УДК 004.8 + 004.93
DOI 10.17726/phillT.2024.2.6

Почему трансформеры не мыслят, как люди?

Александр Б. Хомяков,
магистр физических наук
Санкт-Петербург, Россия

alexander.xom@gmail.com

Аннотация. Большие языковые модели в виде чат-ботов очень правдоподобно имитируют диалог как всезнающий собеседник и поэтому получили широкое распространение. Но даже в чат-боте Google Gemini не советуют доверять тому, что напишет чат-бот, и просят проверять его ответы. В данном обзоре будут проанализированы различные типы ошибок LLM, такие как проклятие инверсии, обработка чисел и др., чтобы выявить их причины. Такой анализ привел к выводу об общих причинах ошибок, заключающихся в том, что трансформеры не обладают глубокой аналогией, абстракцией и избирательностью контента, учитываемого в вычислении ответа (inference). Но наиболее важным выводом является то, что трансформеры, как и другие нейросети, построены по концепции обработки входного сигнала, что создает сильную зависимость от нерелевантной информации, которую не может компенсировать слой внимания трансформера. Концепция нейросетей была заложена в 1950-х идеей перцептрона Ф. Розенблата и не учитывала тех достижений когнитивной психологии, которые появились позже. Согласно же конструктивистской парадигме, входной слой (или перцепция) является только способом проверки правильности сконструированной предиктивной модели для возможных ситуаций. Это же служит причиной самой большой проблемы трансформеров, называемой галлюцинациями. И устранение ее возможно только при изменении архитектуры нейросети, а не за счет большего количества данных в обучении.

Ключевые слова: LLM; трансформеры; мышление; аналогия; когнитивная психология; перцептивный цикл; галлюцинации.

Why don’t transformers think like humans?

Alexander B. Khomyakov,
Master of Physical Sciences
Saint-Petersburg, Russia

alexander.xom@gmail.com

Abstract. Large language models in the form of chatbots very realistically imitate a dialogue as an omniscient interlocutor and therefore have become widespread. But even Google in its Gemini chatbot does not recommend trusting what the chatbot will write and asks to check its answers. In this review, various types of LLM errors such as the curse of inversion, number processing, etc. will be analyzed to identify their causes. Such an analysis led to the conclusion about the common causes of all errors, which is that transformers do not have deep analogy, hierarchy of schemes and selectivity of content taken into account in the inference. But the most important conclusion is that transformers, like other neural networks, are built on the concept of processing the input signal, which creates a strong dependence on superficial noise and irrelevant information that the transformer's attention layer cannot compensate for. The concept of neural networks was laid down in the 1950s by the idea of F. Rosenblatt's perceptron and did not take into account the achievements of cognitive psychology that appeared later. According to the constructivist paradigm, the input word (or perception) is only a way to check the correctness of the constructed predictive model for possible situations. This is the cause of the biggest problem of transformers, called hallucinations. And its elimination is possible only by changing the architecture of the neural network, but not by increasing the amount of data in training.

Keywords: LLM; transformers; thinking; analogy; cognitive psychology; perceptual cycle.

Введение

Современные трансформеры требуют огромных вычислительных ресурсов. Архитектура трансформера масштабируется квадратично по мере увеличения длины входной последовательности. То есть, когда длина последовательности, обрабатываемой трансформером (количество слов в промте), увеличивается, требуемые для обработки вычисления увеличиваются на эту величину в квадрате и быстро становятся неподъемно огромными. Это приводит, например, к невозможности сильно увеличивать окно промта, а его размер в действительности увеличивается за счет алгоритмов сжатия векторов, что неизбежно вызывает потерю информации из промта, особенно фактов, следствием чего является неточность ответа и галлюцинации. Таким образом, трансформеры в действительности не могут работать с длинными последовательностями.

Кроме того, на прошедшей в ноябре 2024 г. конференции NeurlPS Илья Суцкевер заявил, что в мире нет больше данных, чтобы обучить трансформеры в надежде на эффект масштаба (эффект улучшения способностей трансформера при обучении на больших данных). «У нас только один интернет», — сказал он и подчеркнул, что трансформеры, чтобы решить проблему данных, не должны обучаться так, как сейчас [1]. Это ограничение не только данных, но и архитектуры. Это известный факт того, что трансформеры не могут обучаться во время использования, как это делает человек. Но он не ответил на вопрос, а как же должна измениться архитектура нейросети, указав только на то, что они должны стать агентами в среде, как и люди. Необходимость невообразимого количества данных и невозможность обучаться во время использования говорит о том, что трансформеры в своей архитектуре столкнутся с преградой в дальнейшем развитии.

Цель исследования

Но это далеко не все проблемы, которые не преодолимы трансформерами без помощи людей. К ним относят и проблемы с логикой и с вычислениями в больших последовательностях, и следование правилам, и, конечно же, галлюцинации. Такие примеры и будут проанализированы в статье. Отсюда следует все чаще звучащий тезис о том, что необходимо искать новую архитектуру для создания поистине интеллектуальных систем. Данная статья посвящена поиску тех архитектурных принципов, которые могу приоткрыть путь к новой архитектуре. И в этом нам помогут как раз те ошибки трансформеров, которые описаны исследователями. Они прямо указывают на принципиальные недостатки, причины которых могут быть как раз в отсутствии необходимых функций и структур нейросети. И цель исследования состоит в анализе таких ошибок «мышления» трансформеров, чтобы предположить, каких способностей не хватает трансформерам.

Надо отметить, что большинство тестовых заданий, которые приведены ниже, в итоге попадают в Интернет и таким образом оказываются в обучающих данных новых языковых моделей, что позволяет им успешнее их решать. Это называется «утечкой тестовых данных» (data leakage). Более того, компании, создающие большие языковые модели, нанимают десятки тысяч асессоров, которые находят ошибки трансформеров и дообучают их (обу-

чение за счет человеческой обратной связи, RLHF). В результате нейросеть может показывать высокие результаты на тестах не благодаря своим способностям или пониманию заданий, а из-за того, что она натренирована их решать правильно. Такой подход к решению проблем LLM еще раз подчеркивает, что они являются очень хорошими имитаторами с аппроксимацией, но не способны сами прийти к решению.

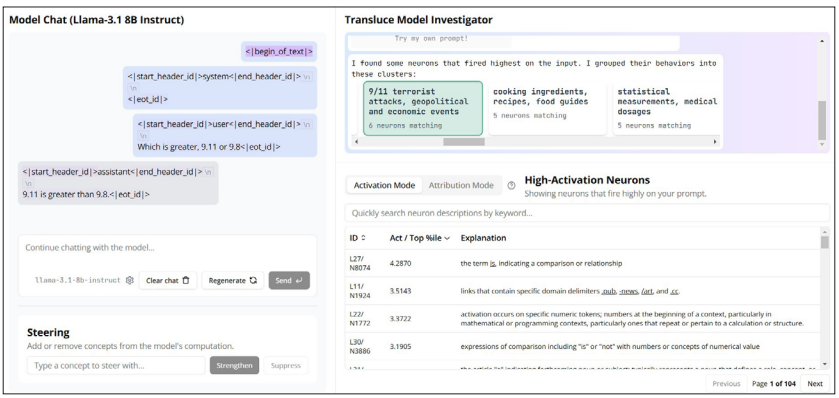
Материалы и методы исследования

Трансформеры называют «черными ящиками», так как интерпретация того, как и почему они генерируют тот или иной ответ, оставалась загадкой. Но с недавнего времени появились инструменты интерпретации LLM, которые позволяют проследить, какие группы нейронов имели большее влияние на результат на выходе трансформера. И удалось также идентифицировать эти группы по тематикам, за которые они отвечают. Такие инструменты появились у Anthropic, у OpenA I. Совсем недавно появился стартап Transluce (https://transluce.org), который создал инструмент Monitor (https://monitor.transluce.org), позволяющий посмотреть, какие группы нейронов активируются при вычислении ответа на введенный промт в llama-3.1-8b-instruct. Это дает возможность увидеть, что повлияло на ответ, какие тематики внесли какой вклад в ответ и на какие слова в ответе больше всего активированы. При этом можно «отключить» или, наоборот, «активировать» эти группы, чтобы посмотреть, как изменится ответ. Это совершенно новый инструмент исследования, который делает трансформеры не такими черными ящиками, как считалось ранее, и позволяет изучать их работу, их «интеллект».

Что больше, 9.8 или 9.11

Например, известная проблема для LLM, заключающаяся в сравнении чисел 9.8 и 9.11. Модели LLM традиционно считают, что 9.11 больше. Так в чем же дело? Мы посмотрели в Монитор Transluce. На рис. 1 видно, что модель (в данном случае Llama-3.1), когда пытается сравнить числа 9.8 и 9.11, активирует несколько неожиданных кластеров: например, про атаку 11 сентября (9.11) и гравитационную константу (9.8), а также нумерацию стихов из Библии, где действительно 9.11 идет раньше 9.8. Вероятно, из-за того, что эти темы появлялись в обучающих данных llama очень

часто, числа 9.8 и 9.11 перестают восприниматься ею как обычные числа: она воспринимает их как другой вид объектов (даты, константы, параграфы).



Но при отключении этих групп нейронов с явно нерелеватными вопросу темами удалось получить верный ответ. Отсюда можно сделать вывод, что трансформеры при определении ответа не могут выбрать ключевую тему и далее не учитывать другие при поиске ответа. Архитектурно трансформер так устроен, что вычисление производится по всем имеющимся параметрам нейросети без избирательности.

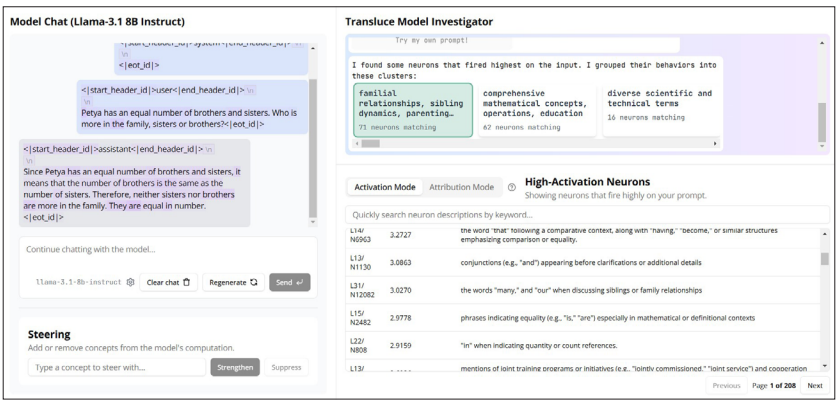
Алиса в стране чудес

В статье «Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models» [4] авторы задают вопрос трансформерам: «У Алисы есть H братья и у нее также есть M сестры. Сколько сестер у брата Алисы?». Они сделали тестовый набор аналогичных задач с разной длиной контекста и провели испытания лучших больших языковых моделей на начало июня 2024 года. Ни одна модель не смогла справиться с заданиями полностью. Только GPT-4 в некоторых случаях при пошаговом объяснении (step-by-step prompt) дает правильный ответ, но ошибается при более сложной формулировке задачи. А люди (с определенного возраста) легко справляются с ней.

В статье Ben Hagag [7] дается анализ того, как LLM Lamma 3.1 отвечает на этот вопрос с помощью монитора Transluce. Автор

подчеркивает, что llama изначально не дает правильный ответ, как и большинство других моделей. Но при отключении активированных нейронов явно неуместных тем (химических соединений и реакций) и усилении активации нейронов темы «гендерные отношения» удалось получить правильный ответ.

Мы взяли другой вопрос для тестирования, на котором ошибаются и такие модели, как ChatGPT-4o. «У Пети братьев и сестер поровну. Кого в семье больше, сестер или братьев». Llama отвечает, как и ChatGPT, что число сестер и братьев в семье одинаково, что ошибочно. Хотя наибольшие активации получают правильные нейроны, отвечающие на семейные отношения. И вторым по количеству нейронов активируется тема математических отношений, что тоже валидно вопросу, так как речь о сравнении количества. Но это не помогает сделать правильный ответ. И попытки отключить или сильнее активировать какие-либо группы нейронов не дали лучшего результата.

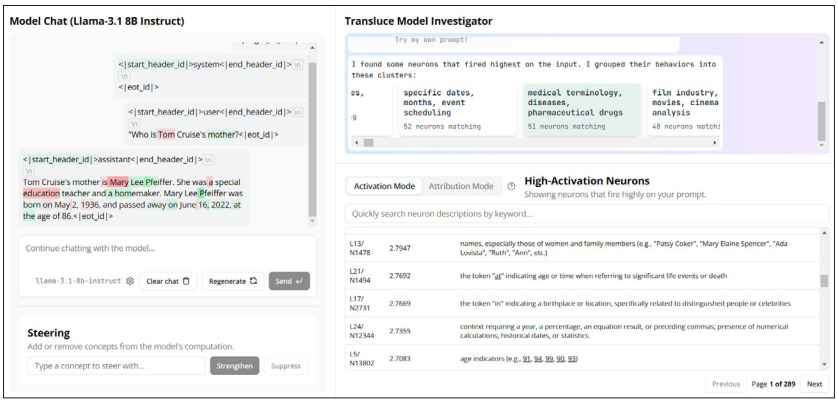


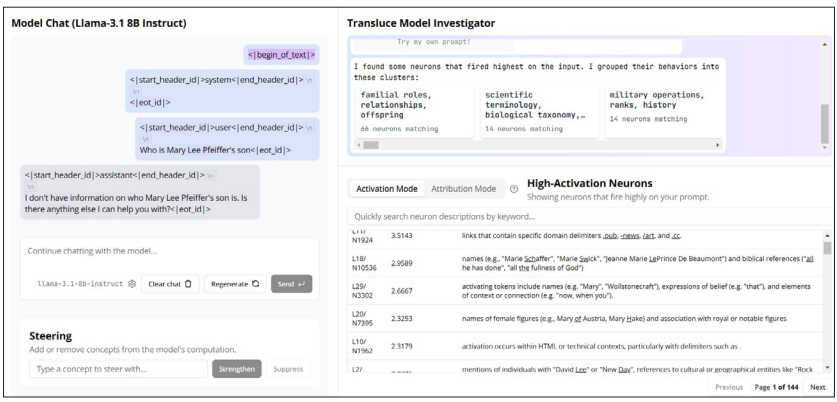
Так в чем проблема решения такой задачи как когнитивного процесса? Представляется, что проблема не только в активации лишних тематических групп нейронов, так как две самые активные группы кажутся достаточными для правильного ответа — семейных отношений и операций сравнения, но и в чем-то ещё, чтобы ответ был верным. Человек выполняет эту задачу в несколько шагов, создавая сначала пространственную картину отношений: вот стоит Петя и вот рядом его, например, три брата и три сестры. Но он тогда четвертый брат в семье, что очевидно при таком пред-

ставлении. И похоже, что таких промежуточных представлений у трансформеров нет. Они пытаются в один шаг найти эвристику, наиболее вероятную для такого запроса. И она неверная.

Проклятие инверсии

Это пример отсутствия у LLM возможности сделать из посылки A->B вывод об отношении B->A, которая в оригинальной статье Lukas Berglund and etc. [6] демонстрируется вопросами “Who is Tom Cruise’s mother? [A: Mary Lee Pfeiffer]” and the reverse “Who is Mary Lee Pfeiffer’s son?”. На второй вопрос LLM не дает ответа, хотя он очевиден. При рассмотрении активации тематических групп в мониторе (рис. 3) можно заметить, что активируется более всего тема ролей в семье и отношений, что верно. Но почему же ответ все равно неверный?

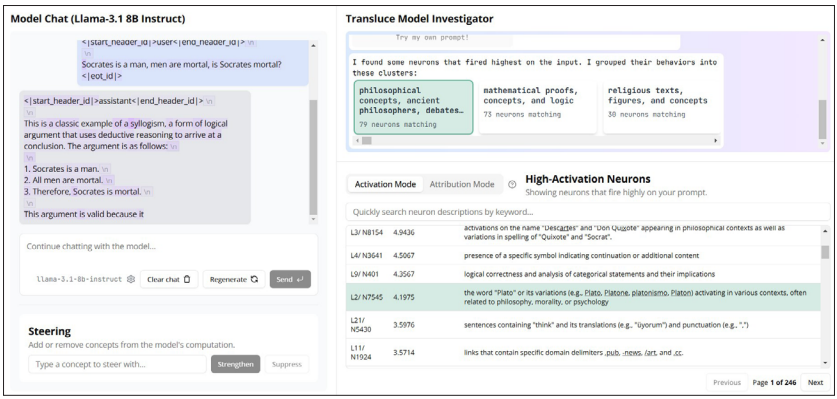


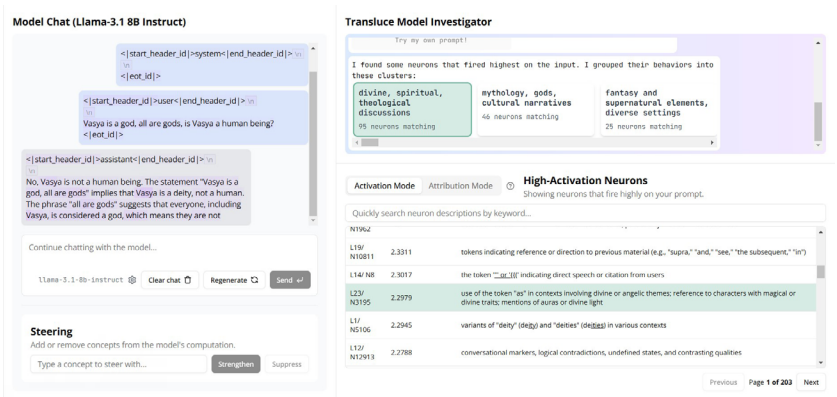


В нейропсихологии известны нарушения в понимании родственных связей, возникающие при повреждении третичных ассоциативных зон в верхней части задневисочной зоны на пересечении с затылочной и теменной. Эта зона отвечает за абстрактные понятия и отношения, включая математические [5]. А. Лурия отмечает, что «общим для всех этих конструкций является то, что ни одну из них нельзя выразить в наглядном образе; все они, таким образом, различными способами кодируют не наглядные, а логические отношения» [5]. Это наталкивает на гипотезу о том, что трансформерам не хватает схем абстрактных обратимых отношений (инверсии), которые особенно характерны для математических операций, пространственных и родственных отношений. Например, если 3+2=5, то 5-2=3. Мы выучиваем не конкретные примеры, а абстрактную схему прибавления и вычитания как обратную. Именно такую абстракцию не может извлечь трансформер, несмотря на обильное обучение на примерах.

Зависимость от формулировок

Мы также решили проверить в мониторе известную проблему трансформеров на зависимость ответа от формулировок запроса в промте. В результате ответ может быть совершенно другой при перефразировке того же вопроса. Мы предложили две формулировки силлогизма: более известную и неизвестную трансформеру (см. рис. 4 и 4а). В первом случае ответ был верен, так как речь шла о Сократе. Трансформер правильно определил философскую тематику — это наиболее активная группа нейронов. Далее активировались нерелевантные темы математических отношений и религиозных текстов, но это не помешало дать правильный ответ. В случае же неизвестной формулировки того же силлогизма трансформер дал неверный ответ. Упоминание бога сделало доминирующей тему «теологического дискурса», а также мифологии и фантастики, что не позволило трансформеру дать правильный ответ. Трансформер при этом не увидел в тексте силлогизм по аналогии, что является ключевой ошибкой. Введение в текст дополнительных слов, не относящихся к вопросу, также нередко меняет ответ трансформеров на неверный. Возникает вопрос, насколько в действительности трансформеры хорошо понимают задание, если их понимание легко сломать переменой слов.





Эти факты означают, что поверхностный слой (вход) нейросети является определяющим для всей последующей цепочки распространения активации и для получения ответа. Понимание человека более устойчиво к перефразировкам. Человек прежде всего воспринимает смысл вопроса, благодаря чему мы легко определяем одинаковый по смыслу вопрос, даже если он выражен иначе. Но что это значит, быть одинаковым по смыслу? Предполагается, что это возможность установить аналогию между высказываниями благодаря функции аналогии, описанной в нашей предыдущей статье [8]. Это означает, что мы мыслим не конкретными выраженными в тексте последовательностями, а некими концептами как совокупностями аналогичных выражений, но которые не привязаны, не зависят от конкретного выражения в поверхностной структуре (тексте). Мы как бы оторваны от него и мыслим на концептуальном уровне при помощи аналогий.

Концепция нейросетей основана на классическом перцептроне. Он был описан Ф. Розенблаттом в 1957 году [9]. В то время в науке царила кибернетика, зарождались компьютерные науки, преобладающим представлением в них была передача и обработка информации, основы его были заложены К. Шенноном в 1948 году. Это подразумевает, что информация содержится вовне нас, мы воспринимаем при помощи перцепции (входной слой нейросетей) и обрабатываем (внутренние слои) для получения результата (выходной слой). На этом основаны все нейросети, но это уже неверно с точки зрения современной когнитивной науки. Мы не воспринимаем концепты извне, а генерируем их в себе. Перцепция же является всего лишь способом выбора и подтверждения той или иной концепции. Так утверждается в многочисленных гипотезах, таких как гипотеза контролируемых галлюцинаций [10] К. Фристона, гипотеза мультимодального пользовательского интерфейса (MUI) Д. Хоффмана [11] и другие конструктивистские концепции в психологии и философии науки. Трансформеры по своей архитектуре не соответствуют этим современным концепциям.

Какой же может быть перспективная архитектура интеллектуальных систем? В этом поиске стоит обратить внимание на перцептивный цикл У. Найсера [2]. Он отличается тем, что совмещает в себе перцептивный подход с конструктивным. В основе перцептивного цикла стоит схема как некий функционал для возможности воспринимать информацию извне. Схема служит для организации информации со всеми ее возможными вариациями, что близко к тому, что утверждается нами про концепты по аналогии выше. Перцепция служит для выбора варианта схемы, но, если ее сочетание не укладывается в какую-либо схему, происходит, по Найсеру, модификация схем. В этом и заключается работа интеллекта, что схоже также с процессами ассимиляции и аккомодации Пиаже. Трансформеры просто предсказывают следующее слово, даже не пытаясь сверять свои концепции с тем, что написано в промте и выдано ими. Но выстраивание такой архитектуры является делом будущего.

Заключение

Приведенные выше примеры неверной работы трансформеров при ответах — это далеко не все примеры, где они ошибаются. Известна проблема понимания отрицания, так как оно реже встре-

чается в обучающих текстах. Это может быть связано с тем же, что указано в разделе про проклятие инверсии: трансформер не может освоить абстрактную обратимую схему утверждения-отрицания. Также известна проблема большой уверенности трансформеров, которые не задают уточняющих вопросов, не сверяются с другими знаниями, имеющимися у них (что выясняется при наводящих вопросах). Это часто является причиной галлюцинаций в ответах трансформеров. У трансформеров возникают проблемы с умножением больших чисел и их сортировкой. Если попросить перемножить большие числа, особенно если числа написаны текстом, даже при рассуждении по шагам трансформеры часто делают ошибки.

Поэтому наша статья не претендует на всеобъемлющий обзор. Она только начало исследований, показывающее возможности, которые открываются для этого такими инструментами, как монитор Transluce. Но уже данное исследование позволяет выдвинуть некоторые гипотезы относительного того, чего же не хватает трансформерам, чтобы приблизиться к интеллекту человеческого уровня не за счет огромного числа примеров, выученных трансформерами на этапе тренировки.

Во-первых, это отсутствие избирательности выбираемых для ответа тем, соответствующих теме вопроса. Это приводит к влиянию нерелевантных групп нейронов и ошибочным ответам.

Во-вторых, трансформерам не хватает общих и доминирующих над контекстом абстракций с обратимыми отношениями, как то: математические, пространственные и родственные отношения, а также отрицания. Это приводит к проблемам с логическими выводами из известных трансформерам фактов.

В-третьих, построение в парадигме обработки информации и полная зависимость от входа (поверхностной структуры) приводят к возможности неверного ответа при иной постановке вопроса.

Общий вывод: нынешние нейросети, включая трансформеры, не являются еще интеллектом уровня человека, и на сегодня очевидно, что для его достижения необходим поиск новой архитектуры интеллектуальных систем.

Литература

1. Robison K. OpenAI cofounder Ilya Sutskever says the way AI is built is about to change // The Verge, Dec 14, 2024. https://www.theverge.com/2024/12/13/24320811/what-ilya-sutskever-sees-openai-model-data-training.

2. Найссер У. Познание и реальность. М.: Прогресс, 1981. С. 42-43. 230 с. (Neisser W. Cognition and Reality. M.: Progress, 1981. P. 42-43. 230 p.)
3. Mitchell M. How do we know how smart ai systems are? // Science, 381(6654). https://www.science.org/doi/10.1126/science.adj5957.
4. Nezhurina M. Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models. https://arxiv.org/html/2406.02061v1#bib.bib12, 04 Jun 2024.
5. Лурия А. Р. Основы нейропсихологии: учеб. пособие. М.: Издательский центр «Академия», 2003. 384 с., с. 123-126. (Luria A. R. Fundamentals of Neuropsychology: textbook. M.: Publishing Center “Academy”, 2003. 384 p., p. 123-126.)
6. Berglund L. and etc. The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”. https://arxiv.org/abs/2309.12288, 26 May 2024.
7. Hagag B. Discover What Every Neuron in the Llama Model Does // Towards Data Science. https://towardsdatascience.com/-0927524e4807, Oct 25, 2024.
8. Хомяков А. Б., Чижик П. Новый способ нахождения аналогов как возможность исследования языка, мышления и построения систем искусственного интеллекта // Философские проблемы информационных технологий и киберпространства. 2024. № 1. С. 77-88. https://doi.org/10.17726/philIT.2024.1.5. (Khomyakov A. B., Chizhik P. A new way of finding analogues as an opportunity to study language, thinking and build artificial intelligence systems // Philosophical Problems of IT & Cyberspace (PhilIT&C). 2024. № 1. P. 77-88. (In Russ.). https://doi.org/10.17726/philIT.2024.1.5.)
9. De La Cruz R. Frank Rosenblatt’s Perceptron, Birth of The Neural Network // Medium. https://medium.com/@robdelacruz/frank-rosenblatts-perceptron-19fcce9d627f, 1 November 2023.
10. Tiehen J. Perception as controlled hallucination // ResearchGate.org. https://www.researchgate.net/publication/359601789_Perception_as_controlled_hallucination march 2022.
11. Хоффман Д. Как нас обманывают органы чувств. М.: АСТ, 2022. (Hoffman D. How Our Senses Deceive Us. M.: AST, 2022.)