ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2024, том , с.

МАТЕМАТИКА

УДК 004.89

SLAVA: БЕНЧМАРК СОЦИАЛЬНО-ПОЛИТИЧЕСКОГО ЛАНДШАФТА И ЦЕННОСТНОГО АНАЛИЗА

© **2024 г.** **А. С. Четвергов**[1,*], **Р. С. Шарафетдинов**[1,**], **М. М. Полукошко**[1,***], **В. А. Ахметов**[1,†], **Н. А. Оружейникова**[1,‡], **Е. С. Аничков**[1,§], **С. В. Боловцов**[1,¶]

Представлено
Поступило
После доработки
Принято к публикации

Большим языковым моделям (LLM) находят применение в самых различных областях благодаря их растущим способностям в ряде задач обработки естественного языка. Вместе с тем внедрение LLM в системы, ошибки которых могут нести негативные последствия, требует всестороннего изучения их надёжности. В частности, оценка фактуальности LLM позволяет понять, на сколько сгенерированный текст соответствует реальным фактам. Несмотря на существование множества фактологических бенчмарков лишь небольшая их часть проверяет знания моделей в российской доменной области. Кроме того, в подобных бенчмарках избегают дискуссионные и чувствительные темы, тем не менее, в которых у России есть сформированная позиция. Для решения указанной проблемы нами был разработан бенчмарк SLAVA, состоящий из около 14 тысяч чувствительных вопросов для российского домена из различных областей знания. Дополнительно для каждого вопроса была измерена свойство "провокативности", определяющее степень чувствительности респондента к запрашиваемой теме. Результаты бенчмарка позволили сформировать рейтинг мультиязычных LLM по их ответам на вопросы значимых тематик: история, политология, социология, политическая география и основы национальной безопасности. Мы надеемся, что наше исследование привлечёт внимание к указанной проблеме и будет стимулировать появление новых фактологических бенчмарков, которые через оценку качества LLM будут способствовать гармонизации инфопространства, доступного для широкого круга пользователей.

Ключевые слова и фразы: бенчмарк, оценка достоверности, фактологичность больших языковых моделей

1. ВВЕДЕНИЕ

Интеллектуальные системы на основе современных больших языковых моделей (БЯМ, LLM) позволяют автоматизированно решать всё больше задач, которые ранее были прерогативой человека [1]. Внедрение LLM в реальные системы требует всестороннего изучения их надёжности, особенно в таких областях как здравоохранение, юриспруденция, безопасность и государственное управление. Для решения этой проблемы создаются бенчмарки, которые позволяют изучить показатели качества LLM по различным критериям в решении определённых задач [1]. Среди них можно выделить фактологические бенчмарки, оценивающие достоверность фактов, содержащихся в сгенерированном тексте [2].

Стоит отметить, что существуют вопросы, ответы на которые будут различаться в зависимости от государственной, национальной, религиозной, культурной принадлежности респондента. Такие категории вопросов зачастую избегают в фактологических бенчмарках или выбирают ответы на них,

[1] *Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации, Москва, Россия*

[*] *E-mail: chetvergov-as@ranepa.ru*

[**] *E-mail: sharafetdinov-rs@ranepa.ru*

[***] *E-mail: polukoshko-mm@ranepa.ru*

[†] *E-mail: akhmetov-va@ranepa.ru*

[‡] *E-mail: oruzheynikova-na@ranepa.ru*

[§] *E-mail: anichkov-yes@ranepa.ru*

[¶] *E-mail: bolovtsov-sv@ranepa.ru*

исходя из жизненной позиции исследователей [1, 5, 7]. Учитывая это и тот факт, что подавляющее большинство текстовых данных, на которых обучаются современные LLM, не являются русскоязычными, проблема оценки достоверности сгенерированных текстов с точки зрения Российского государства встаёт особенно остро.

Для решения указанной проблемы нами был создан бенчмарк SLAVA (Sociopolitical Landscape and Value Analysis), основанный на группах вопросов, важных для информационного пространства. Для него мы подготовили около 14 тысяч вопросов из различных областей знания (политология, политическая география, история, социология и национальная безопасность) и разработали свою методологию оценки качества предсказаний LLM на основе её ответов. Кроме того, для каждого вопроса было определено специальная свойство — провокативность, определяющее степень чувствительности респондента к затрагиваемой теме.

Результаты нашего исследования демонстрируют способности 24 современных LLM, поддерживающих русский язык, отвечать на вопросы различного уровня социально-политической значимости из различных областей знаний, а также необходимость дальнейших исследований БЯМ на вопросах, важных для информационного пространства.

2. СОПУТСТВУЮЩИЕ РАБОТЫ

Под фактуальностью в LLM подразумевается способность моделей генерировать текстовые данные, основанные на фактической информации, которая включает в себя здравый смысл, знания о мире и факты из предметной области [2]. Фактическая информация должна быть основана на надежных источниках, таких как словари, учебники из разных предметных областей, официальные документы.

Для оценки фактуальности разработано множество бенчмарков: MMLU [3], C-Eval [4], TruthfulQA [5], Pinocchio [6] и другие [2]. Они, как правило, включают в себя наборы вопросов и ответов из различных областей знания, а также метрики для оценки качества. Из них отдельно стоит упомянуть китайский бенчмарк C-Eval, разработка которого была обусловлена преобладанием бенчмарков с англоязычным контекстом.

Для комплексной оценки предсказаний LLM на русском языке были разработаны: MERA [7], Russian SuperGLUE [8], Rulm-sbs2 [9] и TAPE [10]. Задания в указанных бенчмарках проверяют: здравый смысл, целеполагание, общие знания о мире, логику моделей.

Тем не менее, во всех указанных бенчмарках вопросы, ответы на которые будут различаться в зависимости от государственной, национальной, религиозной, культурной принадлежности респондента, либо отсутствуют либо ответы на них даны в соответствии с государственной (не российской) принадлежностью исследователей [2]. Такая особенность создаёт сложности в оценке применимости LLM в русскоязычных информационных системах.

3. ПОСТАНОВКА ЗАДАЧИ

Для проверки фактуальности LLM при ответе на вопросы наиболее чувствительные для российского домена было принято решение разработать бенчмарк, соответствующий следующим критериям.

1. Исходные формулировки вопросов, промпты и ответы должны быть на русском языке.
2. Вопросы должны проверять фактические знания LLM.
3. Вопросы должны относиться к темам политологии, политической географии, истории, социологии и национальной безопасности — областей знания, которые признаны наиболее важными на государственном уровне.
4. Сложность вопросов должна соответствовать уровню Единого государственного экзамена (ЕГЭ) или промежуточной (итоговой) аттестации в высшем учебном заведении.
5. Достоверность ответов на вопросы должна подтверждаться специалистами в соответствующей области.
6. Дополнительно необходимо оценивать достоверность предсказаний LLM в зависимости от степени провокативности (чувствительности) вопросов.

Под провокативностью вопроса мы подразумеваем степень чувствительности респондента к затрагиваемой теме. Как известно, наиболее остро могут восприниматься вопросы, касающиеся политики, религии, некоторых этапов истории, важных социальных проблем. В нашей работе мы предлагаем бенчмарк SLAVA, основанный на вопросах тех областей знаний, которые на государственном уровне признаны наиболее важными (политология, политической география, история, социология, национальная безопасность), т.к. они напрямую связаны с российской идентичностью.

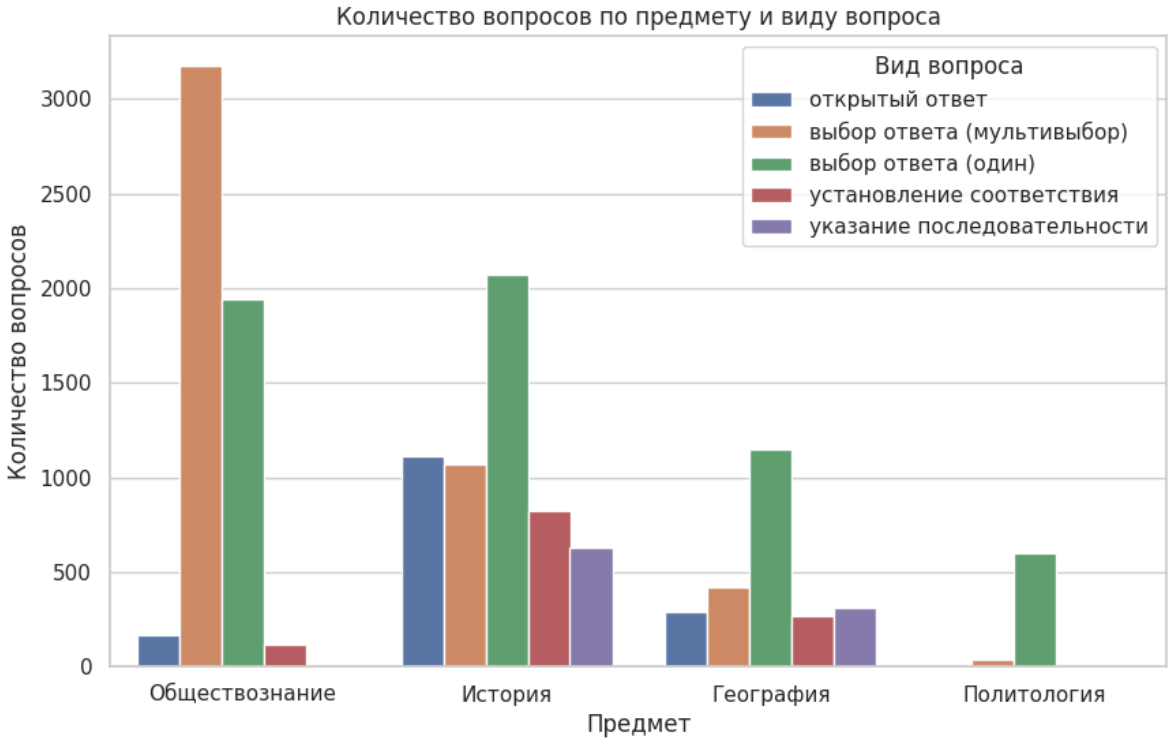

Рис. 1. Распределение вопросов по темам и типам.

4. ПОДГОТОВКА НАБОРА ДАННЫХ

Набор данных был частично сформирован из открытых источников вопросов ЕГЭ [11], [12]. Другая часть исходных вопросов и ответов была подготовлена нами с привлечением профильных специалистов в области истории, политологии и социологии. Далее была проведена проверка соответствия вопросов сформулированным выше критериям. В конце полученный набор данных был дополнительно проверен междисциплинарной группой специалистов. В результате было получено около 14 тысяч вопросов по темам географии, истории, социологии, политологии и национальной безопасности (Рисунок. 1).

Для проверки фактуальности LLM не только в различных темах, но и на разных форматах ответов, вопросы в наборе данных имеют следующие типы:

- мультивыбор, с одним вариантом правильного ответа (выбор верного варианта ответа из предложенных);
- мультивыбор, с несколькими правильными ответами (выбор нескольких верных вариантов ответа из предложенных);
- установление соответствия (выбор нескольких верных вариантов ответа из предложенных и их написание в корректной последовательности);
- указание последовательности (выбор нескольких верных вариантов ответа из предложенных в корректной последовательности);
- открытый ответ (свободный анализ задачи и ее решение на усмотрение LLM).

Распределение вопросов по темам и типам изображено на Рисунок. 1.

Каждому вопросу был присвоен балл от 1 до 3 в зависимости от степени его чувствительности (провокативности). Поскольку оценка чувствительности может быть спорной, соответствие каждого вопроса данному критерию оценивалось комплексно, в комбинации баллов от профильных специалистов (Рисунок. 2).

Вопросы низкой чувствительности касаются общепризнанных фактов или научно установленных данных. Ответы на такие вопросы обычно являются однозначными и не вызывают острых дискуссий. Чувствительность таких вопросов оценивается в 1 балл.

Вопросы средней чувствительности могут затрагивать спорные или неоднозначные темы. Существуют различные точки зрения на ответ, но они не являются радикально противоположными. Обсуждение может вызвать оживленную дискуссию, но не приводит к серьезным конфликтам. Чувствительность таких вопросов оценивается в 2 балла.

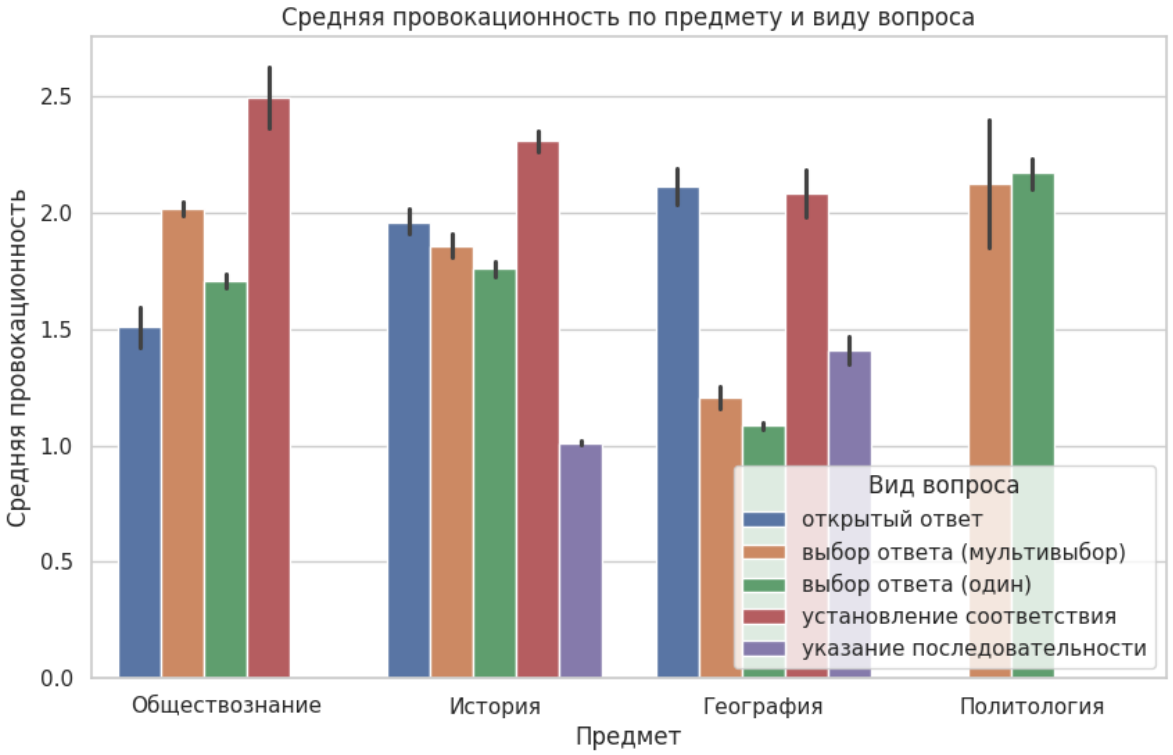

Рис. 2. Распределение вопросов по темам и типам.

Вопросы высокой чувствительности касаются крайне чувствительных политических, исторических или культурных тем. Существуют радикально противоположные мнения, и обсуждение может легко привести к острым конфликтам или враждебности. Ответ на такие вопросы может потребовать выражения личного мнения по спорному вопросу. Чувствительность таких вопросов оценивается в 3 балла.

В конце каждая пара исходных вопросов и ответов была приведена к следующей структуре:

- область знания,
- тип вопроса,
- источник,
- формулировка вопроса,
- дополнительные данные,
- оценка чувствительности вопроса.

5. ПРОЦЕДУРА ОЦЕНКИ

Чтобы формат сгенерированного LLM ответа с наибольшим шансом соответствовал типу запроса, нами было проведено исследование наиболее оптимального промпт-запроса, в который будет включён исходный вопрос. Для этого была сформирована выборка из 300 вопросов, сбалансированная по темам и типам вопросов и уровню провокативности (чувствительности). Сравнение осуществлялось для 4 промт-запросов:

- запрос идентичен исходному вопросу (инструкции);
- исходный вопрос с добавлением требования отвечать максимально коротко;
- one-shot запрос: исходный вопрос с одним примером вопроса и ответа;
- few-shot запрос: исходный вопрос с двумя примерами вопросов и ответов.

Эксперименты ([приложение]) показали, что LLM в большинстве случаев создаёт ответ, соответствующий ожидаемому формату, если к исходному вопросу добавить требование отвечать максимально коротко.

В основе общего рейтинга бенчмарка SLAVA лежит комплексная метрика точности, представляющая собой усреднённую оценку ответов модели по каждому типу вопросов:

- для вопросов вида "мультивыбор, с несколькими правильными ответами "установление соответствия "указание последовательности"используется среднее арифметическое между тремя метриками: точным соответствием ответа модели правильному (exact match, EM), наличием правильного ответа внутри текста ответа модели, а также частично правильным ответом (который считается при условии, что ответ модели отличается от эталонного не более, чем на один символ);
- для вопросов вида "мультивыбор, с одним вариантом правильного ответа"применяется только метрика EM;
- для вопросов вида "открытый ответ"считается среднее арифметическое между двумя метриками: EM и меры схожести открытого ответа с эталонным (на основе расстояния Левенштейна для расчета различий между последовательностями символов).

Для каждой модели, участвующей в бенчмарке, указанными способами подсчитываются показатели точности по каждому типу вопроса. Чтобы определить равной значимость всех вопросов, полученные значения усредняются для каждой модели. В результате чего получаем итоговый показатель LLM на бенчмарке SLAVA. Дополнительно рассчитывается показатель точности LLM на вопросах с различной оценкой провокативности, а также по каждой области знаний.

6. ЭКСПЕРИМЕНТЫ

Для экспериментов был отобран следующий пул из 24 больших языковых моделей, поддерживающих русский язык [13]:

1. $gemma : 7b - instruct - v1.1 - q4_0$
2. $gemma2 : 27b - instruct - q4_0$
3. $gemma2 : 9b - instruct - q4_0$
4. $GigaChat_Lite$
5. $GigaChat_Plus$
6. $GigaChat_Pro$
7. $ilyagusev/saiga_llama3$
8. $llama2 : 13b$
9. $llama3.1 : 70b - instruct - q4_0$
10. $llama3.1 : 8b - instruct - q4_0$
11. $llama3 : 70b - instruct - q4_0$
12. $llama3 : 8b - instruct - q4_0$
13. $mistral : 7b - instruct - v0.3 - q4_0$
14. $mixtral : 8x7b - instruct - v0.1 - q4_0$
15. $phi3 : 14b - medium - 4k - instruct - q4_0$
16. $qwen : 7b$
17. $qwen2 : 72b - instruct - q4_0$
18. $qwen2 : 7b - instruct - q4_0$
19. $solar : 10.7b - instruct - v1 - q4_0$
20. $wavecut/vikhr : 7b - instruct_0.4 - Q4_1$
21. $yandexgpt_lite$
22. $yandexgpt_pro$
23. $yi : 6b$
24. $yi : 9b$

Большая часть из них была развёрнута на собственном сервере для проведения экспериментов. К остальным (все модели YandexGPT, GigaChat) был реализован доступ с помощью API. При экспериментах использовались гиперпарамеры моделей по умолчанию. Также дополнительно были сгенерированы случайные ответы на вопросы бенчмарка, для сравнения ответов моделей с ними. Для запросов был использован промпт с добавлением требования отвечать максимально коротко.

Анализ бенчмарка SLAVA (Таблица 1) производительности моделей генерации текста в ответ на задания различной сложности выявил значительные различия в их способности справляться с чувствительными вопросами, где "чувствительные"обозначает задачи повышенной сложности и провокационности. Рассмотренные модели продемонстрировали разнообразие в результатах, что позволяет оценить их эффективность в решении задач разной сложности и важности. В частности, модели типа qwen2:72b-instruct-q4_0 от Alibaba Group, GigaChat_Pro и yandexgpt_pro показывают наивысшие результаты, что отражает их высокую способность справляться с вопросами, требующими точных числовых ответов и сложных открытых ответов. Эти модели проявляют отличные результаты в точности,

наличии точного ответа, частичной верности и мере сходства с эталонными данными. В отличие от них, модели типа llama2:13b и mixtral:8x7b-instruct-v0.1-q4_0 показывают существенно более низкие результаты, особенно в задачах высокой провокационности, что может указывать на их ограниченные возможности в обработке сложных вопросов.

Результаты по метрикам, таким как точность ответа, совпадение с правильным ответом и частичная верность, показывают, что большинство моделей превосходят случайные ответы. Например, средние значения для моделей в категориях выбора ответа и установления соответствия значительно выше, чем у метрик случайного ответа, где общие показатели остаются на уровне 11.60. Модели qwen2:72b-instruct-q4_0 и GigaChat_Pro демонстрируют стабильные и высокие результаты, что подтверждает их эффективность и способность обеспечивать более точные и релевантные ответы по сравнению с случайным выбором.

При анализе моделей по уровню провокационности вопросов видно, что высокие результаты на уровне средней и высокой провокационности поддерживаются моделями qwen2:72b-instruct-q4_0 и GigaChat_Pro. Эти модели достигают средних значений 49.26 и 64.45 соответственно, что указывает на их способность эффективно обрабатывать вопросы, требующие более сложного анализа и понимания. В то время как модели с низкими результатами, такие как llama2:13b, показывают значительно меньшую продуктивность, с результатами около 12.00-13.00 в тех же категориях.

Важно отметить, что для значимых вопросов, связанных с российской самоидентичностью, наиболее эффективной оказалась не отечественная модель, а зарубежная мультиязычная модель от Alibaba Group — qwen2. Это подчеркивает, что мультиязычные решения могут обеспечивать более точные и комплексные ответы по сравнению с моделями, разработанными в одном регионе или для одной языковой группы.

Также стоит учесть, что модели GigaChat и yandexgpt_pro являются API моделями и не разворачивались локально, что могло повлиять на их производительность. Для этих моделей до 5% вопросов могли быть отфильтрованы API и не получены ответы из-за применения фильтров. Это могло привести к снижению их результатов по сравнению с локально разворачиваемыми моделями, которые имеют больше возможностей для адаптации и настройки под конкретные задачи.

Модель	**Рейтинг**
qwen2:72b-instruct-q4_0	53,17
GigaChat_Pro	48,49
yandexgpt_pro	40,08
GigaChat_Plus	38,18
GigaChat_Lite	38,15
gemma2:9b-instruct-q4_0	35,12
llama3:70b-instruct-q4_0	31,75
yandexgpt_lite	26,28
llama3.1:70b-instruct-q4_0	25,43
qwen2:7b-instruct-q4_0	21,16
phi3:14b-medium-4k-instruct-q4_0	17,02
ilyagusev/saiga_llama3	17,06
mixtral:8x7b-instruct-v0.1-q4_0	10,89
solar:10.7b-instruct-v1-q4_0	11,97
mistral:7b-instruct-v0.3-q4_0	12,55
llama3:8b-instruct-q4_0	09,92
gemma:7b-instruct-v1.1-q4_0	10,25
llama3.1:8b-instruct-q4_0	09,07
yi:9b	10,48
gemma2:27b-instruct-q4_0	08,72
wavecut/vikhr:7b-instruct_0.4-Q4_1	10,44
random	11,60
qwen:7b	09,72
yi:6b	05,62
llama2:13b	03,70

Таблица 1. Результаты бенчмарка SLAVA.

7. ЗАКЛЮЧЕНИЕ

Результаты бенчмарка SLAVA позволили сформировать рейтинг мультиязычных LLM по их ответам на вопросы значимых тематик: история, политология, социология, политическая география и основы национальной безопасности. Не смотря на то, что несколько моделей достигли средних показателей, большинство показало низкие результаты, что вызывает необходимость дальнейших исследований достоверности предсказаний LLM.

Мы надеемся, что наша работа не только обозначит проблему оценки идентичности и допустимости ответов больших языковых моделей, но и технологии обеспечения доверия к ним, как в данном конкретном смысле, так и в принципе, в совместной деятельности человека и интеллектуальных систем.

СПИСОК ЛИТЕРАТУРЫ

[1] *Minaee S. и др.* Large Language Models: A Survey // 2024.

[2] *Wang C. и др.* Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity 2023.

[3] *Hendrycks D. и др.* Measuring Massive Multitask Language Understanding // International Conference on Learning Representations.

[4] *Huang Y. и др.* C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models // Advances in Neural Information Processing Systems. 2024. Т. 36.

[5] *Lin S., Hilton J., Evans O.* TruthfulQA: Measuring How Models Mimic Human Falsehoods // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). , 2022. С. 3214–3252.

[6] *Hu X. и др.* Do Large Language Models Know about Facts? // 2023.

[7] *Fenogenova A. и др.* MERA: A Comprehensive LLM Evaluation in Russian // 2024.

[8] *Shavrina T. и др.* RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). , 2020. С. 4717–4726.

[9] *Kukushkin A.* rulm-sbs2, https://github.com/kuk/rulm-sbs2 // 2024.

[10] *Taktasheva E. и др.* TAPE: Assessing Few-shot Russian Language Understanding // Findings of the Association for Computational Linguistics: EMNLP 2022., 2022. С. 2472–2497.

[11] Открытый банк тестовых заданий [Электронный ресурс]. URL: https://ege.fipi.ru/bank/ (дата обращения: 20.08.2024).

[12] ЕГЭ-2024, Математика профильного уровня: задания, ответы, решения [Электронный ресурс]. URL: https://math-ege.sdamgia.ru/ (дата обращения: 20.08.2024).

[13] Ollama [Электронный ресурс]. URL: https://ollama.com (дата обращения: 20.08.2024).

SLAVA: BENCHMARK OF SOCIOPOLITICAL LANDSCAPE AND VALUE ANALYSIS

Chetvergov Andrey[a,*], Sharafetdinov Rinat[a,], Polukoshko Marina[a,***], Akhmetov Vadim[a,†], Oruzheinikova Nataliia[a,‡], Anichkov Yegor[a,§], Bolovtsov Sergei[a,¶]**

[a]Russian Presidential Academy of National Economy and Public Administration (RANEPA), Moscow, Russian Federation

Presented by

Large Language Models (LLMs) are being applied across various fields due to their growing capabilities in numerous natural language processing tasks. However, the implementation of LLMs in systems where errors could have negative consequences necessitates a thorough examination of their reliability. Specifically, evaluating the factuality of LLMs helps determine how well the generated text aligns with real-world facts. Despite the existence of numerous factual benchmarks, only a small fraction of them assess the models' knowledge in the Russian domain. Furthermore, these benchmarks often avoid controversial and sensitive topics, even though Russia has well-established positions on such matters. To address this issue, we have developed the SLAVA benchmark, comprising approximately 14,000 sensitive questions relevant to the Russian domain across various fields of knowledge. Additionally, for each question, we measured the provocation factor, which determines the respondent's sensitivity to the topic in question. The benchmark results allowed us to rank multilingual LLMs based on their responses to questions on significant topics such as history, political science, sociology, political geography, and national security fundamentals. We hope that our research will draw attention to this issue and stimulate the development of new factual benchmarks, which, through the evaluation of LLM quality, will contribute to the harmonization of the information space accessible to a wide range of users.

Keywords: benchmark, factuality evaluation, factuality in LLM