Потенциал генеративного искусственного интеллекта для решения профессиональных задач

Ярослав Кузьминов
Научный руководитель, kouzminov@hse.ru

Екатерина Кручинская
Старший преподаватель, кафедра высшей математики, ekruchinskaya@hse.ru

Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), 101000, Москва, ул. Мясницкая, 20

Аннотация

Востребованность генеративного искусственного интеллекта (GenAI) стремительно растет ввиду способности быстро обрабатывать масштабные объемы данных, компилировать их и транслировать «общее мнение». Однако дисбаланс между «компетенциями» GenAI препятствует расширению использования этого инструмента для решения сложных профессиональных задач. ИИ работает как гигантский накопитель и средство воспроизводства знаний, однако не способен их интерпретировать и находить правильное применение в зависимости от контекста. Сохраняется критическая вероятность ошибки при генерации ответов даже на самые простые вопросы.

В статье оценивается степень значимости ограничений, присущих GenAI. Тестирование лежащих в его основе языковых моделей, включая новейшие версии — GPT-4o1 и GigaChat MAX, проводилось с помощью авторского набора вопросов, основанного на таксономии Блума. Установлено, что вероятность получения правильного ответа практически не зависит от количества параметров настройки, сложности и таксономии, а при наличии множественного выбора — снижается. Полученные результаты подтверждают предположение о невозможности применения современных инструментов ИИ в профессиональных целях. Предлагаются опции, способные внести значимый вклад в достижение как минимум квазипрофессионального уровня.

Ключевые слова: профессионализм; генеративный искусственный интеллект; профессиональное использование языковых моделей; графики знаний; оркестровка; таксономия Блума

Цитирование: Kouzminov Y., Kruchinskaia E. (2024) The Evaluation of GenAI Capabilities to Implement Professional Tasks. *Foresight and STI Governance*, 18(4), pp. 67–76.
DOI: 10.17323/2500-2597.2024.4.67.76

The Evaluation of GenAI Capabilities to Implement Professional Tasks

Yaroslav Kouzminov
Academic Supervisor, kouzminov@hse.ru

Ekaterina Kruchinskaia
Senior Lecturer, Department of Higher Mathematics, ekruchinskaya@hse.ru

National Research University Higher School of Economics (HSE University), 20, Myasnitkaya str., Moscow 101000, Russian Federation

Abstract

Generative AI (GenAI) or large language models (LLMs) have been running the world since 2022, but despite all the trends surrounding the use of generative models, these cannot yet be used professionally. While they are most valued for 'knowing everything', nonetheless GenAI models cannot explain and prove. In this way we conceptualize the most recent problem of LLMs as the general trend of mistakes even in the core of knowledge and non-causality of mistake via the complexity of question, as the mistake can be named as an accident and be everywhere as the most limitation of professionalism. At their current stage of development, LLMs are not widely used in a professional context, nor have they replaced human workers. They do not event extend workers' professional abilities.. These limitations of GenAI have one general: non-repayment. This article seeks to analyze GenAI's professional viability by examining two models (GigaChatPro, GPT-4) in three fields of knowledge (economics, law, education) based on our unique Bloom's taxonomy benchmark. To prove our assumption concerning the low possibility of its professional usage, we test three hypotheses: 1) the number of parameters of models have low elasticity regarding difficulty and taxonomy with even the right answer; 2) difficulty and taxonomy jointly have no effect on the correctness of an answer, 3) multiple choice is a factor that decreases the number of right answers of a model. We also present the results of GPT-4 and GigaChat MAX on our benchmark. Finally, we suggest what can be done about the limitations of GenAI's architecture to reach at least a quasi-professional use.

Keywords: professionalism; generative artificial intelligence; professional use of language models; graphs of knowledge; orchestration; Bloom's taxonomy

Citation: Kouzminov Y., Kruchinskaia E. (2024) The Evaluation of GenAI Capabilities to Implement Professional Tasks. *Foresight and STI Governance*, 18(4), pp. 67–76. DOI: 10.17323/2500-2597.2024.4.67.76

Технологии генеративного искусственного интеллекта (*generative artificial intelligence*, GenAI), основанные на больших языковых моделях (*large language models*, LLM), получили широкое распространение в последние несколько лет, особенно с появлением ChatGPT. Подобные инструменты используют в основном для быстрого получения справочной информации, написания и переводов текстов, создания изображений и видеороликов. В качестве ассистента они востребованы в различных областях, включая образование, экономику, финансы, юриспруденцию, медицину и фармацевтику (табл. 1).

Рынок GenAI характеризуется стремительной динамикой развития. К концу текущего года его объем приблизится к 67 млрд долл., что почти втрое больше по сравнению с 2022 г., а к 2030 г. может достигнуть 207 млрд долл.[1] (рис. 1). Российские организации тратят на технологии искусственного интеллекта (ИИ) 15% от общего объема расходов на цифровизацию[2].

Ожидается, что к 2030 г. экономический эффект от инструментов ИИ составит 19.9 трлн долл. (3.5% мирового ВВП)[3], а приносимая ими отдача в 4.6 раза превысит вложения (IDC, 2024). Между тем прирост мирового ВВП в последние годы замедляется. В 2023 г. он составлял 2.9%, а к концу текущего года оценивается на уровне 2.7% (OECD, 2024), что станет самым низким значением со времен начала пандемии COVID-19 в 2020 г. Предположения о том, что распространение GenAI позволит переломить тренд, пока не имеют фактических подтверждений. Как следствие, возникают сомнения в их обоснованности.

В статье аргументируются возможные причины сохраняющейся неспособности ИИ внести значимый вклад в экономический рост.

Возможности и ограничения языковых моделей

Для оценки ожиданий в отношении способностей ИИ решать проблемы и стимулировать экономический рост целесообразно обозначить условную шкалу технологического потенциала. С этой целью приведем две исторические аналогии.

Первая относится к алхимическим практикам Средневековья, связанным с попытками получения золота из других металлов при помощи философского камня. Примечательно, что в XX веке все же удалось экспериментальным путем с помощью ядерной реакции преобразовать ртуть в золото, однако данный способ производства оказался чрезвычайно дорогим и не имеющим перспектив окупаемости[4]. Иными словами, «философским камнем» можно назвать технологии, отличающиеся повышенной трудоемкостью и финансовыми затратами, но так и не реализованные в действительности, несмотря на связываемые с ними большие ожидания. Они становятся первой реперной точкой на нашей шкале.

Вторая аналогия касается разработки, возникшей почти случайно, на стыке двух доминантных технологий, доказавших свою полезность и экономическую эффективность. Речь идет об изобретении паровозов и формировании железнодорожной инфраструктуры — процессах, ставших следствием удачной попытки оснастить средства для транспортировки угля паровым двигателем, который изначально создавался с иными целями (Turnock, 1998). Развитие железных дорог преследовало прагматичную и легко наблюдаемую цель — ускорить процесс перевозки грузов и людей, что сократило бы транзакционные и временны*е* издержки. Любая страна с развитой железнодорожной сетью извлекает колоссальные преимущества, как прямые в виде доходов от эксплуатации, так и косвенные за счет снижения затрат[5]. Технологии с подобными атрибутами будут обозначены на нашей условной шкале как «паровоз».

Попытки создать философский камень также преследовали прикладную цель — минимизировать ресурсоемкость получения золота. Но какие перспективы ожидали бы золото как актив, если бы процесс его добычи стал почти беззатратным, а предложение — избыточным? В таком случае, скорее всего, золото потеряло бы свою биржевую цену. Сегодня оно рассматривается как безусловная «валюта» — залог стабильности для финансо-

Табл. 1. **Перспективные сферы применения GenAI**

Сфера	Ассистирующие функции GenAI	Литература
Образование	Обеспечение учебного процесса	Al-Zahrani, Alasmari, 2024; Ogunleye et al., 2024; Al-Zahrani et al., 2023; Gill et al., 2023; Chu et al., 2022; Dai, Ke, 2022; Hassan et al., 2022
Экономика и финансы	Аналитика и разработка финансовых стратегий	Shapira et al., 2024
Юриспруденция	Консультирование	Alimardani, 2024; Lai et al., 2023
Медицина	Постановка диагнозов	Chen, Esmaeilzadeh, 2024
Фармацевтика	Составление новых лекарственных формул	Choi at al., 2024; Mortlock, Lucas, 2024

Источник: составлено авторами.

[1] https://www.statista.com/forecasts/1474143/global-ai-market-size, дата обращения 27.09.2024.

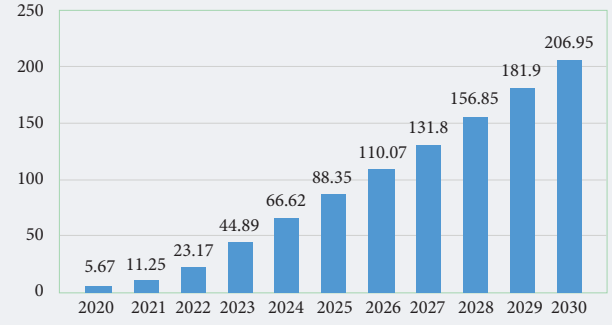
[2] https://issek.hse.ru/news/981416418.html, дата обращения 13.11.2024.

[3] Для сравнения: в 2022 г. доля сельского хозяйства в мировом ВВП составляла 4.27%. https://www.statista.com/statistics/256563/share-of-economic-sectors-in-the-global-gross-domestic-product/, дата обращения 11.10.2024.

[4] https://www.ixbt.com/live/offtopic/pravda-li-chto-mozhno-sdelat-zoloto.html, дата обращения 17.10.2024.

[5] Например, в казну Великобритании за отчетный период с апреля 2022 г. по март 2023 г. за счет пассажирских перевозок поступило около 8.6 млрд ф. ст. (ORR, 2023).

Рис. 1. **Объем мирового рынка GenAI — тенденции и прогноз (*млрд долл.*)**



Источник: https://www.statista.com/forecasts/1474143/global-ai-market-size, дата обращения 18.10.2024.

вого рынка[6]. Возможно, философский камень так и не получил воплощения по причине нереалистичного (и неочевидного) способа использования подобной технологии, поскольку в случае ее гипотетического распространения финансовый рынок очень быстро утратил бы стабильность.

Оценка ресурсных вложений в продуктивизацию упомянутых открытий представляется сложной задачей, не относящейся к цели статьи. Мы стремимся оценить, насколько на данном этапе потенциал ИИ ближе к одному либо другому из них в плане соответствия ожиданиям пользователей, и какие прикладные задачи, выражаемые в измеримых результатах, могут решаться с его помощью. Ответы на поставленные вопросы следуют из выводов обзора литературы и авторского эмпирического эксперимента.

Перед тем как анализировать перспективы профессионального использования языковых моделей, обозначим, что понимается под термином «профессионализм». Речь идет о способности выбирать оптимальные решения в условиях неопределенности, опираясь на накопленную базу знаний, и противостоять распространенным, некомпетентным суждениям, если они противоречат эмпирически проверенным фактам. Так, LLM обычно оперируют широко известными знаниями, исходя из собственной архитектуры (Strachan et al., 2024). Они могут успешно выполнять простые операции, например, агрегировать большие тексты, вычленять их ключевые тезисы (Cheung, 2024; Han et al., 2023). Но подобные функции относятся к рутинным, тогда как профессиональная деятельность подразумевает еще и творчество, работу с новизной. Поскольку генеративные модели в большинстве случаев опираются на поверхностные знания и информацию из низкокачественных источников, возрастает риск получения от них ошибочного ответа даже на элементарный вопрос, что расценивается как неудовлетворительный итог.

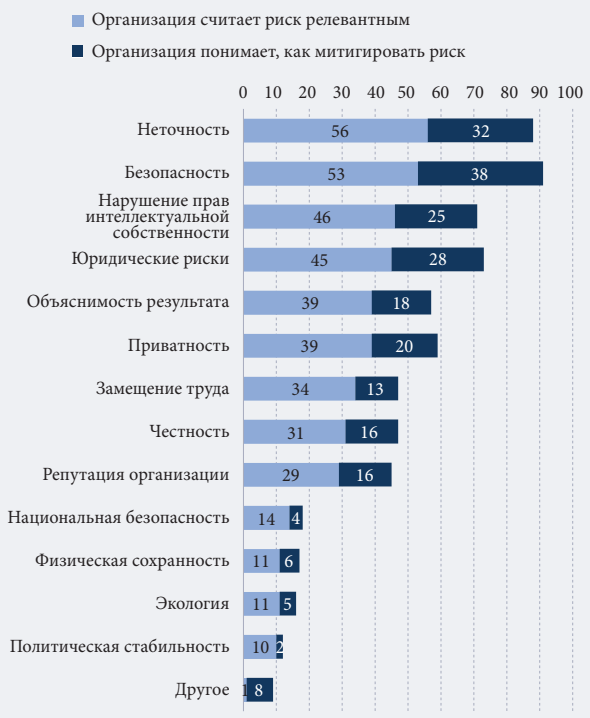
Отсюда следует основное ограничение для полномасштабного делегирования языковым моделям профессиональных задач. Увеличение количества параметров настройки не устраняет общую проблему, что далее иллюстрируется статистической проверкой гипотез. Таким образом, при использовании GenAI для решения рабочих задач обойтись без пристального контроля со стороны человека пока не представляется возможным. Как будет показано эмпирически, тестируемые языковые модели GigaChat Pro и GPT-4 допускают до 50% ошибок в теоретических основах права, образования и экономики, поскольку не обладают базовыми профессиональными знаниями. Все известные методики дообучения пока не могут предложить оптимального решения. Доверие профессиональной работы некомпетентному «помощнику» может быть чревато для компаний не только финансовыми, но и репутационными потерями. Поэтому какой-либо четкой стратегии продуктивизации GenAI до сих пор не выработано.

Другим сдерживающим фактором для расширения профессионального применения LLM является отсутствие эмпирического анализа эффектов от их внедрения в бизнесе, образовании и других прикладных областях (McKnight et al., 2024; Sohail et al., 2023). Как следствие, ни одна генеративная модель в настоящее время не используется как продукт профессионального труда (Noever, Ciolino, 2023). На фоне ожиданий, связываемых с их развитием, возникают очевидные риски неокупаемости, что подтверждается социологическими наблюдениями. По данным Thomson Reuters, почти 60% опрошенных юристов не уверены, что GenAI повлияет на создание добавленной стоимости юридических услуг. Не менее 70% респондентов заявили, что видят наибольшие риски использования LLM в низкой точности ответов, а 57% связывают их с этической некорректностью (Thomson Reuters, 2024). Схожую картину показывает опрос, проведенный McKinsey & Company (рис. 2) – у организаций вызывают беспокойство низкое качество и ошибочность ответов, генерируемых ИИ (56%). Подобный показатель, очевидно, не соответствует критериям профессионализма.

Таким образом, результаты работы ИИ требуют контроля при помощи специальных процедур — промптинга (корректировки запросов), тюнинга (тонкой настройки параметров ответов) и др. Ограничивающими факторами выступают: «смещения» гендерного (Borji, 2023) и лингвистического (Zhang et al., 2024; Zhang et al., 2021) характера, недостаток объективности (Anthis et al., 2024) и логики (Nguyen et al., 2023; Wan et al., 2024). Из-за отсутствия должной производительности, выражаемой в неспособности выполнять функции уровня выше рутинных, генеративные модели все еще нельзя рассматривать как средство для решения сложных задач и извлечения прибыли (Cheung, 2024). По-видимому, ожидания общества в отношении возможностей ИИ пока выглядят необоснованными. Компании-разработчики создали ряд заделов для продвижения ИИ-моделей в сторону профессионализации, но они носят скорее косвенный характер. Например, ведутся разработки по развитию у LLM способностей к логическим рассуждениям (Wei et al.,

[6] Цена декабрьского контракта на золото на Чикагской товарной бирже (Chicago Mercantile Exchange, CME) выросла на 0.77%, поднявшись до 2750.9 долл. за унцию впервые в истории человечества. https://www.barchart.com/futures/quotes/GC*0/profile, дата обращения 22.11.2024.

Рис. 2. **Виды рисков, связанных с использованием больших языковых моделей**



Источник: McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year, дата обращения 17.10.2024.

2022; Liu et al., 2023; Mirzadeh et al., 2024; Xu, 2024). Мы постарались учесть этот тренд, проанализировав новейшие версии языковых моделей, вышедшие во время подготовки публикации, — GPT-4o1 и GigaChat MAX, хотя последние и не являются фокусным объектом исследования. Тем не менее, исходя из общей точности сгенерированных ответов, можно отметить, что GigaChat MAX в целом демонстрирует более значительный прирост по итогам проверки на нашем бенчмарке (Pro vs. MAX +10%) на фоне +5% для GPT-4o1 по сравнению с GPT-4 (см. далее).

Для повышения профессионального уровня языковых моделей чаще всего используют графы знаний (Xu et al., 2024; Luo et al. 2023; Sun et al., 2023). В последнее время активно стали применяться более специализированные инструменты Retrieval Augmented Generation (RAG) или Knowledge Augmented Generation (KAG), направленные на улучшение качества ответов LLM путем формирования «критического мышления» (Sanmartin, 2024; Liang et al., 2024). Чуть менее распространен метод Mixture of Experts (MoE) (Cai et al., 2024; Zhong et al., 2024; Antoniak et al., 2023). Средства оркестровки направлены на совершенствование владения математическими операциями (Rasal, 2024; Zhou et al., 2024). Перечисленные попытки лишь опосредованно содействуют выводу LLM на профессиональное использование, поскольку одного только общего улучшения способности к генерации ответов недостаточно. Возникают риски повышения недостоверности суждений, исходящих от LLM, увеличения ошибок в базовых теориях.

Во многом разрыв между ожиданиями от продуктивизации ИИ и реальностью обусловлен отсутствием культуры машинного обучения. Совершенствованию поддаются только формальные показатели их качества. По данным Института статистических исследований и экономики знаний (ИСИЭЗ) НИУ ВШЭ, большинство студентов в России, имеющих в программе курсы по ИИ, обучаются инженерным специальностям (70.1%). Около четверти (27.2%) связаны с математическими и естественными науками, и только 1.5% имеют отношение к наукам об обществе (НИУ ВШЭ, 2024). Подобный дисбаланс влечет за собой риски в отношении продуктивизации, поскольку разработки по темпам развития опережают продуктовую экспертизу, которой в полной мере владеют только специалисты-прикладники, в большинстве случаев обучающиеся именно на последнем из перечисленных направлений.

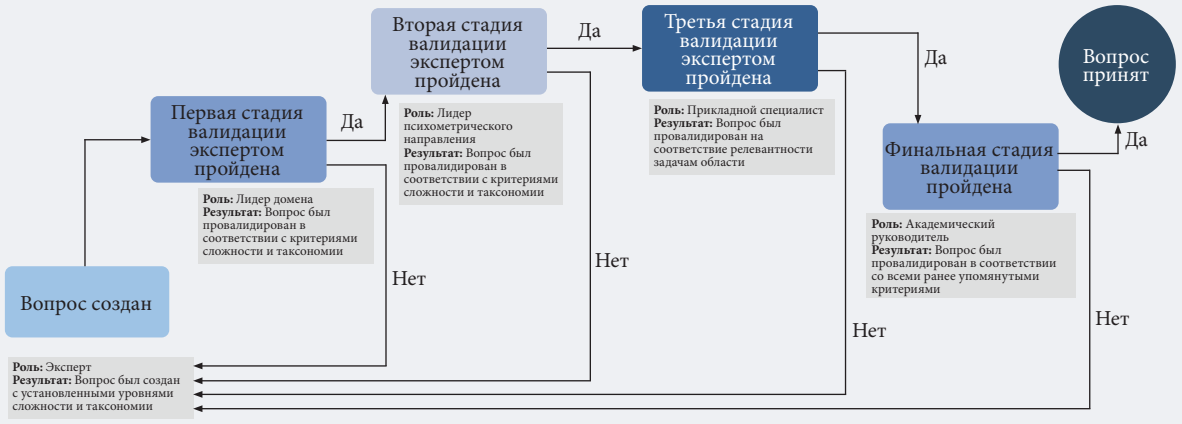
Разработчики LLM нередко ограничиваются увеличением количества параметров настройки и их тестированием на бенчмарках. Как будет продемонстрировано в наших эмпирических экспериментах, подобный подход не имеет ничего общего с проверкой на профессионализм. Между тем компании ожидают от ИИ-моделей качественного выполнения сложных рабочих задач. Наращивание потенциала для профессионального использования LLM невозможно без использования академических принципов в их обучении, которое должно основываться на концепции «ядра знания» (*core of knowledge*) (Kuhn, 1977; Lakatos, 1963; Lakatos, 1970a; Lakatos, 1970b), с тем чтобы ограничить возможность ошибок и не допустить профессиональной некомпетентности ИИ. Предлагается синтезировать увеличение параметров настройки с наращиванием массива качественных академических источников и переосмыслить обучение на ядре знаний путем взаимной увязки предметных областей (возможно, в рамках методики RAG). Необходима разработка способов отслеживания прогресса в обучении ИИ. Профессионализм как критерий является одним из новых поворотов в развитии бенчмарков в целом (Wang, 2024). Наша работа вносит вклад в решение данной задачи, предлагая авторский подход к бенчмаркингу профессионализма и тестированию на нем актуальных версий LLM.

Итак, обзор литературы и результаты социологических опросов указывают на проблему рисков, связанных с профессиональным использованием языковых моделей. В следующем разделе будут приведены эмпирические обоснования в пользу вывода о сохраняющейся невозможности их применения за пределами элементарных рутинных практик.

Эмпирическая оценка степени профессионализма языковых моделей

Для проверки базовых гипотез о низкой способности генеративных моделей выступать в качестве профессионала проводился эксперимент в трех областях знаний (экономика, право, педагогика и образование).

Рис. 3. **Алгоритм экспертной валидации вопросов для тестирования**



Источник: составлено авторами.

Использовался авторский бенчмарк, методологические детали которого описаны в нашем предыдущем исследовании (Kardanova et al., 2024). В соответствии с академическим подходом проверочные задания составлялись на основе следующих критериев:

- *Научность.* Оценивалась способность моделей к работе с профессионально применимыми фундаментальными знаниями и решению практических задач, выходящих за рамки простого изложения фактов.
- *Таксономия и сложность.* Вопросы соотносились с основными измерениями таксономии Блума (Bloom et al., 1956): воспроизведение, понимание, применение; также распределялись по уровням сложности: легкий, средний, трудный.
- *Коллективность.* Сочетались мнения независимых экспертов.
- *Итеративность.* Проводилась поэтапная проверка вопросов по всем критериям таксономии и сложности.

В рамках перечисленных принципов каждое из заданий было создано академическим экспертом в конкретной области, а затем проверено специалистом-психометриком на соответствие классической теории измерения. Вопросы валидировались тремя независимыми экспертами (рис. 3). С помощью созданного набора вопросов оценивались само качество фундаментальных знаний, их сочетаемость и возможность применения в различных контекстах, зачастую неординарных. Разработанная таксономия предназначалась для проверки возможностей языковых моделей на разных уровнях — от запоминания и воспроизведения знаний до их практического применения.

Учитывая тенденцию к наращиванию числа параметров для настройки генеративных моделей, проверке знаний о фактах и теориях посвящалось не более 20% вопросов. В общей сложности создано свыше 13 000 уникальных вопросов, соответствующих стандарту MLUU (Massive Multitask Language Understanding) (Hendrycks et al., 2020). Продемонстрировать их примеры не представляется возможным из-за рисков утечки в претрейн[7]. Сравнивались доли правильных ответов, выдаваемых ИИ в двух режимах: с применением метода Chain-of-Thoughts (CoT) и без него. В первом случае модель «берет время на размышления», во втором — выдает ответ моментально.

В целом, результаты тестирования GPT-4 и GigaChat Pro показывают низкую возможность использования LLM в качестве профессиональных помощников. Наш изначальный тезис заключался в том, что этот фактор статистически значимо связан с вероятностью допущения ИИ ошибок в самых простых вопросах, а ответам нельзя доверять без дополнительной проверки со стороны человека.

Для доказательства предложены следующие гипотезы:

- Шанс получения правильного ответа незначимо связан с количеством параметров настройки моделей (H1), уровнями сложности и таксономии (H2).
- Множественный выбор обусловливает значимое снижение количества верных ответов (H3).

Прежде чем статистически подтверждать гипотезы, следует обратить внимание на средний уровень точности ответов (табл. 2). Значение параметра «температура»[8] для всех тестов равнялось единице. Общие результаты тести-

[7] Претрейн (pre-train) — этап машинного обучения, заключающийся в формировании базы знаний. Наибольшая сложность заключается в выборе максимально полезных сведений из «бесконечного» потока интернет-данных, где возникает парадокс. С одной стороны, если осуществлять обучение модели каждый раз при внесении очередного обновления в набор данных (датасет), то этот процесс окажется слишком ресурсоемким и будет идти медленно. С другой — если проводить «тренинг» после накопления «критического массива» обновлений, то возрастают риски неверного обучающего трека и, как следствие, падения качества генерируемых ответов. Поэтому работа с претрейном — своеобразное искусство балансирования. *Источник*: https://habr.com/ru/companies/yandex/articles/759306/, дата обращения 20.10.2024.

[8] Температура — параметр тонкой настройки случайности для выходных данных языковой модели, измеряемый по шкале от 0 до 1. Понижение его величины приводит к предсказуемым и «традиционным» ответам на запрос пользователя. Напротив, чем выше значение данного показателя, тем больше креативности и разнообразия следует ожидать в выдаваемых результатах. Например, меньший уровень температуры можно использовать для получения фактологических ответов, тогда как ее увеличение оказывается полезным для решения творческих задач. *Источник*: https://learn.microsoft.com/ru-ru/ai-builder/prompt-modelsettings, дата обращения 21.10.2024.

Табл. 2. **Сравнительные показатели доли правильных ответов по тестируемым языковым моделям (%)**

Модель	CoT	Non-CoT
GigaChat Pro	34	38
GPT-4	45	46

Источник: составлено авторами.

Табл. 3. **Матрица сопряженности таксономии и сложности для GigaChat Pro (доля правильных ответов, %)**

Таксономия / уровень сложности	Легкий	Средний	Высокий
Воспроизведение	44.8	37.1	40.3
Понимание	43.3	34.0	31.7
Применение	41.9	33.6	33.7

Источник: составлено авторами.

Табл. 4. **Матрица сопряженности таксономии и сложности для GPT-4 (доля правильных ответов, %)**

Таксономия / уровень сложности	Легкий	Средний	Высокий
Воспроизведение	54.4	46.4	45.8
Понимание	52.7	44.5	40.8
Применение	49.2	45.3	44.3

Источник: составлено авторами.

рования показывают, что GPT-4 не составил серьезной конкуренции GigaChat Pro. Тем не менее, в текущем состоянии ни GigaChat Pro, ни GPT-4 нельзя использовать в профессиональных целях без специальных корректирующих процедур (промптинг, тонкая настройка и т. д.), поскольку фиксируется общая низкая доля правильных ответов (не более 50%). Это означает, что степень владения теоретической базой принципиально важна для дальнейшего дообучения обеих моделей, поскольку пока отсутствует устойчивое владение ядром знаний — дисциплинами, составляющими основу для профессионализма. Отмечена низкая корреляция степеней сложности и таксономии как для GigaChat Pro, так и для GPT-4.

Из табл. 3 видно, что для GigaChat Pro при прочих равных условиях воспроизвести теорию не составляет труда, но не более чем на 45% в самом легком случае. Актуальные характеристики не позволяют говорить о достаточном понимании моделью сути заложенных в нее концепций и выбора контекста для их использования. С задачей простого применения теории, без размышлений о ее соответствии реалиям, GigaChat Pro справляется в целом лучше, хотя разрыв между уровнями понимания и применения в среднем составляет не более 1.5% (в пользу последнего).

Низкий показатель доли верных ответов (менее 50%) указывает на слабое владение теорией, не позволяющее моделям задействовать на практике накопленные знания. Значительный разрыв между компонентами «воспроизведение» и «применение» означает, что само по себе наличие данных в претрейне, вне зависимости от их объема, — недостаточное условие для возможности профессионального использования LLM (табл. 4).

Табл. 5. **Параметры модели для тестирования гипотезы H1**

Зависимая переменная (статус – 0/1)	GPT-4		GigaChat Pro	
	non CoT	CoT	non CoT	CoT
Intercept	0.52***	0.51***	0.45***	0.4***
Difficulty	-0.04***	-0.05***	-0.05***	-0.04***
Taxonomy	-0.01	-0.01	-0.02**	-0.01*
Число наблюдений	13225	13225	13225	13225
AIC	1.665e+04	1.665e+04	1.665e+04	1.665e+04
BIC	1.669e+04	1.669e+04	1.669e+04	1.669e+04

Примечания: Статус: 0 — неправильный ответ; 1 — правильный ответ. * — $p < 0.05$, ** — $p < 0.01$, *** — $p < 0.001$.

Источник: составлено авторами.

Тестовая модель 1 описывается по формуле:

$$Status_i = \beta_0 + \beta_1 * Difficulty_i + \beta_2 * Taxonomy_i + \epsilon_i \quad (1),$$

где: β — константа (значение уравнения при условии, что все переменные будут равны 0); $Difficulty_i$ — уровень сложности; $Taxonomy_i$ — элемент таксономии; ϵ_i — стандартная ошибка (отклонение прогнозной величины от реального значения переменной).

Согласно гипотезе H1, количество параметров настройки почти не влияет на шансы получения верного ответа, что подтверждено в ходе регрессионного анализа по спецификации OLS (см. описание Модели 1 в табл. 5). Статус ответа (правильный либо неправильный) является зависимой переменной, а роль независимых играют степени сложности и таксономия. Несмотря на то что GPT-4 имеет почти 2 трлн параметров против 40 млрд у GigaChat Pro, различия в вероятности получения правильного ответа между ними малозначимые — около 10%. Иными словами, обе модели, несмотря на многократный разрыв в параметрическом многообразии, пока плохо справляются с сопоставлением теории и практики. Более того, в каждом из двух кейсов доли верных ответов оказались выше, когда опция CoT не использовалась. Очевидно, что простое увеличение числа параметров настройки в разработке LLM наталкивается на проблему «потолка производительности», и такого подхода явно недостаточно, чтобы говорить о перспективах дальнейшего совершенствования генеративных моделей.

Другой неожиданный вывод заключается в том, что влияние степени сложности на шанс точного ответа, несмотря на статистическую значимость, имеет низкий коэффициент. При прочих равных условиях усложнение вопроса снижает вероятность корректного ответа лишь на 4–5% для всех моделей. Фактор таксономии статистически значим только для GigaChat Pro, но коэффициент составляет всего 1%.

В соответствии с гипотезой H2 LLM могут совершать ошибки как в сложных, так и в простых задачах. Фактор

Табл. 6. **Параметры модели для тестирования гипотезы H2**

Зависимая переменная (статус – 0/1)	GPT-4		GigaChat Pro	
	non CoT	GPT-4 CoT	non CoT	GigaChat Pro CoT
Intercept	0.54***	0.54***	0.47***	0.41***
Difficulty	-0.06***	–0.1***	–0.07***	–0.05***
Taxonomy	–0.03**	–0.03**	–0.03*	–0.02*
Difficulty* Taxonomy	0.02*	0.03**	0.02	0.01
Число наблюдений	13225	13225	13225	13225
AIC	1.665e+04	1.665e+04	1.665e+04	1.665e+04
BIC	1.669e+04	1.669e+04	1.669e+04	1.669e+04

Примечания: Статус: 0 — неправильный ответ; 1 — правильный ответ. * — p < 0.05, ** — p < 0.01, *** — p < 0.001.

Источник: составлено авторами.

Табл. 7. **Параметры модели для тестирования гипотезы H3**

Зависимая переменная (статус – 0/1)	GPT–4		GigaChat Pro	
	non CoT	CoT	non CoT	CoT
Intercept	0.58***	0.61***	0.58***	0.47***
Difficulty	–0.05***	–0.07***	–0.05***	–0.04***
Taxonomy	–0.03**	–0.03**	–0.03**	–0.02*
Taxonomy* Difficulty	0.01	0.02**	0.01*	0.002
Multiple	–0.23***	–0.41***	–0.24***	–0.35***
Число наблюдений	13225	13225	13225	13225
AIC	1.665e+04	1.665e+04	1.665e+04	1.665e+04
BIC	1.669e+04	1.669e+04	1.669e+04	1.669e+04

Примечания: Статус: 0 — неправильный ответ; 1 — правильный ответ. * — p < 0.05, ** — p < 0.01, *** — p < 0.001.
Источник: составлено авторами.

таксономии не производит заметного эффекта. Гипотеза проверена с помощью тестовой модели 2 (табл. 6):

$$Status_i = \beta_0 + \beta_1 * Difficulty_i + \beta_2 * Taxonomy_i + \beta_3 * (Difficulty_i * Taxonomy_i) + \epsilon_i \quad (2).$$

Результаты показывают, что применительно к GPT-4 значимость взаимодействия таксономии и сложности невелика, а в случае GigaChat Pro она отсутствует. По каждому проверочному вопросу влияние данного фактора незначительно, тем не менее для GPT-4 оно позитивно. Возможное объяснение положительного коэффициента состоит в том, что степень понимания при среднем уровне сложности представляет наибольшую проблему при выполнении заданий GPT-4. В целом, значимые, но небольшие коэффициенты для GPT-4 и несущественные для GigaChat Pro свидетельствуют в пользу гипотезы H2. Есть основания констатировать, что LLM по-прежнему допускают ошибки даже в простейших случаях запоминания, не говоря уже о сложных задачах применения или понимания (процесса, теории и т. д.).

Табл. 8. **Результаты предварительного тестирования новых версий моделей GigaChat и GPT-4**

Языковая модель (режим тестирования)	Доля правильных ответов	Прирост доли по отношению к предыдущим версиям
GigaChat MAX (non-CoT)	49%	+10% в сравнении с GigaChat Pro
GPT-4o1 (non-CoT)	51%	+5% в сравнении с GPT-4

Источник: составлено авторами.

Проверочная модель для гипотезы H3 о том, что множественный выбор обусловливает статистически значимое уменьшение числа корректных ответов для каждого тестируемого объекта, имеет вид:

$$Status_i = \beta_0 + \beta_1 * Difficulty_i + \beta_2 * Taxonomy_i + \beta_3 * (Difficulty_i * Taxonomy_i) + \beta_4 * Multiple_i + \epsilon_i \quad (3),$$

где $Multiple_i$ — фактор множественного выбора.

Расчеты, представленные в табл. 7, показывают статистически значимую, но отрицательную взаимосвязь между фактором множественного выбора и вероятностью правильного ответа для всех языковых моделей. Данное наблюдение, как и значение коэффициента, подтверждает ранее высказанное предположение о том, что LLM хуже справляется с множественным выбором. При использовании CoT вероятность ошибки растет значительно: от 23% до 41%. В целом GigaChat Pro работает с CoT лучше, чем GPT-4. Иными словами, чем разнообразнее задача, тем хуже результаты генеративных моделей в ее выполнении. Налицо противоречие такому атрибуту профессионализма, как умение справляться со сложными ситуациями и неопределенностью.

Напомним, что во время написания статьи появились новые версии LLM — GPT-4o1 и GigaChat MAX, которые нельзя не учитывать. Поэтому они также подверглись оцениванию, хотя и без статистических моделей (табл. 8), что повышает общую актуальность проведенного тестирования.

Полученные результаты подтверждают предположение, согласно которому более высоким потенциалом профессионального развития обладают LLM с меньшим количеством параметров. Дальнейшее увеличение числа последних не приведет к «взрывному» приросту компетенций у ИИ, для обеспечения которого требуются принципиально иные подходы. Детальные испытания языковых моделей будут продолжены.

Заключение

В статье проиллюстрированы современный уровень развития генеративного ИИ и перспективы его профессионального применения. GPT-4 и GigaChat демонстрируют удовлетворительную результативность в переводе и резюмировании текстов, создании иного контента. Тем не менее, работа языковых моделей по-прежнему нуждается в скрупулезном контроле со стороны человека. Следовательно, отсутствуют объективные основания для их признания в качестве самостоятельных профес-

сиональных помощников. Наш эмпирический анализ показал, что LLM ошибаются не только в сложных заданиях, но и в элементарных. Подражая поведению Остапа Бендера[9], они подстраиваются под контекст, но не выявляют научной истины и не используют ее в ответах в качестве норматива. Отсюда вытекает ключевой сдерживающий фактор для профессиональной продуктивизации инструментов ИИ. В качестве аналогий представим, что условный работодатель нанял юриста, допускающего непредвиденные ошибки при толковании положений законодательства, или экономиста, не способного рассчитать оптимум Парето, или преподавателя, не проводящего различий между тьюторством и развивающими методами обучения. Существует неопределенность в отношении того, при решении каких задач помощник допустит больше ошибок. Заранее известно лишь то, что ошибки будут и, строго говоря, их частота не будет зависеть от запаса общих «знаний» ассистента.

Основные гипотезы статьи подтверждены тестированием с применением трех регрессионных моделей и нашли дополнительное обоснование на первичных результатах анализа новых версий LLM — GPT-4o1 и GigaChat MAX. Можно отметить продвижение в их развитии, уровень которого, впрочем, по-прежнему недостаточно.

Новые многообещающие направления в совершенствовании генеративного ИИ связывают с оркестровкой LLM (Zhou et al., 2024), использованием графов для увеличения связности знаний, отбираемых для обучения, имитацией человеческих когнитивных процессов (Jin et al, 2023; Zhu et al., 2024; Wen et al., 2023; Yang et al., 2023), а также разработкой специальных глоссариев, переводящих терминологию ядра знаний в понятную для языковых моделей форму. Обозначенные темы получат раскрытие в наших будущих публикациях.

Авторы благодарят коллег из НИУ ВШЭ Елену Карданову, Элен Юсупову и Ксению Тарасову за организацию работы исследовательского коллектива и уникальную психометрическую экспертизу, ставшую методологической основой статьи. Выражается признательность команде «СалютДевайсы» за открытость к сотрудничеству с академическим сектором, в частности Денису Филиппову, Федору Минькину, Евгению Евдокимову, Сергею Козлову, Валерии Заниной и Ирине Кобзаревой, обеспечившим условия для проведения эксперимента.

Библиография

НИУ ВШЭ (2024) *Подготовка высококвалифицированных кадров в области искусственного интеллекта* (под науч. ред. Л.М. Гохберга), М.: НИУ ВШЭ.

Alimardani A. (2024) Generative artificial intelligence vs. law students: An empirical study on criminal law exam performance. *Law, Innovation and Technology*, 2392932, 1–43. https://doi.org/10.1080/17579961.2024.2392932

Al-Zahrani A., Alasmari T. (2024) Exploring the impact of artificial intelligence on higher education: The dynamics of ethical, social, and educational implications. *Humanities and Social Sciences Communications*, 11(1), 912. https://doi.org/10.1057/s41599-024-03432-4

Al-Zahrani A.M. (2024) From Traditionalism to Algorithms: Embracing Artificial Intelligence for Effective University Teaching and Learning. *IgMin Research*, 2(2), 102–112. https://doi.org/10.61927/igmin151

Anthis J., Lum K., Ekstrand M., Feller A., D'Amour A., Tan C. (2024) *The impossibility of fair LLMs* (ArXiv paper 2406.03198). https://doi.org/10.48550/arXiv.2406.03198

Antoniak S., Krutul M., Pióro M., Krajewski J., Ludziejewski J., Ciebiera K., Król K., Odrzygóźdź T., Cygan M., Jaszczur S. (2023) *Mixture of Tokens: Continuous MoE through Cross-Example Aggregation* (ArXiv paper 2310.15961). https://doi.org/10.48550/arXiv.2310.15961

Bloom B.S., Engelhart M.D., Furst E.J., Hill W.H., Krathwohl D.R. (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals* (Handbook 1: Cognitive Domain), Ann Arbor, MI: Edwards Bros.

Borji A. (2023) *A categorical archive of Chat GPT failures* (ArXiv paper 2302.03494) https://doi.org/10.48550/arXiv.2302.03494

Cai W., Jiang J., Wang F., Tang J., Kim S., Huang J. (2024) *A Survey on Mixture of Experts* (ArXiv paper 2407.06204). https://doi.org/10.48550/arXiv.2407.06204

Chen Y., Esmaeilzadeh P. (2024) Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges. *Journal of Medical Internet Research*, 26, e53008. https://doi.org/10.2196/53008

Cheung M. (2024) *A Reality check of the benefits of LLM in business* (ArXiv paper 2406.10249). https://doi.org/10.48550/arXiv.2406.10249

Choi J., Palumbo N., Chalasani P., Engelhard M.M., Jha S., Kumar A., Page D. (2024) *MALADE: Orchestration of LLM-powered Agents with Retrieval Augmented Generation for Pharmacovigilance* (ArXiv paper 2408.01869). https://doi.org/10.48550/arXiv.2408.01869

Chu H.C., Hwang G.H., Tu Y.F., Yang K.H. (2022) Roles and research trends of artificial intelligence in higher education: A systematic review of the top 50 most-cited articles. *Australasian Journal of Educational Technology*, 38(3), 22–42.

Dai C.-P., Ke F. (2022) Educational applications of artificial intelligence in simulation-based learning: A systematic mapping review. *Computers and Education: Artificial Intelligence*, 3, 100087. https://doi.org/10.1016/j.caeai. 2022.100087

Gill S.S., Xu M., Patros P., Wu H., Kaur R., Kaur K., Fuller S., Singh M., Arora P., Kumar A.P., Stankovski V., Abraham A., Ghosh S.K., Lutfiyya H., Kanhere S.S., Bahsoon R., Rana O., Dustdar S., Sakellariou R., Uhlig S., Buyya R. (2023) Transformative Effects of ChatGPT on Modern Education: Emerging Era of AI Chatbots. *Internet of Things and Cyber-Physical Systems*, 4, 19–23. https://doi.org/10.1016/j.iotcps.2023.06.002

Han S.J., Ransom K.J., Perfors A., Kemp C. (2023) Inductive reasoning in humans and large language models. *Cognitive Systems Research,* 83, 1–28. https://doi.org/10.1016/j.cogsys.2023.101155

Hassan R., Ali A., Howe C.W., Zin A.M. (2022) Constructive alignment by implementing design thinking approach in artificial intelligence course: Learners' experience. *AIP Conference Proceedings*, 2433(1), 0072986. https://doi.org/10.1063/5.0072986

Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., Steinhardt J. (2020) Measuring Massive Multitask Language Understanding (ArXiv paper 2009.03300). https://doi.org/10.48550/arXiv.2009.03300

IDC (2024) *The Global Impact of Artificial Intelligence on the Economy and Jobs*, Needham, MA: IDC Corporate.

Jin B., Liu G., Han C., Jiang M., Ji H., Han J. (2023) *Large Language Models on Graphs: A Comprehensive Survey* (ArXiv paper 2312.02783). https://doi.org/10.48550/arXiv.2312.02783

Kardanova E., Ivanova A., Tarasova K., Pashchenko T., Tikhoniuk A., Yusupova E., Kasprzhak A.G., Kuzminov Y., Kruchinskaia E., Brun I. (2024) *A Novel Psychometrics-Based Approach to Developing Professional Competency Benchmark for Large Language Models* (arXiv paper 2411.00045). https://doi.org/10.48550/arXiv.2411.00045

[9] Остап Бендер — герой романа И. Ильфа и Е. Петрова «Двенадцать стульев», отличающийся способностями к излишнему изобретательству, подстраиванию под любого собеседника, воспроизводству ложных фактов. В иностранной литературе ближайшим аналогом такого персонажа можно считать барона Мюнхгаузена.

Kuhn T.S. (1977) *The Essential Tension*, Chicago: University of Chicago Press.
Lai J., Gan W., Wu J., Qi Z., Yu P.S. (2023) *Large Language Models in Law: A Survey* (ArXiv paper 2312.03718). https://doi.org/10.48550/arXiv.2312.03718
Lakatos I. (1963) Proofs and Refutations (I). *British Journal for the Philosophy of Science,* 14(53), 1–25.
Lakatos I. (1970a) Falsification and the Methodology of Scientific Research Programmes. In: *Criticism and the Growth of Knowledge* (eds. I. Lakatos, A. Musgrave), Aberdeen: Cambridge University Press, pp. 91–195.
Lakatos I. (1970b) History of Science and Its Rational Reconstructions. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, pp. 91–136.
Liang L., Sun M., Gui Z. et al. (2024) KAG: Boosting LLMs in Professional Domains via Knowledge Augmented Generation. ArXiv paper 2409.13731, 1–33. https://doi.org/10.48550/arXiv.2409.13731
Liu N.F., Lin K., Hewitt J., Paranjape A., Bevilacqua M., Petroni F., Liang P. (2023) *Lost in the Middle: How language models use long contexts* (ArXiv paper 2307.03172). https://doi.org/10.48550/arXiv.2307.03172
Luo L., Li Y.F., Haffari G., Pan S. (2023) *Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning* (ArXiv paper 2310.01061). https://doi.org/10.48550/arXiv.2310.01061
McKnight M.A., Gilstrap C.M., Gilstrap C.A., Bacic D., Shemroske K., Srivastava S. (2024) Generative Artificial Intelligence in Applied Business Contexts: A systematic review, lexical analysis, and research framework. *Journal of Applied Business and Economics*, 26(2), 7040. https://doi.org/10.33423/jabe.v26i2.7040
Mirzadeh I., Alizadeh K., Shahrokhi H., Tuzel O., Bengio S., Farajtabar M. (2024) *GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models* (ArXiv paper 2410.05229). https://doi.org/10.48550/arXiv.2410.05229
Mortlock R., Lucas C. (2024) Generative artificial intelligence (Gen-AI) in pharmacy education: Utilization and implications for academic integrity: A scoping review. *Exploratory Research in Clinical and Social Pharmacy,* 15, 100481. https://doi.org/10.1016/j.rcsop.2024.100481
Naveed H., Khan A.U., Qiu S., Saqib M., Anwar S., Usman M., Akhtar N., Barnes N., Mian A. (2023) *A comprehensive overview of large language models* (ArXiv paper 2307.06435). https://doi.org/10.48550/arXiv.2307.06435
Nguyen H., Fungwacharakorn W., Satoh K. (2023) *Enhancing logical reasoning in large language models to facilitate legal applications* (ArXiv paper 2311.13095). https://doi.org/10.48550/arXiv.2311.13095
Noever D., Ciolino M. (2023) *Professional Certification Benchmark Dataset: The first 500 jobs for large language models* (ArXiv 2305.05377). https://doi.org/10.48550/arXiv.2305.05377
OECD (2024) *OECD Economic Outlook* (Interim Report, September 2024), Paris: OECD.
Ogunleye B., Zakariyyah K.I., Ajao O., Olayinka O., Sharma H. (2024) A Systematic Review of Generative AI for Teaching and Learning practice. *Education Sciences*, 14(6), 14060636. https://doi.org/10.3390/educsci14060636
ORR (2023) *Rail industry finance (UK): April 2022 to March 2023,* London: Office of Rail and Road.
Rasal S., Hauer E.J. (2024) Navigating Complexity: Orchestrated Problem Solving with Multi-Agent LLMs (ArXiv paper 2402.16713). https://doi.org/10.48550/arXiv.2402.16713
Sanmartin D. (2024) *KG-RAG: Bridging the gap between knowledge and creativity* (ArXiv paper 2405.12035). https://doi.org/10.48550/arXiv.2405.12035
Shapira E., Madmon O., Reichart R., Tennenholtz M. (2024) *Can LLMs replace economic choice prediction labs? The case of language-based persuasion games* (ArXiv paper 2401.17435). https://doi.org/10.48550/arXiv.2401.17435
Sohail S.S., Faiza Farhat F., Himeur Y., Nadeem M., Madsen D.O., Singh Y., Atalla S., Mansoor W.. (2023) Decoding ChatGPT: A taxonomy of existing research, current challenges, and possible future directions. *Journal of King Saud University – Computer and Information Sciences*, 35(8). https://doi.org/10.1016/j.jksuci.2023.101675
Strachan J., Albergo D., Borghini G., Pansardi O., Scaliti E., Gupta S., Saxena K., Rufo A., Panzeri S., Manzi G., Graziano M.S.A., Becchiol C. (2024) Testing theory of mind in large language models and humans. *Nature Human Behaviour,* 8(7), 1285–1295. https://doi.org/10.1038/s41562-024-01882-z
Sun J., Xu C., Tang L., Wang S., Lin C., Gong Y., Ni L.M., Shum H.Y., Guo J. (2023) *Think-on-Graph: Deep and responsible reasoning of large language model on knowledge graph* (ArXiv paper 2307.07697). https://doi.org/10.48550/arXiv.2307.07697
Thomson Reuters (2024) *2024 Generative AI in Professional Services*, Toronto: Thomson Reuters Institute.
Turnock D. (1998) *An Historical Geography of Railways in Great Britain and Ireland* (1st ed), New York: Routledge.
Wan Y., Wang W., Yang Y., Yuan Y., Huang J., He P., Jiao W., Lyu M.R. (2024) $A \wedge B \Leftrightarrow B \wedge A$: *Triggering logical reasoning failures in large language models* (ArXiv paper 2401.00757). https://doi.org/10.48550/arXiv.2401.00757
Wang Y., Ma X., Zhang G., Ni Y., Chandra A., Guo S., Ren W., Arulraj A., He X., Jiang Z., Li T., Ku M., Wang K., Zhuang A., Fan R., Yue X., Chen W. (2024) *MMLU-Pro: A more robust and challenging Multi-Task Language Understanding benchmark* (ArXiv paper 2406.01574). https://doi.org/10.48550/arXiv.2406.01574
Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., Ed H., Quoc C.V., Zhou L.D. (2022) *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (ArXiv paper 2201.11903). https://doi.org/10.48550/arXiv.2201.11903
Wen Y., Wang Z., Sun J. (2023) *MindMap: Knowledge Graph prompting sparks graph of thoughts in large language models* (ArXiv paper 2308.09729). https://doi.org/10.48550/arXiv.2308.09729
Xu Z., Cruz M.J., Guevara M., Wang T., Deshpande M., Wang X., Li Z. (2024) *Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering* (ArXiv paper 2404.17723). https://doi.org/10.48550/arXiv.2404.17723
Yang L., Chen H., Li Z., Ding X., Wu X. (2023) *Give Us the Facts: Enhancing Large Language Models with Knowledge Graphs for Fact-aware Language Modeling* (ArXiv paper 2306.11489). https://doi.org/10.48550/arXiv.2306.11489
Zhang Y., Ding H., Shui Z., Ma Y., Zou J., Deoras A., Wang H. (2021) *Language models as recommender systems: Evaluations and limitations*. Paper presented at the NeurIPS 2021 Workshop on I (Still) Can't Believe It's Not Better.
Zhang Y., Sun R., Chen Y., Pfister T., Zhang R., Arik S.O. (2024) *Chain of Agents: Large language models collaborating on Long-Context Tasks* (ArXiv paper 2406.02818). https://doi.org/10.48550/arXiv.2406.02818
Zhong Z., Xia M., Chen S., Lewis M. (2024) *Lory: Fully Differentiable Mixture-of-Experts for Autoregressive Language Model Pre-training* (ArXiv paper 2405.03133). https://doi.org/10.48550/arXiv.2405.03133
Zhou J.P., Luo K.Z., Gu J., Yuan J., Weinberger K.Q., Sun W. (2024) *Orchestrating LLMs with Different Personalizations* (ArXiv paper 2407.04181). https://doi.org/10.48550/arXiv.2407.04181
Zhu Y., Wang X., Chen J., Qiao S., Ou Y., Yao Y., Deng S., Chen H., Zhang N. (2023) *LLMS for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities* (ArXiv paper 2305.13168). https://doi.org/10.48550/arXiv.2305.13168