DOI: 10.15514/ISPRAS-2025-37(3)-12

SLAVA: бенчмарк социально-политического ландшафта и ценностного анализа

[1] *А.С. Четвергов, ORCID: 0009-0004-9787-6785 <chetvergov-as@ranepa.ru>*
[1] *Р.С. Шарафетдинов, ORCID: 0009-0001-2170-8855 <sharafetdinov-rs@ranepa.ru>*
[1] *М.М. Полукошко, ORCID: 0009-0000-8568-041X <polukoshko-mm@ranepa.ru>*
[1] *В.А. Ахметов, ORCID: 0009-0008-2354-8948 <akhmetov-va@ranepa.ru>*
[1] *Н.А. Оружейникова, ORCID: 0009-0007-8783-4444 <oruzheynikova-na@ranepa.ru>*
[1] *Е.С. Аничков, ORCID: 0009-0005-1030-4491 <anichkov-es@ranepa.ru>*
[2] *И.С. Алексеевская, ORCID: 0009-0006-8833-441X <alekseevskaia@ispras.ru>*
[1] *С.В. Боловцов, ORCID: 0000-0003-2342-2663 <bolovtsov-sv@ranepa.ru>*
[1] *П.Е. Голосов, ORCID: 0000-0003-4313-0887 <golosov-pe@ranepa.ru>*

[1] *Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации, Россия, 119571, г. Москва, пр. Вернадского, д. 82.*
[2] *Институт системного программирования им. В.П. Иванникова РАН, Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*

Аннотация. Большим языковым моделям (LLM) находят применение в самых различных областях благодаря растущим способностям в задачах обработки естественного языка. Внедрение LLM в системы, ошибки которых могут нести негативные последствия, требует всестороннего изучения достоверности их работы. Оценка фактуальности LLM позволяет понять, насколько сгенерированный текст соответствует реальным фактам. Существует множество фактологических систем сравнения, но лишь небольшая их часть проверяет знания моделей в российской доменной области. В подобных оценочных стандартах избегают дискуссионных и чувствительных тем, в отношении которых у России существует вполне сформированная позиция. Для преодоления проблемы неполноты чувствительных оценок нами был разработан бенчмарк SLAVA, состоящий из четырнадцати тысяч вопросов в российском домене, представляющих различные области знания. При оценке фактуальности для каждого вопроса измерялось свойство провокативности, определяющее степень чувствительности респондента к запрашиваемой теме. Результаты исследования позволили сформировать рейтинг мультиязычных LLM по ответам на вопросы значимых тематик: истории, политологии, социологии и географии. Проведенное исследование может стимулировать появление новых фактологических систем сравнения, которые будут способствовать гармонизации инфопространства, формированию мировоззренческого суверенитета.

Ключевые слова: система сравнения (бенчмарк); оценка достоверности; фактологичность больших языковых моделей.

Для цитирования: Четвергов А.С., Шарафетдинов Р.С., Полукошко М.М., Ахметов В.А., Оружейникова Н.А., Аничков Е.С., Алексеевская И.С., Боловцов С.В., Голосов П.Е. SLAVA: бенчмарк социально-политического ландшафта и ценностного анализа. Труды ИСП РАН, том 37, вып. 3, 2025 г., стр. 171–184. DOI: 10.15514/ISPRAS–2025–37(3)–12.

Благодарности: Институт общественных наук Российской академия народного хозяйства и государственной службы при Президенте Российской, Институт системного программирования РАН.

SLAVA: Benchmark of Sociopolitical Landscape and Value Analysis

[1] *Chetvergov Andrey, ORCID: 0009-0004-9787-6785 <chetvergov-as@ranepa.ru>*
[1] *Sharafetdinov Rinat, ORCID: 0009-0001-2170-8855 <sharafetdinov-rs@ranepa.ru>*
[1] *Polukoshko Marina, ORCID: 0009-0000-8568-041X <polukoshko-mm@ranepa.ru>*
[1] *Akhmetov Vadim, ORCID: 0009-0008-2354-8948 <akhmetov-va@ranepa.ru>*
[1] *Oruzheinikova Nataliia, ORCID: 0009-0007-8783-4444 <oruzheynikova-na@ranepa.ru>*
[1] *Anichkov Yegor, ORCID: 0009-0005-1030-4491 <anichkov-es@ranepa.ru>*
[2] *Alekseevskaia Irina, ORCID: 0009-0006-8833-441X <alekseevskaia@ispras.ru>*
[1] *Bolovtsov Sergei, ORCID: 0000-0003-2342-2663 <bolovtsov-sv@ranepa.ru>*
[1] *Golosov Pavel, ORCID: 0000-0003-4313-0887 <golosov-pe@ranepa.ru>*

[1] *The Russian Presidential Academy of National Economy and Public Administration, Russia, 119571, Moscow, Vernadsky Ave., 82.*

[2] *Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.*

Abstract. 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 assesses 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 overcome the problem of incompleteness of sensitive assessments, 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 and geography. 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 and the formation of ideological sovereignty.

Keywords: benchmark; factuality evaluation; factuality in LLM.

For citation: Chetvergov A.S., Sharafetdinov R.S., Polukoshko M.M., Akhmetov V.A., Oruzheinikova N.A., Anichkov Ye.S., Alekseevskaia I.S., Bolovtsov S.V., Golosov P.E. SLAVA: benchmark of Sociopolitical Landscape and Value Analysis. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 3, 2025, pp. 171-184 (in Russian). DOI: 10.15514/ISPRAS-2025-37(3)-12.

Acknowledgements. Institute of Social Sciences of the Russian Presidential Academy of National Economy and Public Administration, Institute for System Programming of the Russian Academy of Sciences.

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 достоверности, особенно в таких областях как здравоохранение, юриспруденция, безопасность и государственное управление. Для решения задачи оценки создаются бенчмарки (эталонные системы сравнения), которые позволяют изучить показатели качества LLM по различным критериям при решении определённых задач [1]. Среди них можно выделить фактологические системы сравнения, 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 системах сравнения или выбирают ответы на них, исходя из жизненной позиции исследователей [1, 3, 4]. 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 мировоззренческого суверенитета встаёт особенно остро.

Для решения указанной проблемы нами был создан бенчмарк SLAVA (Sociopolitical Landscape and Value Analysis), - набор оценочных показателей, основанный на группах вопросов, важных для информационного пространства. Для него мы подготовили около 14 тысяч вопросов из различных областей знания, которые на государственном уровне признаны наиболее важными (история, обществознание, география, политология) и разработали свою методологию оценки качества ответов LLM. Кроме того, для каждого вопроса было определено специальное 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 систем сравнения: MMLU [5], C-Eval [6], TruthfulQA [3], Pinocchio [7] 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 набор стандартизированных тестов C-Eval, разработка которого была обусловлена преобладанием систем сравнения с англоязычным контекстом.

Для комплексной оценки предсказаний LLM на русском языке были разработаны: MERA [4], Russian SuperGLUE [8], Rulm-sbs2 [9] и TAPE [10]. Задания в указанных оценочных системах проверяют здравый смысл, целеполагание, общие знания о мире, логику моделей.

Стоит отметить, что во всех указанных оценочных системах чувствительные вопросы, эталонные ответы на которые будут различаться в зависимости от государственной, национальной, религиозной, культурной принадлежности респондента, либо [4, 8-10], либо эталонные ответы на них даны в соответствии с государственной (не российской) принадлежностью исследователей [2, 3, 5-7]. Такая особенность создаёт сложности в оценке применимости LLM в русскоязычных информационных системах.

По результатам проведения обзора существующих русскоязычных сравнительных системах по оценке фактуальности было установлено:

- отсутствие задач для проверки знаний LLM по отечественной истории, обществознанию, политологии и географии;
- отсутствие задач на указание последовательности в наборах вопросов по гуманитарным дисциплинам (например, USE в стандарте оценки MERA [4] по русскому языку);

- отсутствие характеристики чувствительности (степени значимости) вопросов для российской системы знаний и ценностей.

3. Постановка задачи

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

1. Исходные формулировки вопросов, инструкции для LLM и ответы должны быть на русском языке.
2. Вопросы должны проверять фактические знания LLM.
3. Вопросы должны относиться к темам истории, социологии, географии и политологии, – 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 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 ключевых тематик, т.к. они напрямую связаны с российской идентичностью.

4. Подготовка набора данных

Набор данных в первую очередь был сформирован (более 88% всех вопросов) из открытых источников вопросов ЕГЭ [11, 12]. Следовательно, это даёт нам возможность утверждать, что тематика и фактологичность этих данных проверены на соответствие критериям посредством аккредитации такого набора вопросов Федеральной службой по надзору в сфере образования и науки. Другая часть исходных вопросов и ответов была подготовлена нами с привлечением профильных специалистов в области истории, политологии и социологии. Далее была проведена проверка соответствия этих вопросов сформулированным выше критериям. Стоит отметить, что и фактологичность, и соответствие подготовленных вопросов критериям оценивалась группой экспертов коллегиально, учитывая мнение каждого. Состав экспертной комиссии включал профильных специалистов с необходимым научно-педагогическим стажем. В конце полученный набор данных был дополнительно проверен междисциплинарной группой специалистов. В результате было получено около 14 тысяч вопросов по темам истории, обществознания, географии, политологии. Распределение вопросов по темам и типам изображено на рис. 1.

Открытая часть созданного набора данных размещена на сайте Hugging Face [13].

Для проверки фактуальности LLM не только в различных темах, но и на разных форматах ответов, вопросы в предлагаемом нами 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).

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

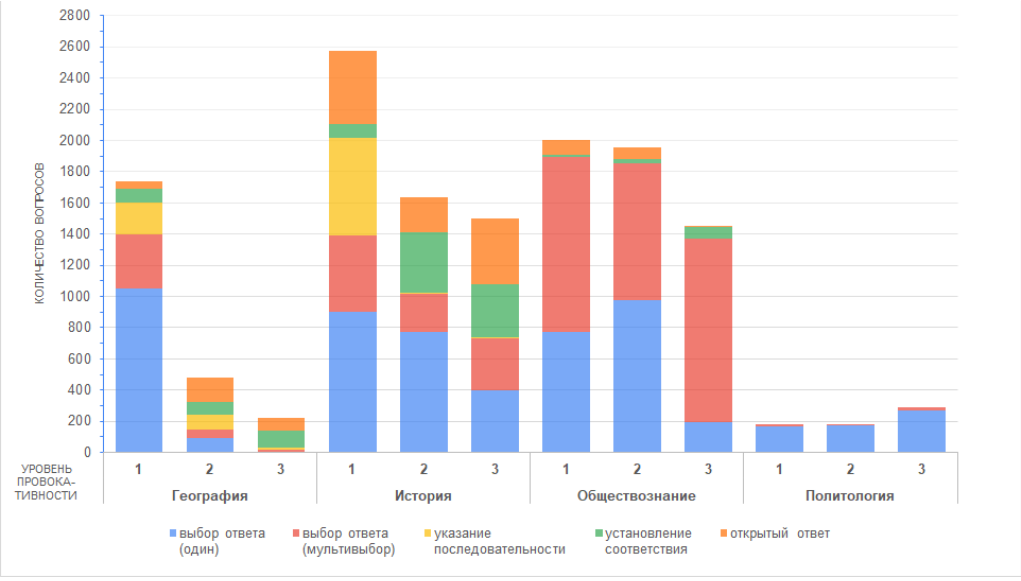

Рис. 1. Распределение вопросов по темам и типам (весь набор данных).
Fig. 1. Distribution of questions by topics and types (entire dataset).

Вопросы низкой чувствительности (провокативности) 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.

Вопросы высокой провокативности 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.

4.1 Примеры вопросов

4.1.2 По области знаний:

1) География:

   С какой из перечисленных стран Россия имеет сухопутную границу?

   - Туркмения;
   - Армения;
   - Эстония;
   - Киргизия.

2) Обществознание:

   Гарантом Конституции РФ выступает:

   - Председатель Конституционного Суда РФ;
   - Вице-спикер Совета Федерации РФ;
   - Президент РФ;
   - Председатель Правительства РФ.

3) История:

   Расположите в хронологической последовательности события:

   - Первое упоминание Москвы в летописи;
   - Разгром немецких рыцарей на Чудском озере;
   - Подавление Тверского восстания ордынско-московским войском;
   - Битва на р. Шелони;
   - Начало правления Владимира Мономаха в Киеве.

4) Политология:

   В каком году в России проводился референдум по новой Конституции?

   - в 1993;
   - в 1996;
   - в 1985;
   - в 1989.

4.1.2 По виду вопроса

1) выбор ответа (один):

   Какая из стран в период 1992−2011 гг. стала членом НАТО?

   - Грузия;
   - Австрия;
   - Черногория;
   - Болгария.

2) выбор ответа (мультивыбор):

Ниже приведён ряд терминов. Все они, за исключением двух, относятся к царствованию Екатерины II. Выберите лишние элементы.

- Вольное экономическое общество;
- Просвещенный абсолютизм;
- Вольные хлебопашцы;
- Уложенная комиссия;
- Государственный совет;
- Вооружённый нейтралитет.

3) указание последовательности:

Расположите в хронологической последовательности события:

- Ялтинская конференция «Большой тройки»;
- Бородинский бой;
- Присоединение Левобережной Украины к России.

4) установление соответствия:

Установите соответствие между событиями и их датами. В ответе запишите номера правильных ответов в порядке, соответствующем буквам А, Б, В, Г.

А) начало кампании по «борьбе с космополитизмом»

Б) волнения и расстрел рабочих в Новочеркасске

В) избрание Генеральным секретарем ЦК КПСС Ю. А. Андропова

Г) прекращение существования СССР

- 1982 г.;
- 1977 г.;
- 1991 г.;
- 1949 г.;
- 1962 г.

5) открытый ответ:

Напишите пропущенное слово (словосочетание).

Изданная Петром I _____, устанавливая принципиально новые критерии служебной годности, предоставляла возможность пополнять дворянское сословие талантливыми выходцами из других социальных групп.

4.1.3 По уровню провокативности

1. 1 уровень:

Создание какого произведения литературы относится к 1920–1930-м гг.?

- Романа В. П. Аксёнова «Остров Крым»;
- Романа А. И. Солженицына «Архипелаг ГУЛАГ»;
- Романа К. М. Симонова «Живые и мёртвые»;
- Романа М. А. Булгакова «Мастер и Маргарита».

2. 2 уровень:

Выберите подходящее окончание для фразы.

Социальное неравенство проявляется в обществах традиционного типа в:

- различном доступе представителей разных сословий к власти и обладанию привилегиями;
- возможности для всех граждан получить образование и социальное обеспечение;
- равных правах граждан в политической сфере, но различии в доходах, в обладании собственностью;
- делении общества на страты, определяемые уровнем образования, дохода, профессией.

3. 3 уровень:

   Какие государства на сегодняшний день являются лидерами в мире по количеству санкционных запретов в отношении России (по состоянию на 2024 год)?

   - страны Евросоюза и США;
   - Белоруссия и Китай;
   - Украина и Казахстан;
   - Сирия и Иран.

5. Процедура оценки

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

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

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

В основе общего рейтинга SLAVA лежит комплексная метрика точности (по шкале от 0 до 100, где 100 – наивысший балл рейтинга), представляющая собой усреднённую оценку ответов модели по каждому типу вопросов:

1. 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);
2. для вопросов вида «мультивыбор, с одним вариантом правильного ответа» применяется только метрика EM;
3. 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 рейтинге, 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 по набору вопросов 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 [14]:

1) gemma: 7b − instruct − v1.1 − q40;
2) gemma2: 27b − instruct − q40
3) gemma2: 9b − instruct − q40
4) GigaChatLite
5) GigaChatP lus
6) GigaChatP ro
7) ilyagusev/saigallama3
8) llama2: 13b
9) llama3.1: 70b − instruct − q40
10) llama3.1: 8b − instruct − q40
11) llama3: 70b − instruct − q40
12) llama3: 8b − instruct − q40
13) mistral: 7b − instruct − v0.3 − q40
14) mixtral: 8x7b − instruct − v0.1 − q40
15) phi3: 14b − medium − 4k − instruct − q40
16) qwen: 7b
17) qwen2: 72b − instruct − q40
18) qwen2: 7b − instruct − q40
19) solar: 10.7b − instruct − v1 − q40
20) wavecut/vikhr: 7b − instruct0.4 − Q41
21) yandexgptlite
22) yandexgptpro
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 гиперпараметры моделей по умолчанию. Также дополнительно были сгенерированы случайные ответы на вопросы для сравнения ответов моделей с ними. Для запросов был использован LLM-запрос с добавлением требования отвечать максимально коротко.

Анализ экспериментов SLAVA (табл. 1) выявил значительные различия в способности LLM справляться с заданиями по ключевым тематикам в разрезе трёх уровней провокативности (на IV квартал 2024 года). Рассмотренные модели продемонстрировали разнообразие в

результатах, что позволяет оценить их способность решать задачи 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 демонстрируют 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.

Табл. 1. Результаты применения различных стратегий.
Table 1. Results for different strategies.

№	Модель	Область знаний	Провокативность	Итоговый рейтинг (по виду вопроса)
1	qwen2:72b-instruct-q4_0	58,78	53,51	**52,39**
2	GigaChat_Pro	**61,02**	**57,60**	48,43
3	yandexgpt_pro	57,81	52,75	40,90
4	GigaChat_Plus	53,86	49,74	38,33
5	GigaChat_Lite	53,84	49,69	38,32
6	gemma2:9b-instruct-q4_0	54,23	46,13	34,88
7	llama3:70b-instruct-q4_0	51,95	47,23	32,76
8	yandexgpt_lite	51,41	46,78	28,09
9	llama3.1:70b-instruct-q4_0	47,46	44,60	27,24
10	qwen2:7b-instruct-q4_0	36,07	28,48	20,42
11	ilyagusev/saiga_llama3	38,15	29,97	17,44
12	phi3:14b-medium-4k-instruct-q4_0	35,37	28,37	17,00
13	mistral:7b-instruct-v0.3-q4_0	27,96	21,75	12,48
14	solar:10.7b-instruct-v1-q4_0	27,76	22,38	11,93
15	random	20,01	17,49	11,60
16	wavecut/vikhr:7b-instruct_0.4-Q4_1	25,35	21,77	11,11
17	llama3:8b-instruct-q4_0	28,21	23,33	10,47
18	mixtral:8x7b-instruct-v0.1-q4_0	22,94	18,55	10,45
19	yi:9b	19,55	15,44	9,94
20	gemma:7b-instruct-v1.1-q4_0	23,18	18,03	9,93
21	llama3.1:8b-instruct-q4_0	25,39	21,86	9,85
22	qwen:7b	18,13	13,75	9,16
23	gemma2:27b-instruct-q4_0	20,52	16,81	8,68
24	yi:6b	12,95	10,54	5,36
25	llama2:13b	9,84	8,07	3,81
Среднее значение		**35,27**	**30,58**	**20,84**

Результаты по метрикам, таким как точность ответа, совпадение с правильным ответом и частичная верность, показывают, что оценки большинства моделей превосходят метрики случайных ответов. 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 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 значения метрик около 12.00-13.00 в тех же категориях.

Важно отметить, что для значимых вопросов, связанных с российской идентичностью, 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 моделями с доступом по API 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.

7. Заключение

Результаты исследования SLAVA позволили сформировать рейтинг мультиязычных LLM по их ответам на вопросы значимых тематик: история, обществознание, география, политология. Несмотря на то, что несколько моделей достигли средних показателей, большинство показало низкие результаты, что вызывает необходимость дальнейших исследований достоверности предсказаний LLM.

К дальнейшим направлениям нашей работы мы относим расширение корпуса вопросами, разработанными на основе официальных источников (нормативно-правовые акты Российской Федерации и др.), добавление в рейтинг новых моделей, включение оценки ответов людей в итоговую таблицу результатов («human baseline»), ежеквартальный пересмотр содержания системы тестов, заключающийся в добавлении новых вопросов по актуальным темам, удалении или обновлении устаревших вопросов), а также пересмотр оценок провокативности с учётом изменений в общественном контексте.

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 / References

[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: 07.11.2024).
[12]. ЕГЭ-2024, Математика профильного уровня: задания, ответы, решения [Электронный ресурс]. URL: https://math-ege.sdamgia.ru/ (дата обращения: 07.11.2024).
[13]. SLAVA: Benchmark of the Socio-political Landscape and Value Analysis, открытая часть набора данных [Электронный ресурс]. URL: https://huggingface.co/datasets/RANEPA-ai/SLAVA-OpenData-2800-v1 (дата обращения: 07.11.2024).
[14]. Ollama [Электронный ресурс]. URL: https://ollama.com (дата обращения: 07.11.2024).

Информация об авторах / Information about authors

Андрей Сергеевич ЧЕТВЕРГОВ – специалист Лаборатории интеллектуальной аналитики Исследовательского центра искусственного интеллекта ИОН Президентской академии. Сфера научных интересов: разработка и оптимизация моделей машинного обучения, глубокое обучение, обработка естественного языка, автоматизация процессов машинного обучения, исследование новых алгоритмов искусственного интеллекта, междисциплинарные исследования.

Andrey Sergeevich CHETVERGOV is a specialist at the Laboratory of Intelligent Analytics at the Research Center for Artificial Intelligence, Institute for Social Sciences (ISS), Presidential Academy. Research interests include the development and optimization of machine learning models, deep learning, natural language processing, automation of ML processes, exploration of novel artificial intelligence algorithms, and interdisciplinary research.

Ринат Саярович ШАРАФЕТДИНОВ - специалист Лаборатории интеллектуальной аналитики Исследовательского центра искусственного интеллекта ИОН Президентской академии. Сфера научных интересов: проектирование и улучшение моделей машинного обучения, изучение методов обработки естественного языка, исследование новых подходов в работе больших языковых моделей.

Rinat Sayarovich SHARAFETDINOV is a specialist at the Laboratory of Intelligent Analytics at the Research Center for Artificial Intelligence, Institute for Social Sciences (ISS), Presidential Academy. Research interests include the design and enhancement of machine learning models, the study of natural language processing methods, and the exploration of novel approaches in working with large language models (LLMs).

Марина Михайловна ПОЛУКОШКО – заведующий Лаборатории интеллектуальной аналитики Исследовательского центра искусственного интеллекта ИОН Президентской академии. Сфера научных интересов: стратегическое управление проектами в области анализа данных и машинного обучения, развитие и применение больших языковых моделей, исследования доверенности систем ИИ, междисциплинарные исследования.

Marina Mikhailovna POLUKOSHKO is the Head of the Laboratory of Intelligent Analytics at the Research Center for Artificial Intelligence, Institute for Social Sciences (ISS), Presidential

Academy. Research interests include strategic management of data analysis and machine learning projects, the development and application of large language models (LLMs), trustworthiness studies of AI systems, and interdisciplinary research.

Вадим Аксанович АХМЕТОВ – эксперт Лаборатории интеллектуальной аналитики Исследовательского центра искусственного интеллекта ИОН Президентской академии. Сфера научных интересов: анализ больших данных, построение прогнозных моделей, временные ряды, компьютерное зрение, интерпретация моделей машинного обучения.

Vadim Aksanovich AKHMETOV is an expert at the Laboratory of Intelligent Analytics at the Research Center for Artificial Intelligence, Institute for Social Sciences (ISS), Presidential Academy. Research interests include big data analysis, development of predictive models, time series analysis, computer vision, and interpretability of machine learning models.

Наталия Андреевна ОРУЖЕЙНИКОВА – аналитик Лаборатории интеллектуальной аналитики Исследовательского центра искусственного интеллекта ИОН Президентской академии. Сфера научных интересов: обработка больших данных, статистическое моделирование, визуализация данных, прогнозная аналитика и оптимизация бизнес-процессов.

Natalia Andreevna ORUZHEYNIKOVA is an analyst at the Laboratory of Intelligent Analytics at the Research Center for Artificial Intelligence, Institute for Social Sciences (ISS), Presidential Academy. Research interests include big data processing, statistical modeling, data visualization, predictive analytics, and business process optimization.

Егор Сергеевич АНИЧКОВ – ведущий специалист Лаборатории интеллектуальной аналитики Исследовательского центра искусственного интеллекта ИОН Президентской академии. Сфера научных интересов: обработка естественного языка, научная экспертиза проектов, связанных с большими языковыми моделями, исследование факторов доверенности интеллектуальных систем.

Egor Sergeevich ANICHKOV is a leading specialist at the Laboratory of Intelligent Analytics at the Research Center for Artificial Intelligence, Institute for Social Sciences (ISS), Presidential Academy. His research interests include natural language processing, scientific evaluation of projects related to large language models (LLMs), and the study of trustworthiness factors in intelligent systems.

Ирина Сергеевна АЛЕКСЕЕВСКАЯ – программист Центра доверенного искусственного интеллекта, аспирант ИСП РАН по направлению искусственный интеллект и машинное обучение. Сфера научных интересов: большие языковые модели, состязательные атаки, бэкдор атаки, выравнивание больших языковых моделей.

Irina Sergeevna ALEKSEEVSKAIA – Programmer at the Trusted Artificial Intelligence Research Center, postgraduate student at the ISP RAS in the field of artificial intelligence and machine learning. Research interests: large language models, adversarial attacks, backdoor attacks, alignment of large language models.

Сергей Владимирович БОЛОВЦОВ – директор Исследовательского центра искусственного интеллекта ИОН Президентской академии. Сфера научных интересов: оптимизация и масштабирование инфраструктуры для работы с большими данными и ML, анализ данных и управление качеством данных, продвинутые методы обработки естественного языка, применение больших языковых моделей в междисциплинарных исследованиях.

Sergey Vladimirovich BOLOVTSOV is the Director of the Research Center for Artificial Intelligence, Institute for Social Sciences (ISS), Presidential Academy. Research interests include

the optimization and scaling of infrastructure for big data and machine learning workflows, data analysis and quality management, advanced natural language processing (NLP) methods, and the application of large language models (LLMs) in interdisciplinary research.

Павел Евгеньевич ГОЛОСОВ – директор Института общественных наук Президентской академии. Сфера научных интересов: технологические вызовы и искусственный интеллект, экономика данных и внедрение искусственного интеллекта, индивидуализированный подход в высшем образовании, применение искусственного интеллекта в образовании.

Pavel Evgenievich GOLOSOV is the Director of the Institute for Social Sciences (ISS) at the Presidential Academy. Research interests include technological challenges and artificial intelligence, the data economy and AI implementation, an individualized approach in higher education, and the application of AI in education.