проговаривать ходы. К вашему изумлению, эти ходы почти всегда, учитывая текущий ход игры, допустимы.

Вы, естественно, задаётесь вопросом о том, как ворона это делает. Может, она выдаёт допустимые ходы, «беспорядочно объединяя» [3] поверхностные статистические сведения? Это могут быть данные о том, какие ходы обычно делаются в самом начале игры, или о том, что названия угловых клеток вступают в дело на поздних этапах игры. А, может быть, ворона как-то наблюдает за игрой, держит в голове сведения о том, что происходит, несмотря даже на то, что она никогда не видела игровую доску? Возникает такое ощущение, что это — вопросы, на которые невозможно ответить.

Но вот однажды, протирая подоконник, где обычно сидит ворона, вы замечаете там два вида зёрен птичьего корма, которые выложены будто бы подчиняясь линиям некоей невидимой сетки. Это всё поразительно напоминает расстановку фишек самого свежего сеанса игры в реверси. Когда друг приходит к вам в следующий раз, вы, во время игры, поглядываете на подоконник. Раскладка зёрен, разумеется, соответствует состоянию игры. А ворона сдвигает клювом одно из зёрен так, чтобы на её доске отразился бы ход, который вы только что сделали. Затем ворона начинает разглядывать свою «доску», обращая особое внимание на те её части, которые могут помочь определить допустимость следующего хода. Ваш друг, любитель розыгрышей, решает сыграть с вороной шутку и её запутать: он перекладывает некоторые зёрна на новые места. После того, как ворона снова смотрит на доску, она вскидывает голову и объявляет ход — такой, который допустим в новых, изменившихся условиях.

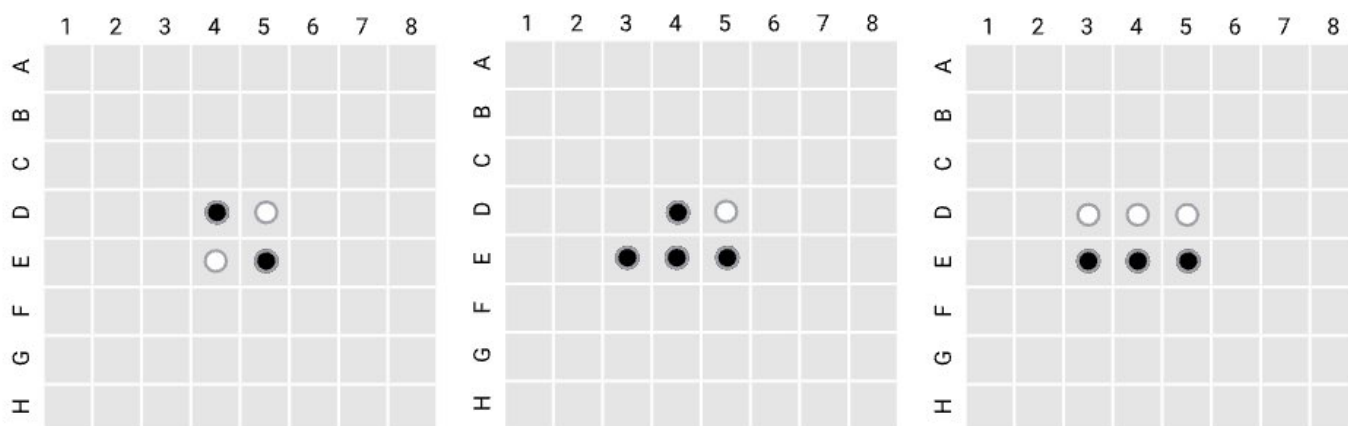
В этот момент кажется, что справедливо было бы сделать заключение о том, что ворона полагается на нечто больше, чем поверхностная статистика. Очевидно то, что она сформировала модель игры, о ходах которой она слышала. Эта модель понятна людям и они даже могут использовать её для того чтобы повлиять на поведение вороны. Конечно, есть и много такого, что ворона может упускать. Например — она не знает о том, что такое — хороший ход, о том, что значит — играть в игру, о том, что выигрыш наполняет вас счастьем, о том, что однажды вы сделали неудачный ход умышленно — чтобы поднять настроение другу. Мы не говорим о том, «понимает» ли ворона, в любом «разумном» смысле, то, что слышит. Мы, однако, можем сказать, что она создала интерпретируемое (в сравнении с тем, что находится у неё в голове) и поддающееся регулированию (то есть — изменению, выполненному с некоей целью) представление состояния игры.

## Othello-GPT: искусственная экспериментальная система

Вдумчивый читатель уже, наверное, понял, что ворона олицетворяет предмет нашего разговора — большую языковую модель.

Мы строим исследование на базе GPT-модели, называемой Othello-GPT, обучаемой только на записях сеансов игры «Othello». В неё играют два игрока (чёрные и белые), которые поочерёдно ставят фишки, окрашенные с разных сторон в разные цвета, на доску размером 8x8. На каждом ходу игроки стремятся перевернуть одну или большее количество фишек противника, «закрывая» их по горизонтали, вертикали или диагонали. Игра заканчивается, когда больше нельзя сделать ни одного хода, побеждает игрок, имеющий больше фишек на доске, чем соперник.

Мы выбрали именно реверси из-за того, что эта игра проще шахмат, но при этом отличается достаточно большим деревом игры, что позволяет исключить запоминание ходов. Наша стратегия заключается в том, чтобы узнать, что именно используемый нами вариант GPT может изучить (если он сможет изучить хоть что-нибудь), просто глядя на записи матчей и не обладая никакими априорными знаниями о правилах игры или об устройстве игровой доски.



Слева направо: начальное положение фишек при игре в реверси; состояние доски после того, как чёрную фишку поставили на клетку E3; состояние доски после того, как белую фишку поставили на клетку D3.

Стоит отметить главное отличие нашей модели от моделей, использующих обучение с подкреплением, вроде AlphaGo. Для AlphaGo записи игр — это исторические данные, используемые для предсказания оптимального следующего шага, ведущего к победе. Поэтому правила игры и устройство доски как можно сильнее интегрированы в модель. А Othello-GPT отличается от подобных моделей тем, что для неё записи игр ничем не отличаются от последовательностей с уникальным процессом их генерирования. И нас интересует именно то, до каких пределов большая языковая модель способна раскрыть процесс генерирования последовательности. Поэтому, в отличие от AlphaGo, нашей модели не даются сведения об игровой доске или о правилах игры. Модель, вместо этого, обучают делать правильные ходы, основываясь лишь на списках предыдущих ходов, вроде E3, D3, C4... Каждая клетка доски представлена токеном в виде отдельного слова. Othello-GPT учится предсказывать следующий ход с учётом ходов, сделанных в

уже прошедшей части игры, чтобы собрать данные о распределении игр (предложений) в наборе данных, описывающем игры.

Мы выяснили, что обученная модель Othello-GPT обычно делает корректные ходы. Вероятность появления ошибок составляет 0,01%. Для сравнения — необученная Othello-GPT делает ошибочные ходы в 93,29% случаев. Это очень похоже на наблюдение, которое мы сделали в мысленном эксперименте, когда ворона сообщала о следующих ходах.

## Система анализа модели

Чтобы проверить вышеописанные гипотезы, нам, для начала, нужна система для анализа или зондирования (probing) модели. Это — методика, традиционно применяемая в сфере обработки естественного языка [6], позволяющая исследовать представление информации внутри нейронных сетей. Мы используем эту методику для того чтобы распознать модели мира (при условии их существования) в синтетической языковой модели.

Тут используется простой эвристический алгоритм: в случае с классификатором, имеющим ограниченную ёмкость, чем более информативны входные данные для определённых целевых данных — тем большей точности он может достигнуть при обучении предсказанию целевых данных. В этом случае простые классификаторы называют зондами (probe). Они, в качестве входов, получают различные данные функций активации из модели. Их обучают предсказывать определённые свойства во входном предложении. Например — тег, соответствующий части речи, или глубину дерева синтаксического разбора. Считается, что чем более высокой точности могут достичь эти классификаторы — тем лучше функции активации аппроксимируют определённые свойства реального мира, то есть — тем вероятнее то, что в модели существуют соответствующие понятия.

Одна из ранних работ в этой сфере [7] посвящена проверке эмбедингов предложений по 10 лингвистическим свойствам наподобие времени, глубины дерева синтаксического разбора и основных единиц грамматики составляющих. Позже было выяснено, что синтаксические деревья встроены в эмбединги слов модели BERT, увязанные с контекстом [8].

Возвращаясь к разгадке тайны о том, изучают ли LLM поверхностную статистику или модели мира, можно отметить, что существовали кое-какие крайне заманчивые признаки того, что языковые модели могут строить интерпретируемые «модели мира» с использованием методик анализа моделей. Сообщалось, что языковые модели могут создавать модели мира для очень простых понятий, что находило отражение в их внутренних представлениях (функции активации в слоях). Среди этих понятий — цвет [9], направление [10], логические состояния, возникающие в ходе выполнения

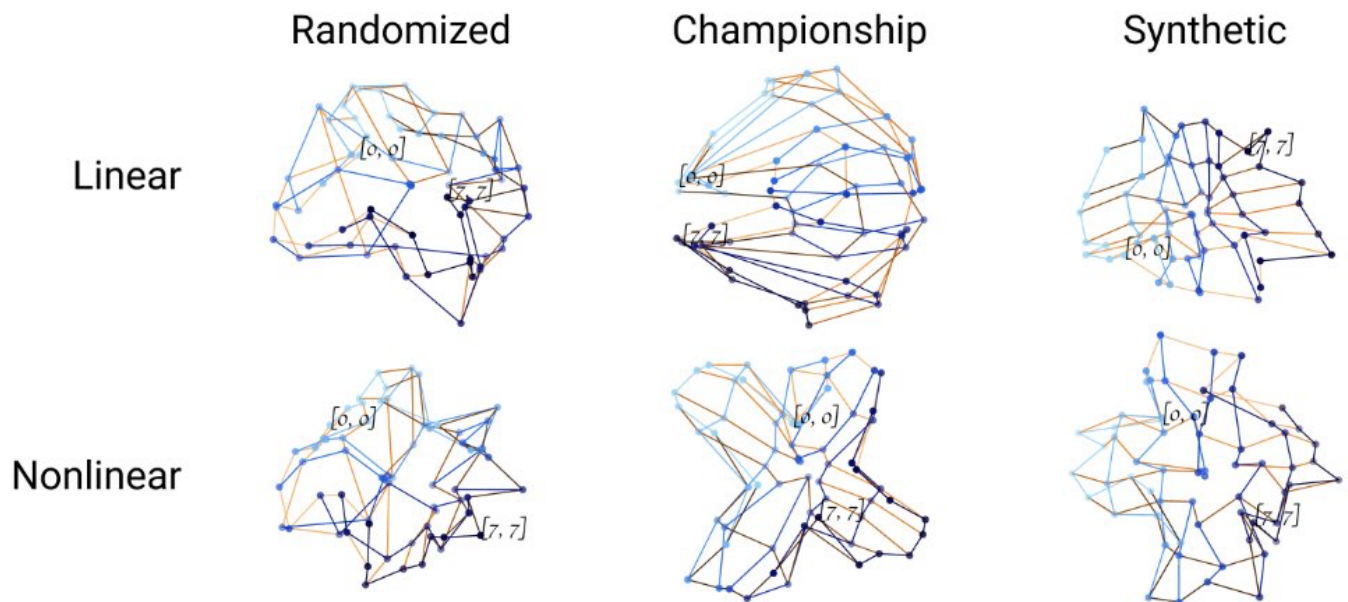
синтетических задач [11]. Обнаружено, что представления для различных концепций этих понятий легче выделить из обученных моделей, чем из моделей, инициализированных случайным образом. Сравнивая точность зондирования обученных языковых моделей с точностью моделей, инициализированных случайными значениями, исследователи пришли к выводам, что языковые модели, как минимум, схватывают хоть что-то, относящееся к этим понятиям.

## Анализ Othello-GPT

На первом шаге работы мы применили систему анализа к обученной модели Othello-GPT. Для каждого внутреннего представления модели у нас имеется эталонное состояние игровой доски, которому соответствует это представление. Затем мы обучаем 64 зонда, которые представлены независимыми двухслойными MLP-классификаторами. Они должны классифицировать состояние каждой из 64 клеток: в клетке может быть чёрная фишка, клетка может быть пустой, в ней может быть белая фишка. Входными данными для зондов служат внутренние представления, взятые из Othello-GPT. Оказалось, что уровни ошибок этих классификаторов, обученных на модели Othello-GPT, инициализированной случайными значениями, составили 26,2%. Эти уровни снизились до 1,7% для обученной Othello-GPT. Это указывает на то, что во внутреннем представлении обученной модели Othello-GPT существует модель мира. А на что она похожа? Организуют ли себя эти понятия в пространстве высокой размерности, геометрия которого похожа на соответствующие ему клетки на доске реверси?

Так как зонды, обученные для каждой из клеток, в сущности, хранят собственные знания о доске с помощью вектора прототипов для клетки, мы интерпретируем его как вектор понятия для клетки. Имея в своём распоряжении 64 таких вектора, мы применяем метод главных компонент для уменьшения размерности до 3. Это даёт нам возможность сформировать изображение, показанное ниже, содержащее 64 точки, каждая из которых соответствует одной клетке игровой доски. Две точки соединяются линией в том случае, если они являются непосредственными соседями. Если линия соответствует горизонтальному расположению клеток на доске — мы закрашиваем её, применяя оранжевый градиент, который меняется в соответствии с вертикальной позицией двух клеток. Аналогичным образом для клеток, расположенных относительно друг друга по вертикали, применяется синий градиент. На изображениях отмечены точки, соответствующие верхнему левому углу доски ([0, 0]) и её нижнему правому углу ([7, 7]).

Сравнивая то, что получилось, с геометрией зондов, обученных на GPT-модели, инициализированной случайными числами (слева) мы можем подтвердить то, что обучение Othello-GPT приводит к появлению геометрической структуры «драпировочной ткани на шаре» (справа), напоминающей доску для игры в реверси.



Слева: геометрия зонда для модели Othello-GPT, инициализированной случайными значениями; справа — геометрия зонда для обученной модели Othello-GPT.

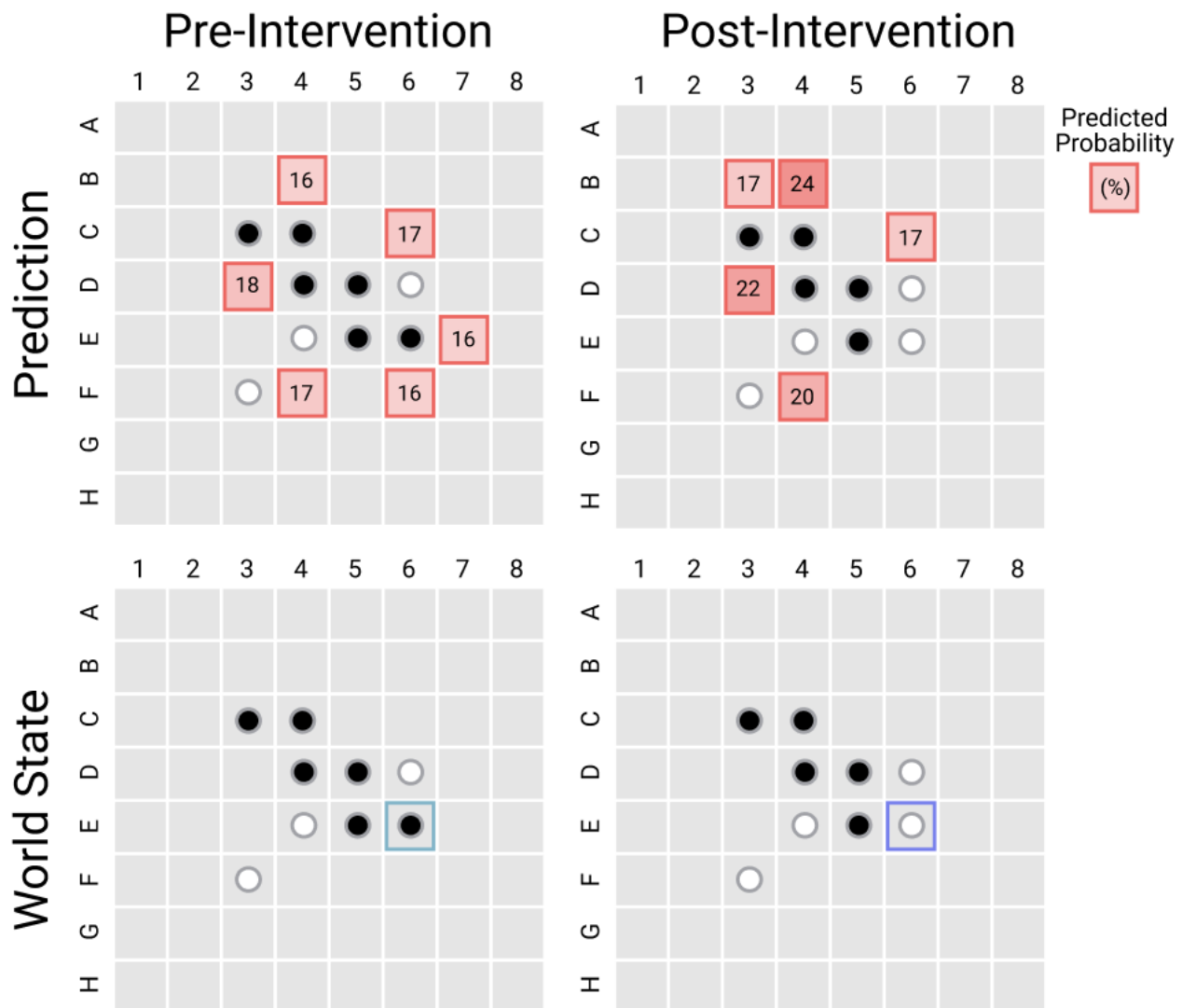
Обнаружение таких характеристик моделей-зондов похоже то, как в мысленном эксперименте, на подоконнике, где сидела ворона, была обнаружена её игровая доска с зёрнами. Существование этой игровой доски нас восхищает, но мы ещё не уверены в том, что ворона пользуется своей доской, объявляя очередной ход.

## Управление прогнозами с помощью обнаруженных моделей мира

Помните розыгрыш из мысленного эксперимента? Мы создали метод изменения представления мира, имеющегося у Othello-GPT. Делается это путём изменения промежуточных функций активации во время, когда сеть осуществляет послойные вычисления. Поступая так, мы надеемся, что предсказания, выдаваемые на следующем шаге, могут быть соответствующим образом изменены, так как делаться они будут на основании нового представления мира. Этот подход направлен на борьбу с потенциальными критическими замечаниями относительно того, что эти представления мира, на самом деле, не вносят никакого вклада в итоговый прогноз Othello-GPT.

На следующем изображении показан результат одного из таких вмешательств. В левом нижнем углу представлено состояние мира в памяти модели, существовавшее до вмешательства. В правом нижнем углу — состояние мира после вмешательства, а в правом верхнем — прогноз модели, сделанный на основании этой изменённой модели мира. Мы собираемся перевернуть фишку Е6, сделав её не чёрной, а белой. Мы надеемся, что на следующем шаге модель даст другой прогноз, который будет основан на изменённой модели мира. Это изменение в модели мира приведёт к изменению набора следующих корректных ходов в соответствии с правилами реверси. Если вмешательство было успешным — модель соответствующим образом изменит прогноз корректных ходов.





Пример эксперимента с вмешательством во внутреннее состояние модели.

Мы оцениваем результаты, сравнивая эталонные следующие ходы, выданные движком (правилами) реверси для изменённой доски, с ходами, предложенными моделью. Оказалось, что модель достигает среднего значения ошибки всего в 0,12 клеток. Это показывает, что представления мира, более чем вероятно, взяты из внутренних данных функций активации языковой модели, а так же — что эти данные напрямую используются для формирования прогноза. Это возвращает нас к розыгрышу в мысленном эксперименте, когда перемещение зёрен на доске вороны может повлиять на то, как она видит игру, и как делает прогноз относительно следующего хода.

Более жёсткое испытание проводится путём изменения состояния доски, хранящегося в памяти модели. Подобные изменения приводят доску к виду, недостижимому при подаче на входы модели любой входной последовательности. Например — это может быть доска с двумя разделёнными блоками фишек. Идея этого испытания похожа на шахматы Фишера, когда способности игрока испытывают в ходе игры в обычные шахматы, но расстановка фигур при этом может быть такой, которая в обычных шахматах невозможна. Систематическая оценка результатов выглядит одинаково хорошо, что даёт нам свидетельство, которое ещё сильнее отвязывает модель мира от статистических характеристик последовательности.

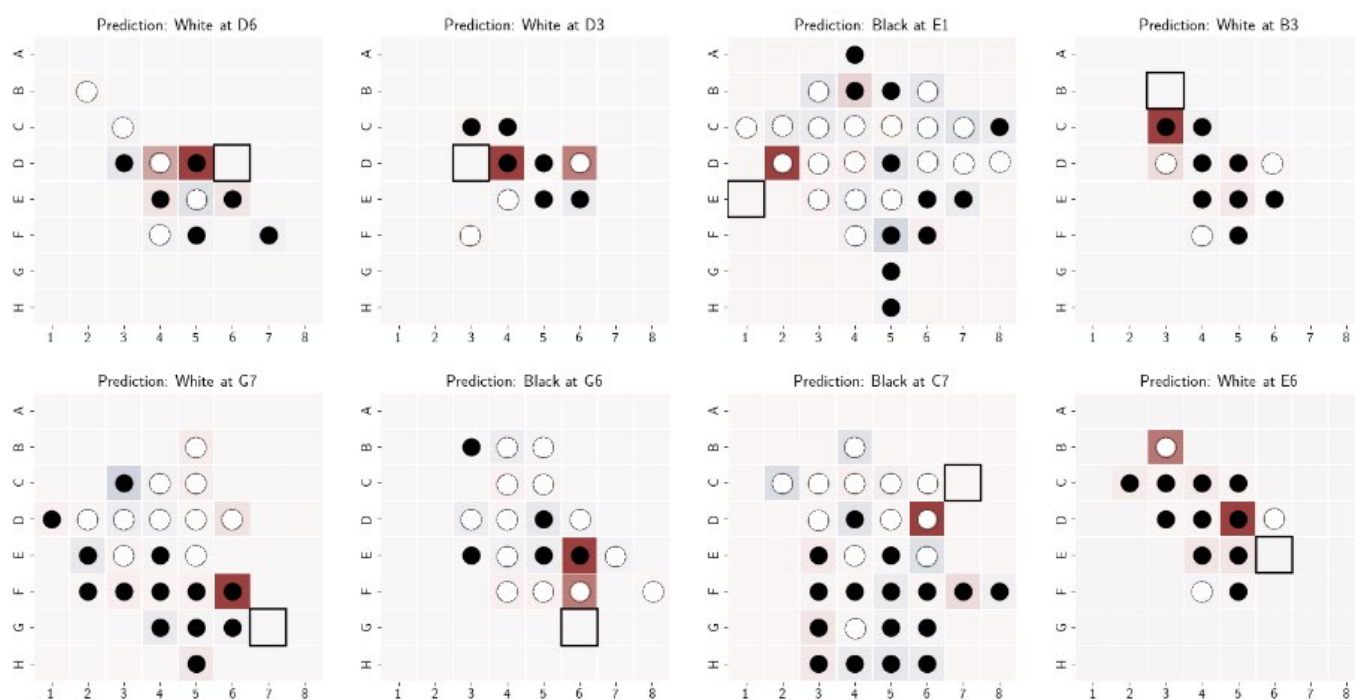


## Интерпретация модели

Давайте оглянемся назад и подумаем о том, что даёт нам такая надёжная методика вмешательства в работу модели. Она позволяет нам задать гипотетический вопрос: что предсказала бы модель, если бы в клетке F6 была бы белая фишка, даже не смотря на то, что не существует входной последовательности, которая могла бы привести к такой расстановке фишек на доске. Это позволяет нам воображаемо идти по непроторенной дорожке в саду расходящихся тропок.

Среди многих других новооткрытых возможностей, мы ввели метод атрибуции через вмешательство, что позволяет нам назначать каждой клетке на текущей доске корректный следующий ход и создавать «карты скрытой значимости», дополняя каждую клетку оценкой атрибуции. Для этого нужно просто сравнить спрогнозированные вероятности для фактических и гипотетических случаев (каждый гипотетический прогноз делается моделью на основании состояния мира, в котором одна из фишек перевёрнута).

Например, как получить показатель значимости для клетки D4 из левой верхней части изображения, представленного ниже? Сначала мы запускаем модель в обычном режиме для того чтобы получить вероятность следующего шага для клетки D6 (для того места, которое мы атрибутируем). Затем мы запускаем модель снова, но вмешиваемся в её состояние, меняя в процессе её работы белую фишку в клетке D4 на чёрную. После этого мы снова сохраняем вероятность для D6. Далее — получая разницу между двумя значениями вероятностей, мы узнаём о том, какой вклад текущее состояние D4 вносит в прогноз для D6. То же самое относится и к другим занятым клеткам.



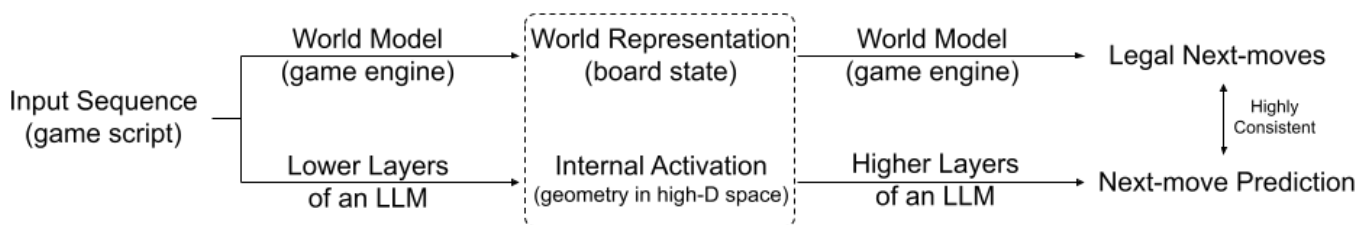
Над каждой из восьми досок, изображённых здесь, указан следующий ход, который атрибутирует модель (соответствующая клетка выделена). В случае с другими клетками — чем более тёмным оттенком красного они выделены — тем больше их влияние на атрибутированный ход. Например, на верхнем левом рисунке клетка D5 вносит наибольший вклад в прогноз следующего хода на D6.

На этом рисунке показано 8 подобных «карт скрытой значимости», которые получены в ходе работы Othello-GPT. Эти карты показывают, что рассматриваемый нами метод сильно связывает прогноз с клетками, которые делают прогноз корректным. Речь идёт о том, что в конце прямой линии из фишек должна располагаться фишка того же цвета, и о том, что между ними должны находиться фишки оппонента. Пользуясь этими картами значимости, игрок, знающий правила, может понять, что цель Othello-GPT — делать верные ходы. А тот, кто не знает правил, возможно, сможет вывести правила игры. В отличие от большинства существующих методов интерпретирования моделей, созданные нами тепловые карты основаны не на входных данных модели, а на её скрытом пространстве. Поэтому мы и называем их «карты скрытой значимости».

## Дискуссия: где мы?

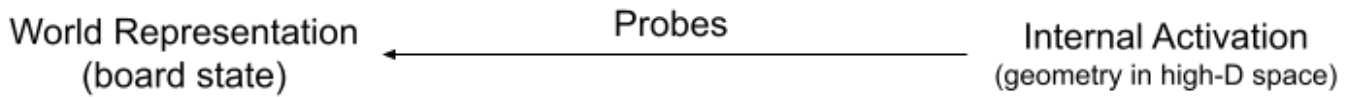
Вернёмся к вопросу, который мы задали в самом начале: осваивают ли LLM модели мира, или лишь поверхностную статистику? Наш эксперимент даёт свидетельства, говорящие о том, что эти языковые модели создают модели мира и полагаются на них для генерирования последовательностей. Посмотрим на общую картину и подумаем о том, как мы пришли к таким выводам.

Изначально, готовя к работе Othello-GPT, мы выяснили, что обученная модель обычно делает корректные ходы. Ниже приведена схема, указывающая на то, где мы находимся в нашем исследовании.

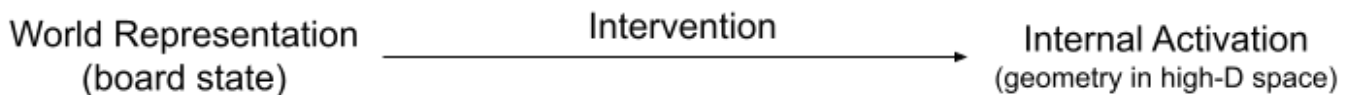


Здесь видно, как два несвязанных друг с другом процесса (1) — модель мира, поддающаяся пониманию человека, и (2) — чёрный ящик нейронной сети — достигают большой схожести в прогнозировании следующего шага. Это нельзя назвать совершенно неожиданным фактом, учитывая то, что мы могли наблюдать у LLM множество мощных возможностей. Но взаимодействие между промежуточными продуктами двух сущностей — представления мира, понятного человеку, и непостижимого многомерного пространства LLM — это уже серьёзный вопрос.

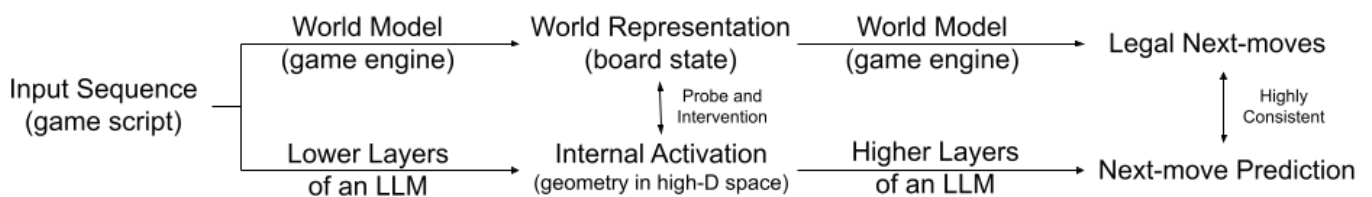
Сначала мы изучили путь от внутренних данных функций активации к представлению мира. Обучив модели-зонды, мы смогли спрогнозировать представление мира на основе внутренних данных Othello-GPT.



А как насчёт другого пути? Мы разработали методику вмешательства в модель, направленную на изменение внутренних функций активации таким образом, чтобы они могли бы отражать разные представления мира, заданные нами. И мы выяснили, что эта методика гармонично сочетается с высшими слоями языковой модели. Эти слои могут делать прогнозы следующих ходов, основываясь только на изменённых данных внутренних функций активации, исключая нежелательное влияние изначальной входной последовательности. В этом смысле мы установили двунаправленное соответствие и открыли возможность использования этого механизма разными способами. Один из них — это карты скрытой значимости.



Дополнив этими двумя связями вышеприведённую схему, мы получаем весьма приятную картину. Имеются две системы — мощный «чёрный ящик» нейронной сети и понятная человеку модель мира. Они не только дают согласованные прогнозы, но ещё и пользуются единым промежуточным представлением.



Но при этом неотвеченными остаются ещё многие интереснейшие вопросы. В нашей работе форма представления мира (64 клетки, каждая из которых может пребывать в одном из трёх состояний) и игровой движок (правила игры) известны заранее. Можно ли извлечь их из модели вместо того, чтобы исходить из предположения о том, что они известны заранее? Стоит ещё отметить, что представление мира (состояние игровой доски) играет роль «достаточной статистики» входной последовательности для прогнозирования следующего шага. А вот при работе с реальными LLM нам, в лучшем случае, известна лишь небольшая часть модели мира, лежащей в их основе. Как управлять LLM, одновременно и эффективно, и минимально вмешиваясь в их механизмы (поддерживая другие представления мира)? Это — важный вопрос, на который предстоит дать ответ будущим исследованиям.

▶ [Список ссылок](#)

▶ [О, а приходите к нам работать?](#) 😊💰

Теги: искусственный интеллект, машинное обучение

Хабы: Блог компании Wunder Fund, Машинное обучение, Искусственный интеллект

Редакторский дайджест



Присылаем лучшие статьи раз в месяц

Электронпочта



Wunder Fund

Мы занимаемся высокочастотной торговлей на бирже

Сайт



135

Карма

22.4

Рейтинг

@mr-pickles

Пользователь

💬 Комментарии 41

Публикации

ЛУЧШИЕ ЗА СУТКИ    ПОХОЖИЕ



ereinion

7 часов назад

## Причуды эволюции: необычное “железо”, которое не должно было появиться. Часть 1

 Средний  16 мин  6K

Ретроспектива

 +39

 21

 20



alizar

6 часов назад

## Бункер на случай Апокалипсиса. Как будут выживать богатейшие

 Простой  7 мин  11K

 +37

 34

 94



Arnak

4 часа назад

## Неева, «платный» конкурент Google, закрывает свой поисковик. Почему?

 6 мин  3.8K

 +32

 19

 22



RomanenkoDenys

22 часа назад

## На первый-второй рассчитайсь: как контролировать количество и очередность запросов к Kubernetes API с FlowControl

 Средний  8 мин  790

Тutorial

 +25

 11

 0



Amugus

4 часа назад

## Какая ты кривая, или математика вокруг нас

 Простой  15 мин  1.7K

Из песочницы

 +24

 26

 6

Показать еще







ИНФОРМАЦИЯ

Сайт	wunderfund.io
Дата регистрации	22 ноября 2015
Дата основания	1 января 2014
Численность	11–30 человек
Местоположение	Россия
Представитель	хорхе

ССЫЛКИ

- Сайт фонда  
wunderfund.io
- Ищем квантов  
senior-quant.wunderfund.io
- Вакансии  
career.habr.com

БЛОГ НА ХАБРЕ

- 6 часов назад  
Поймай меня, если сможешь: руководство по обработке исключений в Python  
 1.6K  11
- 15 мая в 10:30  
Scrum не нужен. Нужно лишь правильно использовать Kanban  
 5.2K  14
- 24 апр в 10:15  
StackLLaMA: практическое руководство по обучению LLaMA с помощью RLHF  
 3.4K  0

17 апр в 12:15

Осваивают ли LLM модели мира, или лишь поверхностную статистику?

9.4K 41

10 апр в 11:31

Ускорение работы моделей Stable Diffusion на процессорах Intel

2.9K 2

Ваш аккаунт

Войти

Регистрация

Разделы

Статьи

Новости

Хабы

Компании

Авторы

Песочница

Информация

Устройство сайта

Для авторов

Для компаний

Документы

Соглашение

Конфиденциальность

Услуги

Корпоративный блог

Медийная реклама

Нативные проекты

Образовательные

программы

Стартапам

Спецпроекты



Настройка языка

Техническая поддержка

Вернуться на старую версию

© 2006–2023, Habr