Оценка моделей LLM по степени готовности решать задачи управления в области ESG

Л.А. Мыльников[1], М. А. Сторчевой[1],
[1]Национальный исследовательский университет «Высшая школа экономики»

В. В. Чернышев[2], А.В. Булатов[2], Н.Е. Холоднов[2]
[2]Федеральная служба по надзору в сфере природопользования

А.А. Баютина[3], В.И. Кондратьева[3]
[3]ФГБУ «Государственный научно-исследовательский институт промышленной экологии»

Аннотация. Внимание к охране природы принимает все большую значимость для бизнеса с одной стороны в связи с ужесточением в природоохранном законодательстве, а с другой в связи с использованием ESG рейтингов при принятии решений о коммерческой деятельности компаний. Составление рейтинга LLM систем, способных оказывать консультационные услуги в области природоохраны и ESG, позволяет осуществить выбор такой системы для использования в своей деятельности, что позволит как сократить текущие расходы на обеспечение этой деятельности, так и снизить объем возможных штрафов от принятия неверных решений. Ранжирование существующих LLM осуществляется на основе эталонных ответов. Для ранжирования выбраны LLM, использующие разные архитектуры нейронных сетей, а также сформулированы группы вопросов, сгруппированные по тематикам, предполагаемой форме ответа и сложности. Для ранжирования использован подход на основе оценки когерентности ответов LLM с эталонными ответами подготовленными экспертами.

Ключевые слова: LLM, ESG, рейтинг, сравнение текстов, когерентность, Q&A, ранжирование, экспертная деятельность

Введение

Использование чат-ботов и генеративных моделей ИИ является перспективным направлением в организации консультационной поддержки в самых разных сферах, и в том числе – в области выработки управленческих решений в частных компаниях или в государственном секторе. Выгода данной технологии заключается в том, что они могут заменить большое количество экспертов [4] или сократить объем их работы, подготовив часть информации в автоматическом режиме. Подготовка экспертов и оплата их труда может приводить к значительным финансовым издержкам при отсутствии обоснованности, а иногда и согласованности выбираемых ими решений и в условиях недостатка или недоступности в оперативном режиме соответствующего эксперта. Кроме того, использование LLM позволяет обеспечить и более высокий уровень консультационной поддержки за счет с одной стороны скорости ответа, а с другой стороны полноты предоставляемой информации (более развернутых ответов). На большом количестве интернет-ресурсов и мобильных приложений уже используются чат-боты, которые позволяют пользователям быстро получить подсказку, не ожидая возможности контакта с оператором, что экономит компаниям значительные ресурсы. С появлением LLM данные возможности значительно расширяются – теперь вместо специализированных чат-ботов можно развивать универсальные системы, которые можно применять к самым разным прикладным отраслям, использовать уже существующую документацию для их обучения, получать ответы на вопросы со сложными формулировками.

Традиционные вопросно-ответные системы такие как Information Retrieval-based Question Answering (IR-based) и Named Entity Linking (NEL) хорошо себя показывают при ответах на закрытые вопросы, а также на вопросы об отношениях между сущностями (например авторстве, месте рождения известных людей, наличии того или иного компонента в рецепте, именах литературных персонажей, столицах стран и т. п.). Для этого они

используют базы знаний (такие как DBpedia[1] и WikiData[2]) и специализированные базы данных (OpenStreetMap и т.д.).

Использование LLM позволяет получать ответы, в том числе, и на открытые вопросы. Для этого необходимо их обучение на больших объемах текстовых данных и наличие механизма ассоциативного поиска в них. Использование для обучения открытых источников и баз данных таких как Википедия, образовательные ресурсы университетов, юридические базы данных, базы часто сдаваемых вопросов и ответов, внутренние базы знаний организаций, информации из новостных сообщений и т.п. даёт LLM возможность давать ответы на самые разные вопросы, в том числе, и на вопросы ответы на которые могут быть получены только по результатам сбора и обработки больших объемов информации. Для ответов на вопросы данные должны содержать достоверную информацию. Особенности используемых моделей и данных для их обучения отражается в разной степени точности их ответов. Это связано с тем, что при обучении не осуществляется проверка на согласованность используемых текстов [10], что может проявляться у разных моделей в разных ответах на одни и те же вопросы (может быть даже противоречащих друг другу как случается и у реальных экспертов), а также в появлении явления получившего название галлюцинации LLM (когда факты или источники «придумываются» или искажаются моделью, что так же бывает и у людей и не всегда намеренно, а например, в силу забывчивости).

Во многих случаях ответ LLM может быть крайне важным для принятия правильного управленческого решения, поскольку ошибка может привести как к существенным затратам для организации (штрафы, компенсации, вынужденные расходы и т. п.), так и к чувствительному негативному воздействию на заинтересованные стороны организации (прекращение поставки услуг, некачественные услуги, экологический вред и т. д.).

Компетентность LLM в области управления экологической и социальной ответственностью компании (ESG) имеет с одной стороны огромное значение с точки зрения оказания помощи менеджерам и государственным служащим, а с другой стороны представляет собой довольно сложную задачу, поскольку многие вопросы в области ESG имеют нечеткий характер, свойственный для гуманитарного или философского знания, при этом они связаны с необходимостью знания законодательства и стандартов в конкретных областях деятельности. Например, природоохранное законодательство различно в разных странах не только с точки зрения требований по охране природе, правоприменительной практике, но и по своей структуре, проявляющейся в множестве контролируемых параметров, применяемых для сохранения природы решениях, способах оценки влияния на природу загрязняющих веществ и так далее. Также в сфере экологии могут наблюдаться культурные различия даже в разных регионах одной страны, большая динамика изменений и недостаток данных в свободном доступе.

Ранжирование LLM по степени компетентности в сфере природопользования необходимо при использовании для повышения компетентности сотрудников компаний в области устойчивого развития, экологической и социальной ответственности компаний; для оценки возможности их использования для получения необходимых сведений при подготовке отчетности в области экологии в контрольно-надзорные органы; контролирующие органы, в свою очередь, заинтересованы в инструменте, позволяющем ускорять процесс проверки сведений предоставляемых в отчетной документации.

Источники информации в области природопользования в настоящее время ограничены юридическими базами, справочными документами, которые готовят юридические компании по разным областям деятельности (например, см. пояснения к закону о плате за негативное воздействие на окружающую среду — https://www.consultant.ru/document/cons_doc_LAW_154375/, экологический сбор — https://www.consultant.ru/document/cons_doc_LAW_210784/), информационными

[1] https://www.dbpedia.org/

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

материалами Росприроднадзора (например, по уплате экологического сбора — https://rpn.gov.ru/activity/rop/ecological-fee/, по уплате сборов за негативное воздействие на окружающую среду — https://rpn.gov.ru/activity/environment-fee/) и сборниками нормативной документации подготовленными организациями работающими в области охраны природы (см. например, https://greenium.ru/regulation/).

Наличие таких ресурсов не закрывает потребности в получении информации в области природопользования, что подтверждается большим число обращений и запросов, поступающих в центральный аппарат Росприроднадзора. Всего за 2023 год только в центральный аппарат Росприроднадзора поступило 8 212 обращений (из Администрации Президента Российской Федерации – 1 565 обращений, из Минприроды России – 1 550 обращений; с официального сайта Росприроднадзора – 3 781 обращения; из других источников – 1 316). Более половины поступивших обращений касается загрязнения окружающей среды выбросами и сбросами различных предприятий, а также санкционированных и несанкционированных размещений всех видов отходов (4 731 обращение). Далее следуют обращения о нарушениях водного законодательства, строительства в водоохранных зонах, осуществляющегося с нарушением законодательства об охране окружающей среды и нанесением ущерба экологии и населению (682 обращения), о жестоком обращении с животными (589 обращений), по вопросу проведения государственной экологической экспертизы (365 обращений[3]).

Кроме обозначенных информационных ресурсов и возможности обращения в ответственную за природоохрану организацию существуют и специализированные организации, оказывающие консультационные услуги и услуги по подготовке документов в области охраны природы, которые тоже находят своих клиентов. Многие крупные компании имеют в своем составе соответствующих специалистов или экспертные подразделения в области охраны природы.

Методология оценивания LLM

Ранжирование и оценка моделей LLM и систем Q&A на данный момент остается затруднительной в связи со сложностью оценки качества ответов без привлечения экспертов. Существующие LLM не раскрывают информацию об используемых ими источниках данных, а широко используемые в Q&A системах на таких наборах данных как KGQA, LC-QuAD и QALD, не вызывают доверия в связи с неполнотой и недостаточной актуальностью информации. В связи с этим появляются работы, посвященные оценкам существующих систем как по качеству ответов так и использующие выводы полученные в отдельных исследованиях для получения обобщенных оценок [9].

Решение задачи оценки правильности ответов связано с задачей оценки степени похожести текстов. При этом многие смысловые связи могут быть сформулированы с использованием разной лексики и различных способах построения фраз. Для оценки степени похожести выделяют тематическую, сущностную и риторическую когерентности. Все из них измерить не представляется возможным. Поэтому на практике оценивают отличия в используемых словах и структурах фраз.

Для тестирования Q&A и LLM систем в отдельных областях разработаны специализированные наборы вопросов-ответов. Например, разработан MMLU (Measuring Massive Multitask Language Understanding) тест с использованием которого исследователи измеряют понимание языка изучаемыми LLM. Тест опирается на большую базу размеченных по категориям вопросов и вариантов ответов (множество возможных корректных ответов), что делает возможным применение традиционных для ML моделей метрик (Precision и Recall [5]) [3], а также использование рейтингов ответов (MRR, mean reciprocal rank) для

[3] https://rpn.gov.ru/upload/iblock/da8/x8cgrbq0y6fk8ziy6cdmf51s9e09nk9q/Doklad-2023-_1_.pdf

определения того как хорошо система выбирает свой ответ, если варианты ответов могут быть отранжированы.

Контекст таких тестов использует общие знания и не учитывает специфику и особенности манипуляции с информацией и данными в специальных предметных областях. При этом существуют специализированные базы задач по математике (например, GSM8k) и программированию HumanEval (тест на решение задач программирования). Такие тесты предполагают только один или ограниченный тип ответов (например, число, которое можно сравнить с ответом, программа, которую можно запустить тем самым проверив её работоспособность и результат, который она выдаст и т.п.).
Составлять подобные тесты для всех областей знания затруднительно и не имеет смысла в связи с тем, что это потребует с одной стороны огромных ресурсов на их создание и актуализацию, а с другой поставит вопрос о необходимости LLM.

Другая группа оценки эффективности LLM и Q&A является группа методов использующая токенизацию ответов. В этом случае при наличии эталонных ответов на задаваемые вопросы становиться возможным применение метрик используемых в системах машинного перевода — BLEU, ROUGE, METEOR, TER [2], а также метрик, опирающихся на меры когерентности [1], количества информации, энтропию Шеннона, а также такие традиционные метрики как корреляция, TF-IDF, косинусное расстояние, расстояние Левенштейна, сходства Жаккара, Рэнда и т.п.

Применение метрик возможно только при наличии материала, который можно сравнивать; таким образом, необходимо определиться с типами вопросов, которые будут задаваться LLM и перечнем LLM, которые будем сравнивать.

Еще одной возможностью оценки качества является использование LLM, таких как BERT и др. для оценки соответствия между векторами термов. Преимуществом такого подхода может быть использование эмбединнгов для получения промежуточных оценок, похожих или аналогичных описанным выше, и возможность их корректировки за счет учета мнения экспертов и построения сложных нелинейных зависимостей между значениями оценок. Однако на практике для этого необходимы большие объемы размеченной информации, собрать которые затруднительно.

Принципы формирования множества вопросов для оценки LLM

Особенности вопросов, которые передаются для обработки в LLM могут являться темой отдельных исследований. Размер подаваемого на вход LLM сообщения является одной из характеристик языковых моделей. Список контрольных вопросов для LLM должен отвечать нескольким условиям:

- характеризоваться полнотой, то есть охватывать все возможные области деятельности в области природопользования, в которых у представителей бизнеса или государственных органов власти могут возникать потребность;
- вопросы должны включать в себя все наиболее актуальные проблемы, которые возникают в практической деятельности – от концептуальных (напр., почему существует данное ограничение) до очень практических (как заполнить заявление или как рассчитать размер платежа).

Все вопросы можно объединить в группы и подгруппы, которые будут как облегчать составление вопросов, так и последующий анализ ответов, поскольку лексические характеристики ответов на вопросы из одной группы должны быть очевидным образом похожи.

Такие группы могут быть сформированы:

1) **по тематикам вопросов**:

- нормативно-правовые вопросы (о содержании нормативно-правовых актов);
- расчётно-финансовые вопросы (о применении методик расчёта платы за негативное воздействие на окружающую среду и экологического сбора);
- терминологические вопросы (об определениях терминов и их трактовках);

- гносеологические вопросы (о знаниях о природе вещей).

2) **по форме ответа**: с бинарным ответом (да / нет); с числовым и с текстовым ответами.

3) **по уровню сложности будем выделять**:

- *Уровень 1.* Вопросы, ответ на которые содержится в текстах, на которых обучалась LLM в явном виде.
- *Уровень 2:* Вопросы, ответ на которые можно получить логическим выводом или используя вычисления с использованием к информации (присутствующей в явном виде) из текстов, на которых обучалась модель.
- *Уровень 3:* Вопросы, ответ на которые можно получить логическим выводом или с использованием математических операций, обладая дополнительной информацией к той, на которой обучалась модель (правила надо достать из документов, правила предметной области).
- *Уровень 4:* Вопросы, ответ на которые можно получить, обладая дополнительной информацией об объекте или субъекте, о котором задается вопрос. Для ответа на вопрос нужны обобщения, которые можно собрать из общедоступной информации (новостей, статистических данных и т.п.).

4) **по вопросительному слову**, на основании которого строится запрос, например: «Кто?», «Когда?», «Где?», «Почему?», «Сколько?», «Что?», «Как?» и т.д.

Подготовив сбалансированное множество вопросов по каждой из выделенных групп, можно ожидать, что будут получены характеристики, показывающие способность LLM отвечать на вопросы в рамках предлагаемой классификации (см. табл.1).

Таблица 1. Примеры вопросов, ответов и разметки по категориям вопросов.

Вопрос	**Ответ**	**Тематика**	**Форма ответа**	**Сложность**	**Вопросительное слово**
Что будет, если объект 1 категории не получит комплексное экологическое разрешение?	Расчет платы за негативное воздействие на окружающую среду производится с повышающим коэффициентом 100; на природопользователя налагается штраф от 50 000 до 100 000 рублей в соответствии со статьей 8.47 КоАП РФ.	Нормативно-правовой	Текстовый	2	Что?
Должен ли импортер шин 2024 года подавать отчетность в Росприроднадзор? Если да, то какую?	Отчетность о массе товара, отчет о выполнении нормативов утилизации (при наличии), расчет суммы эко-сбора (при необходимости). Также, в случае отсутствия 100 процентов утилизации, необходимо произвести плату эко-сбора.	Нормативно-правовой	Текстовый	2	Требуется ли?
У меня свиноводческий комплекс на 5000 голов. Производится мясная продукция. Должен ли я вставать на учет в качестве объекта негативного воздействия на окружающую среду?	Да, согласно Постановлению Правительства РФ № 2398, объект должен быть включен в государственный реестр объектов, оказывающих негативное воздействие на окружающую среду в случае разведения свиней, убоя и производства мяса и мясной продукции.	Нормативно-правовой	Бинарный	2	Требуется ли?

Кто отвечает за упаковку в контексте утилизации? Производитель упаковки или производитель товара?	С 2024 года за утилизацию упаковки отвечают производитель товара, производитель упаковки или импортеры товаров, в том числе товаров в упаковке.	Нормативно-правовой	Текстовый	2	Кто?
…	…	…	…	…	…

Подготовка вопросов включает в себя и подготовку эталонных ответов. При этом для использования алгоритмического подхода в оценке качества LLM эталонные ответы должны быть развернутыми. Учитывая трудоёмкость процесса подготовки вопросов и эталонных ответов количество вопросов, которое будет подготовлено, окажется небольшим. Во многих областях знания такие выборки могут быть подготовлены на основе множеств часто задаваемых вопросов. Ответы при этом окажутся представлены лишь одним ответом на каждый вопрос, что делает процесс составления рейтингов без привлечения экспертов еще более сложным, так как правильные ответы могут отличаться по форме их выражения. Например, при ответе на вопрос «Можно ли что-то сделать?» могут быть получены на первый взгляд противоположные ответы, которые по сути будут говорить об одном и том же: «Да, можно за исключением определенных случаев» и «Нет, нельзя за исключением перечня случаев». Сложная для анализа ситуация возникает и с вопросами с бинарными ответами (если они не содержат пояснений к ответу). Имея одну лексическую единицу в эталонном ответе, их невозможно использовать для анализа ответов LLM методами, использующими токенизацию (т.е. автоматически). Анализ таких ответов можно провести только методами экспертного анализа.

Выбор LLM для ранжирования

Область, связанная с LLM, развивается настолько стремительно, что с 2022 года, когда архитектура GPT была применена для работы с текстами, на её основе и с использованием других архитектур нейронных сетей созданы сотни моделей с самими разными характеристиками. Они отличаются используемой архитектурой, количеством параметров с которыми могут работать, набором технологий, который используется для подготовки данных для основной модели и декодирования получаемых результатов, размером входного и выходного наборов данных, данными, на основе которых обучалась модель.

Большинство моделей используют для обучения данные, находящиеся в свободном доступе в сети интернет. В этих данных ответы на интересующие нас вопросы (вопросы 1-го уровня сложности) могут быть найдены в явном виде. Тогда ответы, которые мы будем получать от LLM, должны быть не хуже тех, что могут быть найдены поисковыми системами с использованием ставших классическими алгоритмами поиска (например, PageRank алгоритм [7]). С другой стороны будет находиться абсолютный ИИ (AGI) [13], который на данный момент для нас недостижим, а значит нам необходима подготовка эталонных ответов на сформированное множество вопросов с использованием экспертов.

Для сравнения будем рассматривать уже обученные LLM, к которым имеется свободный доступ через сеть Интернет, не установлены ограничения (правовые/технические) для установки модели в закрытом контуре пользователя. При этом современные LLM могут дообучаться, а это значит, если компания хочет использовать свои данные для подготовки ответов на вопросы 3-го и 4-го уровней сложности, то модель должна быть доступна в свободном доступе (преимущественно open source модели) для скачивания и развертывания в локальной инфраструктуре компании для исключения утечки конфиденциальной информации.

Для того, чтобы сравнивать архитектуры и оценить влияние на результат данных, используемых для обучения, отберем модели с близкими характеристиками, а именно с числом токенов контекста от 20 тыс. и числом параметров моделей от 7 млрд. (см. табл. 2).

Таблица 2. Перечень LLM и их характеристик для сравнения по степени их готовности для решения задач в области ESG

Система	**Разработчик**	**Архитектура нейронной сети**	**Ссылка на обученную модель / модель для скачивания и использования**	**Число параметров модели/длина сообщения (контекст)**	**Данные для обучения**
DeepSeek	DeepSeek (Китай)	Mixture of Experts (MoE)	https://www.deepseek.com / https://github.com/deepseek-ai	671 млрд. параметров / до 128 тыс. токенов	14,8 трлн. токенов
Falcon	Институт технологических инноваций (TII) (ОАЭ)	Основана на GPT-3	https://falconllm.tii.ae / https://huggingface.co/blog/falcon	180 млрд. параметров / до 32 тыс. токенов	5,5 трлн. токенов с использованием набора данных RefinedWeb [8]
GigaChat	Сбер (Россия)	MoE. За текст отвечают модели ruGPT-3 и FRED-TP, за оценку семантической близости ruCLIP. Ансамбль назвали NeONKA (NEural Omnimodal Network with Knowledge-Awareness).	https://giga.chat / https://github.com/ai-forever/gigachat (библиотека для взаимодействия с API) https://huggingface.co/ai-sage/GigaChat-20B-A3B-base (Версия Lite)	29 млрд. параметров (во время инференса задействовано только 3 млрд.) / до 131 тыс. токенов	Книги, новости и статьи на русском и английском языках
Grok AI	xAI (США)	Mixture of Experts (MoE) (8 экспертов, 2 активных)	https://x.ai / https://github.com/xai-org/grok-1	314 млрд. параметров (активных параметров 86 млрд.) / до 128 тыс. токенов	Данные из социальной сети X (Twitter)
Mistral	Private (Франция)	Трансформер использующий Mixture of Experts (MoE)	https://chat.mistral.ai/chat / https://github.com/mistralai	7 млрд. параметров / до 128 тыс. токенов	н/д
Qwen	Alibaba (Китай)	Qwen (модель основана на Llama от Meta) Transformer, имеет такие компоненты как активация SwiGLU, внимание QKV bias	https://chat.qwenlm.ai / https://github.com/QwenLM/Qwen	32 млрд. параметров / до 128 тыс. токенов	18 трлн. токенов
Yandex GPT	Компания Яндекс (Россия)	Архитектура похожа на Llama или Qwen	https://yandex.cloud/ru/services/foundation-models / https://huggingface.co/yandex/YandexGPT-5-Lite-8B-pretrain	100 млрд. параметров / до 32 тыс. токенов	15 трлн. токенов

Методика оценки ответов

После сбора ответов LLM на вопросы из составленного множества необходимо, провести оценку качества ответа на каждый вопрос и получить агрегированные оценки для получения рейтинга. В собранном множестве вопросов в качестве эталонного эксперты подготовили только один ответ, который обладает краткостью и содержит при этом всю необходимую информацию (прежде всего необходимые цифры и ссылки на нормы законодательства).

При оценке ответов экспертными методами учитывают такие факторы как 1) точность информации, 2) глубина ответа, 3) контекстуальная релевантность, 4) актуальность, 5) практическая применимость, 6) ясность и структурированность, 7) этичность и нейтральность. Абсолютно полноценно и надежно оценить по данным критериям ответы способен только эксперт или группа экспертов.

Однако при наличии словарей терминов, учет этих факторов возможен и при использовании методов, использующих токенизацию.

Принимая в качестве эталонного ответа ответ эксперта будем считать, что он содержит всю необходимую информацию. Тогда для ранжирования ответов будем искать функцию $d(A_э, A)$ которая будет показывать меру эталонного ($A_э$) и полученного (A) ответов.

Для эталонного ответа проведем токенизацию, лемматизацию и выполняем очистку удаляя стоп слова, знаки препинания и сохраняя только уникальные токены. В результате получим множество токенов W состоящее из $|W|$ элементов, по которым будем анализировать ответы.

Вероятность появления токена $P(w)$ будем определять как вероятность Лапласа — $P(w) = \frac{n_w}{N}$, где n_w – число раз, которое встретился токен, N – число токенов в анализируемом тексте (тексте, составленном из всех ответов включая эталонный.

Аналогично посчитаем и вероятности для появления пары токенов $P(w_i, w_j)$, где $i \neq j, i \in [1, |W|], j \in [1, |W|]$.

Можно предположить, что с увеличением сложности вопроса разница между эталонным ответом и ответами LLM будет возрастать. При этом само значение вероятности наступления события, при котором ответ идеален, будет уменьшаться для сложных вопросов (то есть значения когерентности для сложных вопросов должны убывать). Задействуем две формулы для расчета когерентности, которые обоснованы и протестированы в работе [10] и ведут себя немного по разному в зависимости от объема текста ответа (в первой формуле значение когерентности C_{UCI} понижается при увеличении объема текста, а во второй формуле C_{UMass} — нет). Далее мы проведем расчеты и определим, какой из способов расчета в большей степени приближается к экспертной оценке качества ответа.

1) $C_{UCI}(T) = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j)$, где $PMI(w_i, w_j) = \log_2 \frac{P(w_i, w_j) + \varepsilon}{P(w_i) \cdot P(w_j)}$, N – общее количество лемм, w – токен, T – оцениваемый ответ (текст).

2) $C_{UMass}(T) = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log_2 \frac{P(w_i, w_j) + \varepsilon}{P(w_j)}$.

Рейтинг качества ответов получим по величине разности значений когерентности для эталонного и оцениваемого ответов.

Алгоритм оценки описывается следующими шагами:

Шаг 1. Для эталонного ответа проводим токенизацию, лемматизацию и выполняем очистку удаляя стоп слова, знаки препинания и сохраняя только уникальные токены (получаемый эталонное множество токенов W).

Шаг 2. Для всех ответов LLM на анализируемый вопрос проводим токенизацию, лемматизацию и выполняем очистку удаляя стоп слова, знаки препинания.

Шаг 3. Для каждого токена (w_i) и уникальных пар токенов (w_i, w_j) определяем вероятности $p(w_i)$ и $p(w_i, w_j)$ $\forall i, j$.

Шаг 4. Используя значения вероятностей для каждого токена $w_i \in W$ и их сочетаний из эталонного ответа, определяем значение когерентности каждого ответа LLM на анализируемый вопрос.

В качестве еще одной меры возьмем косинусное расстояние, которое будем считать попарно для эталонного ответа и ответа LLM, используя для этого вектор Bag of Words для рассматриваемой пары.

Результаты, ранжирования моделей по качеству ответов

Полученные результаты можно разделить на качественные и количественные. Качественные результаты были получены в результате экспертного анализа ответов (см. табл. 3), а количественные при использовании формальных методов.

Таблица 3. Перечень LLM и качественная оценка их характеристик.

	DeepSeek	**GigaChat**	**Grok AI**	**Mistral**	**Qwen**	**Yandex GPT**
Адаптивность	—	При последовательном задании вопросов из одной сферы делает ответы более короткими и конкретными.	Запоминает ранее заданные вопросы и последующие ответы строит с примерами опыта ответа на ранее заданные вопросы.	—	—	—
Лингвистические проблемы в ответах	—	—	—	Да	—	—
Достаточность знаний о предметной области	+	+	+	—	+	+

Для анализа результатов были собраны ответы на 81 вопрос, относящихся к разным тематикам, с разными формами вопросов, сложностью и отвечающие на разные вопросительные слова (см. пример в табл. 4).

Ответы, которые дают LLM - более развернутые, чем ответы эксперта-человека. Для оценки работы алгоритмических методов ранжирования необходимо провести их проверку с использованием экспертной разметки. Ответы эксперта могут не совпадать на 1-2 позиции с оценками, полученными с использованием алгоритмов. Это связано с наличием порога безразличия, который можно продемонстрировать так называемым парадоксом чашек кофе или парадоксом Люче [6].

Суть парадокса состоит в том, что если у нас, например, есть 402 чашки кофе: c_0 чашка из нуля зерен кофе, c_1 чашка из одного зерна кофе и т.д. То мы не можем сказать, что нет разницы между c_0 и c_{401}. Однако разница между соседними элементами последовательности может оказаться для нас незаметна. Этот пример показывает, что при оценке LLM может нарушаться свойство транзитивности, что и приводит к необходимости не обращать внимание на незначительные нарушения порядка, а задача сводиться к определению некоторого квази-порядка, который характеризуется:

- интервальными порядками (пороги предпочтений и безразличия);
- псевдо-порядком (имеются несравнимые альтернативы);
- частичной упорядоченностью (ограниченная структура и множество предпочтений).

Сравнивая полученные с применением оценки когерентности результаты и результаты экспертной оценки мы наблюдаем нарушение порядка при сохранении лидеров и аутсайдеров (см. табл. 5 и 6).

Таблица 4. Примеры сокращенных ответов LLM (вопросы и эталонные ответы были подготовлены ФГБУ Научно-исследовательский институт промышленной экологии).

Вопрос	Ответ эксперта	DeepSeek	GigaChat	Grok AI	Mistral	Qwen	Yandex GPT
Что будет если объект 1 категории не получит комплексное экологическое разрешение?	Расчет платы за негативное воздействие на окружающую среду производится с повышающим коэффициентом 100; на природопользователя налагается штраф от 50 000 до 100 000 рублей в соответствии со статьей 8.47 КоАП РФ.	Если объект I категории не получит комплексное экологическое разрешение (КЭР), это может повлечь за собой ряд серьезных последствий: Запрет на эксплуатацию: Объекты I категории, оказывающие значительное негативное воздействие на окружающую среду, обязаны получить КЭР. …	Объект I категории — это объекты, оказывающие значительное негативное воздействие на окружающую среду. Если такой объект не получит комплексное экологическое разрешение (КЭР), возможны следующие последствия: …	Если объект I категории, который обязан получить комплексное экологическое разрешение (КЭР) в соответствии с законодательством Российской Федерации, не сделает этого, это повлечет за собой ряд юридических и финансовых последствий. Вот что может произойти: …	Если объект 1 категории не получит комплексное экологическое разрешение (КЭР), это может привести к ряду серьезных последствий: Административная ответственность: Отсутствие КЭР может привести к административным штрафам и санкциям. …	Если объект 1 категории (объект высокого экологического риска) не получит комплексное экологическое разрешение (КЭР), это повлечет ряд правовых, экономических и организационных последствий. Рассмотрим их подробнее: 1. Прекращение деятельности …	В соответствии с Федеральным законом от 10.01.2002 № 7-ФЗ «Об охране окружающей среды» объекты I категории подлежат федеральному государственному экологическому контролю. Для осуществления деятельности на таких объектах необходимо получить комплексное экологическое разрешение (КЭР). …
Должен ли импортер шин 2024 года подавать отчетность в Росприроднадзор? Если да, то какую?	Отчетность о массе товара, отчет о выполнении нормативов утилизации (при наличии), расчет суммы экосбора (при необходимости). Также в случае невыполнения нормативов утилизации необходимо произвести плату экосбора.	Да, импортер шин в 2024 году обязан подавать отчетность в Росприроднадзор, если шины подлежат утилизации после утраты ими потребительских свойств. Это связано с требованиями расширенной ответственности производителей и импортеров (РОП) в рамках экологического законодательства России. …	Да, импортёры шин обязаны подавать отчётность в Росприроднадзор. Согласно действующему законодательству Российской Федерации, компании, занимающиеся ввозом товаров, подлежащих утилизации после окончания срока службы, включая шины, должны соблюдать требования по учёту и отчётности в области обращения с отходами. …	Да, импортер шин в 2024 году обязан подавать отчетность в Росприроднадзор в рамках механизма расширенной ответственности производителей и импортеров (РОП), установленного Федеральным законом № 89-ФЗ "Об отходах производства и потребления". …	Да, импортер шин в 2024 году обязан подавать отчетность в Росприроднадзор. В частности, импортеры должны представлять следующие виды отчетности: …	Согласно российскому законодательству, начиная с 2021 года, введена система производственного экологического контроля (ПЭК) для операторов, занимающихся оборотом товаров, подлежащих утилизации. …	Импортёры шин 2024 года должны соблюдать требования законодательства Российской Федерации в области охраны окружающей среды. В соответствии с Федеральным законом от 10.01.2002 № 7-ФЗ «Об охране окружающей среды», импортёры могут быть отнесены к объектам, подлежащим федеральному государственному экологическому контролю. …
…	…	…	…	…	…	…	…

Таблица 5. Результаты ранжирования LLM при получении ответа на вопрос №8 «Как Россия участвует в Парижском соглашении?»

LLM	C_{uci}	C_{UMass}	Косинусное расстояние	C_{uci} , рейтинг	C_{UMass}, рейтинг	Косинусное расстояние, рейтинг	Экспертное ранжирование
DeepSeek	3.3928	-0.2574	0.0426	2	3	4	2
GigaChat	1.9328	-0.1840	0.0851	4	5	3	4
Grok AI	4.5932	-0.4196	0.0261	1	1	5	1
Mistral	1.6017	-0.1741	0.3284	5	6	2	5
Qwen	2.3781	-0.2753	0.0222	3	2	6	3
Yandex GPT	1.5706	-0.2409	0.7180	6	4	1	6

Таблица 6. Результаты ранжирования LLM при получении ответа на вопрос №10 «Нужно ли выбрасывать батарейки в отдельный контейнер?»

LLM	C_{uci}	C_{UMass}	Косинусное расстояние	C_{uci}, рейтинг	C_{UMass}, рейтинг	Косинусное расстояние, рейтинг	Экспертное ранжирование
DeepSeek	6.1128	-0.7488	0.0247	1	1	5	3
GigaChat	4.2593	-0.5709	0.2448	2	2	3	2
Grok AI	6.1128	-0.7488	0.0765	1	1	4	1
Mistral	6.1128	-0.7488	0.7058	1	1	1	6
Qwen	6.1128	-0.7488	0.0217	1	1	6	4
Yandex GPT	2.7535	-0.4205	0.6946	3	3	2	5

Из полученных результатов видно, что разница между когерентностью C_{uci} и экспертными оценками меньше, чем между другими рассмотренными способами автоматического получения рейтинга (косинусное расстояние, например, дает практически противоположный результат). Для построения рейтинга проведем оценку по всем собранным вопросам и построим скрипичную диаграмму (см. рис. 1). На вертикальной оси будем откладывать значение когерентности C_{uci} (чем выше значение, тем выше качество ответа). Белая точка показывает медианное значение когерентности. Ширина фигуры в каждом месте показывает частоту ответов данной LLM с данным значением когерентности. Как видим, на этом графике наилучший результат показывает модель Grok AI.

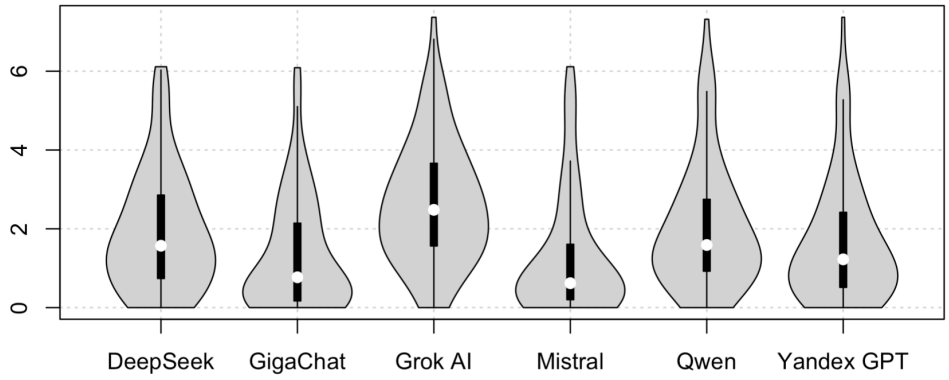
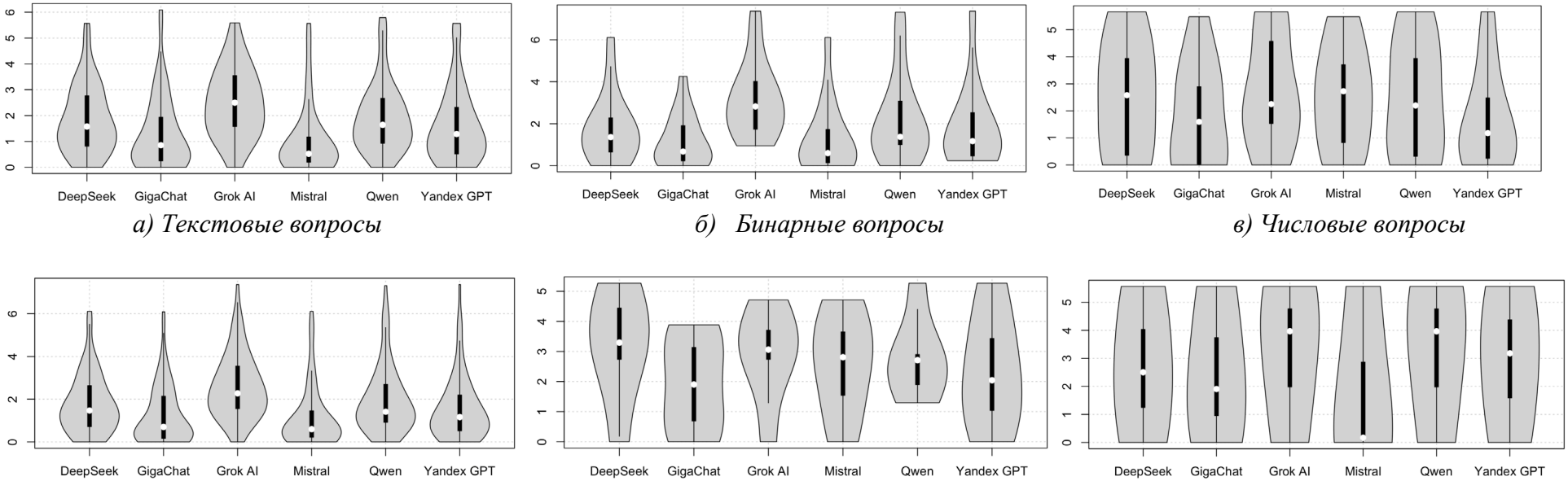


Рис. 1. Рейтинг LLM по всем вопросам в виде скрипичных диаграмм, полученных на основе оценки когерентности C_{uci}.

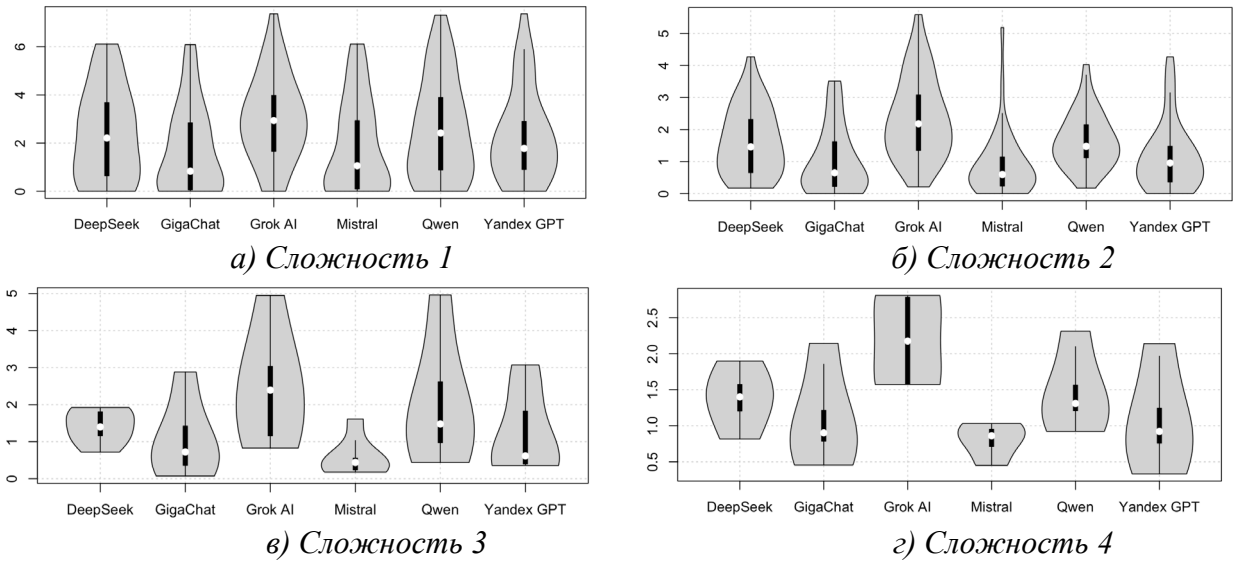
Построив рейтинги по отдельным категориям вопросов, получим результаты, приведенные на рис. 3-4.



а) Текстовые вопросы *б) Бинарные вопросы* *в) Числовые вопросы*

г) Нормативно-правовые вопросы *д) Гносеологические вопросы* *е) Терминологические вопросы*

Рис. 2. Рейтинг LLM в виде скрипичных диаграмм, полученных на основе оценки когерентности C_{uci}: а) по текстовым вопросам, б) по бинарным вопросам, в) по числовым вопросам, г) по нормативно-правовым вопросам, д) по гносеологическим вопросам, е) по терминологическим вопросам



а) Сложность 1 *б) Сложность 2*

в) Сложность 3 *г) Сложность 4*

Рис. 3. Рейтинг LLM в виде скрипичных диаграмм, полученных на основе оценки когерентности C_{uci}: а) по вопросам 1-й сложности, б) по вопросам 2-й сложности, в) по вопросам 3-й сложности, г) по вопросам 4-й сложности

В результате построения рейтингов по отдельным категориям видно, что LLM ведут себя по-разному. Для построения рейтингов по вопросам «Почему?» и «Куда?» оказалось недостаточно данных. Можно заметить, что Grok AI, являясь показывая наилучшие результаты по общей массе вопросов без различения категорий, показывает снижение результатов до нулевых значений по таким категориям как бинарные вопросы, вопросы «Когда?», «Требуется ли?» и «Кто?», а также по вопросам 1-й и 4-й сложности, что может говорить о том, что эта LLM по этим категориям вопросов не будет давать недостоверных ответов. При этом все LLM показывают примерно равные результаты по терминологическим вопросам.

В процессе проведения экспериментов эталонные ответы уточнялись т.к. за время обсуждения произошли незначительные изменения законодательства, а также эксперты, которые готовили ответы, согласовывали свое мнение по формулировкам эталонных ответов. Это привело к необходимости пересчета рейтинга, что в свою очередь повлияло на получаемые значения и вид распределений на скрипичных диаграммах. Однако не оказало влияние на оценки, которые получили LLM как по всем вопросам, так и различным видам врезов по вопросам. Такая ситуация говорит с одной стороны о чувствительности метода к изменениям и отсутствия эффекта переобучения (в данном случае избыточности данных), с другой стороны из-за сохранения мест в рейтинге говорит о достаточности данных для составления рейтингов LLM.

Повторение эксперимента с выбором в качестве меры косинусного расстояния показало нечувствительность метода к изменениям, что еще раз подтверждает о том, что данная мера не может быть применена к рассматриваемой задаче.

Таким образом, еще одним результатом, полученным в результате проведенного эксперимента, является то, что не все меры близости применимы для оценки качества работы LLM по эталонным ответам.

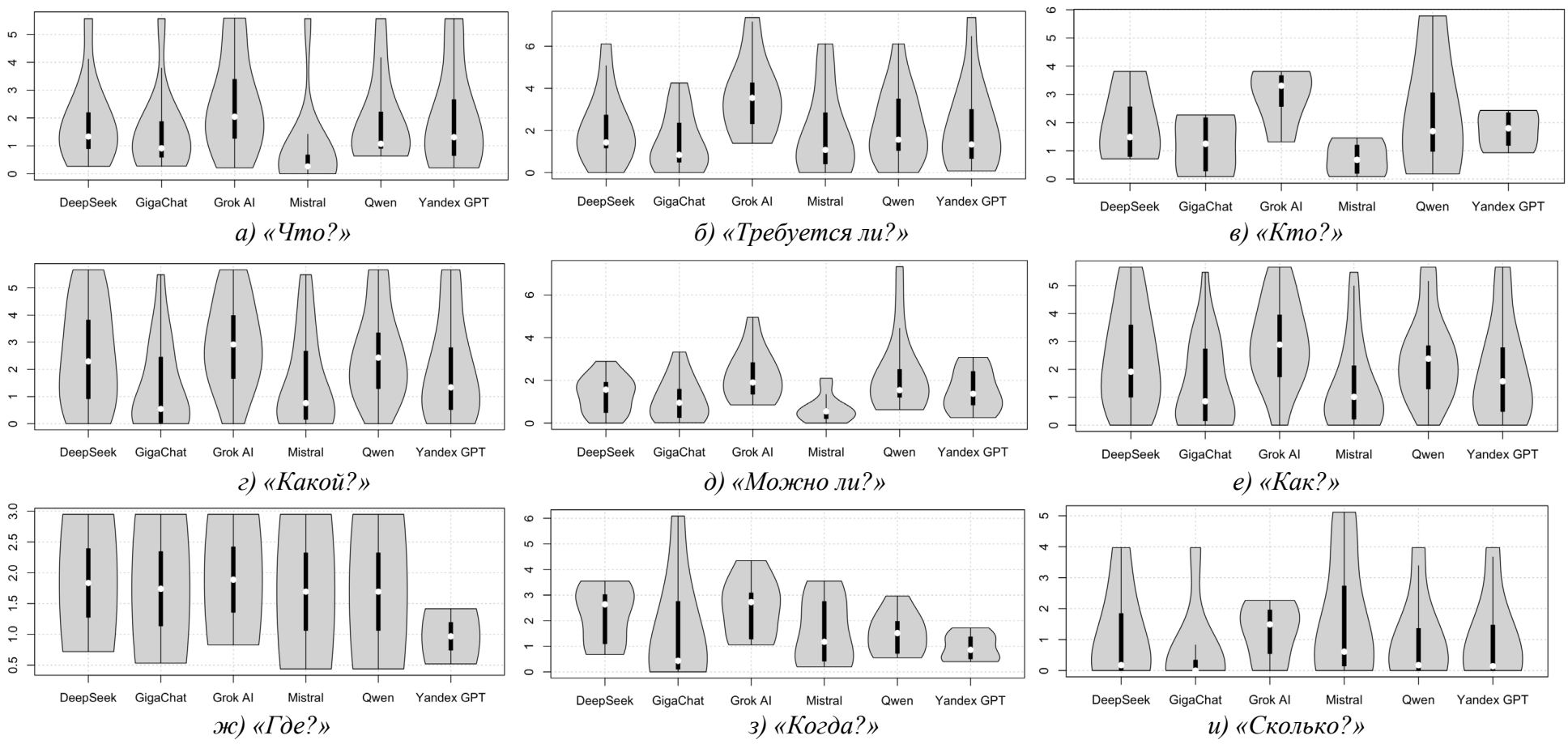


Рис. 4. Рейтинг LLM в виде скрипичных диаграмм, полученных на основе оценки когерентности C_{uci}: а) по вопросам, отвечающим на вопрос «Что?», б) по вопросам, отвечающим на вопрос «Требуется ли?», в) по вопросам, отвечающим на вопрос «Кто?», г) по вопросам, отвечающим на вопрос «Какой?», д) по вопросам, отвечающим на вопрос «Можно ли?», е) по вопросам, отвечающим на вопрос «Как?», ж) по вопросам, отвечающим на вопрос «Где?», з) по вопросам, отвечающим на вопрос «Когда?», и) по вопросам, отвечающим на вопрос «Сколько?».

Теперь мы можем построить обобщающий рейтинг LLM на основе среднего значения C_{uci} и с использованием всех данных. Для оценки устойчивости данных результатов посчитаем и другие параметры описательной статистики (см. табл. 7).

Таблица7. Рейтинг LLM и значения параметров описательной статистики для C_{uci}.

Рейтинг	LLM	Среднее значение	Значение медианы	Дисперсия	Стандартное отклонение	Мин. значение	Макс. значение
1	Grok AI	2.636218	2.480794	2.384537	1.544195	0	7.367779
2	Qwen	2.054098	1.588951	2.58365	1.607374	0	7.316624
3	DeepSeek	1.912428	1.570291	2.266501	1.50549	0	6.112856
4	Yandex GPT	1.660196	1.224961	2.483691	1.575973	0	7.367779
5	GigaChat	1.331862	0.770178	2.099291	1.448893	0	6.088549
6	Mistral	1.265062	0.619849	2.358404	1.535709	0	6.112856

Как видно, лидирующее место занимает Grok AI, на втором и третьем месте находятся очень близкие по значению Qwen и Deepseek. Из данной таблицы также видно, лидеры рейтинга могут иногда ошибаться, давая неверные ответы (не согласующиеся с ответами эксперта), что необходимо учитывать и закладывать эти риски при выборе системы.

Внимательный анализ ответов LLM также показывает минимальное использование маркеров неопределенности и субъективной оценки информации, что свидетельствует о трудностях моделирования рефлексивных аспектов познания. Наблюдается низкое лексическое разнообразие дискурсивных маркеров, высокая повторяемость контекстуальных шаблонов. Отсутствуют эмоциональные слова, аддитивные и эпистемические маркеры, маркеры переформулирования, наблюдается переизбыток логических связей, что в живой речи людей говорит об отсутствии целостного понимания предметной области. Ответы LLM сформулированы таким образом, что не дают повода усомниться в знаниях LLM и сами LLM не сообщают о своих ограниченных знаниях.

Из проделанного анализа следует, что неспециалистам использовать LLM все-таки нужно с должной осторожностью, поскольку ответы модели могут быть неполны или неточны. Проблема усложняется тем, что в ответе, как правило, отсутствует информация об использованных источниках (напр., первоисточников в виде нормативных актов или из уже подготовленных экспертами или другими LLM ответов) и о способе составления ответа (напр., путем рассуждений на основе ограниченных сведений или источников, содержащих ответ; были ли в этих источниках противоречия и как они были разрешены и т. п.). Без указания данной информации оценка достоверности становится трудоемкой. Уже известно такое явление как деградация LLM [11], которое наблюдается, когда их ответы попадают в сеть и используются для дальнейшего обучения. В перспективе такая проблема будет усиливаться, поэтому указание источников и способа составления ответа LLM может иметь принципиальное значение.

Заключение

В результате проведенной работы была получена группа рейтингов LLM при ответе на вопросы связанные с охраной природы. Особенностью полученных рейтингов является не только их ранжирование, по интегральной оценке, но и оценка возможных отклонений в лучшую или худшую стороны, а также оценка функции распределения ответов и наличие «хвостов» при оценке качества ответов.

Таким образом рейтинг учитывает не только некоторое место, но и стабильность ответов и величину ошибки, которая может быть допущена системой при подготовке ответа. Такое построение рейтинга позволяет подбирать систему исходя из выбранной стратегии поведения компании.

Предложенный в статье подход может быть масштабирован на другие области знания и сферы деятельности, что позволит находить эффективные способы применения LLM [12]

в практической деятельности, увеличивая производительность труда, а не попадая в ситуации, связанные с таким явлением как Парадокс 2.0.

Благодарности

Авторы благодарят сотрудников ФГБУ «Научно-исследовательский институт промышленной экологии», подведомственное учреждение Росприроднадзора за экспертную поддержку, оказанную связанную с подготовкой вопросов, ответов и многочисленные дискуссии, связанные с выбором множества рейтингуемых LLM и выбором способа их оценивания исходя из специфики выбранной для исследования области. Авторы благодарны Алексею Масютину (Центр искусственного интеллекта НИУ ВШЭ) за координацию данного исследования. Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ в 2025 году.

Список литературы

1. Bureš L. [и др.]. Semantic text segmentation from synthetic images of full-text documents // SPIIRAS Proceedings. 2019. № 6 (18). С. 1380–1405.
2. Chauhan S., Daniel P. A Comprehensive Survey on Various Fully Automatic Machine Translation Evaluation Metrics // Neural Processing Letters. 2023. № 9 (55). С. 12663–12717.
3. Hu T., Zhou X.-H. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions 2024.
4. Jackson P. Introduction to Expert System / P. Jackson, NY: Addison-Wesley, 1998. 560 с.
5. Juba B., Le H. S. Precision-Recall versus Accuracy and the Role of Large Data Sets // Proceedings of the AAAI Conference on Artificial Intelligence. 2019. № 01 (33). С. 4039–4048.
6. Luce R. D. The choice axiom after twenty years // Journal of Mathematical Psychology. 1977. № 3 (15). С. 215–233.
7. Page L. [и др.]. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report / L. Page, S. Brin, R. Motwani, T. Winograd, Stanford InfoLab,.
8. Penedo G. [и др.]. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only 2023.
9. Perevalov A. [и др.]. Knowledge Graph Question Answering Leaderboard: A Community Resource to Prevent a Replication Crisis 2022.
10. Röder M., Both A., Hinneburg A. Exploring the Space of Topic Coherence Measures New York, NY, USA: ACM, 2015.С. 399–408.
11. Shumailov I. [и др.]. Author Correction: AI models collapse when trained on recursively generated data // Nature. 2025. № 8058 (640). С. E6–E6.
12. Strassmann P. A. The Economics of Corporate Information Systems: Measuring Information Payoffs / P. A. Strassmann, Information Economics Press, 2007. 224 с.
13. Tegmark M. Life 3.0: Being Human in the Age of Artificial Intelligence / M. Tegmark, Knopf, 2017. 384 с.
14. Новикова А. В. Референциально-ситуативный анализ семантики возможных миров // Вестник Челябинского государственного университета. 2008. № 26. С. 101–107.