ПРИМЕНЕНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ РЕШЕНИЯ ЛИНГВИСТИЧЕСКИХ ЗАДАЧ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ НА МАТЕРИАЛЕ ТЕСТОВЫХ ЗАДАНИЙ ПО РУССКОЙ ЛЕКСИКОЛОГИИ

А. А. Голиков, Ю. Ю. Данилова, Д. А. Акимов
Елабужский институт Казанского федерального университета
ООО «Мастерская цифровых решений»
г. Елабуга, Москва

В настоящее время большие языковые модели находят все большее применение в различных отраслях знаний, при этом для оценки качества работы больших языковых моделей применяются те или иные виды тестирования, бенчмарков (наборов задач, на которых тестируются модели и сопоставляются ответы). Одним из наиболее распространенных бенчмарков для тестирования больших языковых моделей на сегодняшний день является бенчмарк MMLU, который предполагает ответы моделей на вопросы из разных областей знаний в формате выбора одного правильного ответа из нескольких вариантов. Однако в данном и иных основных используемых бенчмарках не тестируется способность моделей глубоко понимать русский язык, его единицы (лексемы и фразеологизмы), их дифференциальные признаки, лексико-семантические варианты, внутреннюю форму, системные связи, социолингвистическую обусловленность. В данной работе производится сравнительный анализ качества работы основных зарубежных и отечественных больших языковых моделей для решения лингвистических задач в виде тестовых заданий по курсу «Лексикология современного русского языка» в системе высшего филологического образования. В итоге было выявлено, что наилучшие результаты как суммарно, так и по отдельным блокам вопросов продемонстрировала модель Claude 3 Opus от компании Anthropic, за ней следуют GPT-4 от OpenAI и GigaChat Pro от Сбера. Анализ результатов по тематическим блокам показал, что наибольшие трудности у моделей вызвали задания по фразеологии, а наилучшие результаты были достигнуты в блоке вопросов по лексикографии.

Ключевые слова: большая языковая модель, бенчмарк, сравнительный анализ, лингвистика, русский язык, лексикология

APPLICATION OF LARGE LANGUAGE MODELS TO SOLVE LINGUISTIC PROBLEMS: A COMPARATIVE ANALYSIS ON THE MATERIAL OF TEST TASKS IN RUSSIAN LEXICOLOGY

A. A. Golikov, Yu. Yu. Danilova, D. A. Akimov
Elabuga Institute of Kazan Federal University
LLC «Digital solutions workshop»
Elabuga, Moscow

Nowadays, large language models are increasingly used in various branches of knowledge, and in order to assess the quality of large language models' performance, one or another type of testing, benchmarks (i.e. sets of tasks on which the models are tested and their answers are compared) are used. One of the most widespread benchmarks for testing large language models today is the MMLU benchmark, which involves answering questions from different areas of knowledge in the format of selecting one correct answer from several answers. However, this and other main benchmarks in use do not test the models' ability to deeply understand the Russian language, its units (lexemes and phraseological phrases), their differential features, lexico-semantic variants, internal form, systemic relations, sociolinguistic conditionality, etc. The present paper makes a comparative analysis of the quality of work of the main foreign and domestic large language models for solving linguistic problems in the form of test tasks for the course "Lexicology of the modern Russian language" in the system of higher philological education. As a result of the study, it was found that the Claude 3 Opus model from Anthropic showed the best results both in total and in individual question blocks, followed by GPT-4 from OpenAI and GigaChat Pro from Sber. Analysing the results by subject blocks showed that the models had the most difficulty in phraseology, while the best results were achieved in the lexicography block of questions.

Keywords: large language models, benchmark, comparative analysis, linguistics, Russian language, lexicology

Большие языковые модели в настоящее время являются быстро развивающейся технологией, обладающей значительным потенциалом. Разные страны, в том числе Россия, стремятся развивать свои большие языковые модели для обеспечения технологического суверенитета. Одним из важных вопросов является оценка качества их работы [1; 2; 3] — для этого используются тесты, бенчмарки, содержащие большое количество вопросов по разным отраслям знаний во всех сферах жизни. Наиболее известные мировые бенчмарки — MMLU [4] и MT-Bench [5] с вопросами и ответами на английском языке, также существует отечественный бенчмарк MERA [6] с вопросами на русском.

Помимо этого, в литературе можно найти примеры использования специализированных тестов и экзаменационных заданий для оценки конкретных аспектов языковых моделей. Так, в [7] использовались задачи из математических олимпиад для тестирования способности моделей к математическому рассуждению. В [8] авторы применяли задания из вступительных экзаменов по математике для оценки способности моделей к количественным рассуждениям. В [9] оценивалась возможность моделей по решению задач из области физики, математики и информатики, используя вопросы из университетских курсов и олимпиад. Эти исследования показывают, что использование узкоспециализированных тестов позволяет получить более глубокое понимание сильных и слабых сторон языковых моделей.

Несмотря на значительное количество работ в этой области, большинство из них не фокусируется на оценке глубокого понимания моделями самого языка, его структуры и специфических лингвистических аспектов. И остается открытым вопрос о том, насколько хорошо большие языковые модели способны справляться с узкоспециализированными лингвистическими задачами, требующими не только общих знаний, но и понимания устройства языка, например, в области лексикологии и лексикографии. Настоящее исследование призвано восполнить этот пробел и оценить возможности использования языковых моделей для решения таких задач на материале тестов по русской лексикологии.

Отметим, что лексикология современного русского литературного языка является основным разделом в системе вузовского филологического образования, предметом изучения которого становятся лексемы (слова) и фразеологизмы во всем многообразии их признаков, лексико-семантических вариантов, синтагматических и парадигматических отношений внутри различных групп лексики и фразеологии, а также с учетом социолингвистической природы данных единиц языка. В данной работе в качестве эмпирического материала исследования обозначенного вопроса были выбраны контрольные тесты по курсу «Лексикология» в рамках предметного модуля «Современный русский язык», использующиеся в учебном процессе для студентов-филологов 2 курса отделения филологии и истории Елабужского института Казанского федерального университета (ЕИ КФУ). Общий объем тестовых заданий составил 185 вопросов. Тематика, примеры вопросов и ответов приведены в табл. 1.

Таблица 1. Примеры вопросов и ответов для тестирования качества работы больших языковых моделей с целью решения лингвистических задач (на материале тестовых заданий по курсу «Лексикология современного русского языка»)

Блок знаний	Общее количество вопросов в блоке	Пример вопроса	Правильный ответ
Лексикография	25	Какой из словарей не относится к аспектным? а) словарь сочетаемости слов; б) словарь эпитетов; в) словарь синонимов; г) словарь паронимов; д) толковый словарь	д
Слово как единица языка. Лексическое значение слова	30	На каком основании выделяются такие типы значений, как: свободное и связанное (фразеологически связанное, конструктивно ограниченное, синтаксически обусловленное)? а) по соотношению языка и речи; б) по возможности лексической сочетаемости; в) по способу номинации (соотнесенности с называемым предметом, реалией); г) по степени семантической мотивированности; д) по характеру выполняемых функций	б
Системные отношения в лексике	30	Выделите явление, которое не отражает парадигматические отношения в лексике: а) антонимическая пара; б) синонимический ряд; в) лексическая сочетаемость; г) тематическая группа; д) семантическое поле	в

Продолжение таблицы 1

Блок знаний	Общее количество вопросов в блоке	Пример вопроса	Правильный ответ
Активная и пассивная лексика	30	Выделите ряд, в котором представлены архаизмы: а) прасол, продразверстка, престолонаследие; б) камергер, кичка, целовальник; в) опричник, шишак, ротмистр; г) смерд, брадобрей, бортничать; д) лапти, армяк, шапокляк	г
Происхождение слов	30	Определите ряд слов, имеющих признаки собственно русских слов: а) единство, алиби, какаду; б) младенец, пещера, герцог; в) молочник, летчик, урод; г) хетты, сундук, башлык; д) бювар, хаос, фарисей	в
Фразеология	40	Выделите словосочетания, свободный или фразеологический характер которых выявляется только в контексте: а) точить лясы, бить баклуши; б) развязать узел, махнуть рукой; в) яко тать в нощи, паче чаянья; г) тянуть канитель, не видно ни зги; д) бить челом, попасть впросак	б

Для тестирования были выбраны передовые мировые и отечественные большие языковые модели:

- GPT-4 (проприетарная модель от компании OpenAI);
- GPT-3.5 (проприетарная модель от компании OpenAI);
- Claude 3 Opus (проприетарная модель от компании Anthropic);
- Qwen-1.5-72B (open source модель от компании Alibaba, на текущий момент лучшая open source языковая модель согласно рейтингу [10]);
- GigaChat Pro (проприетарная модель от компании Сбер, на текущий момент лучшая отечественная языковая модель согласно бенчмарку MERA);
- GigaChat Lite (проприетарная модель от компании Сбер, упрощенная версия GigaChat Pro).

Важно отметить, что дополнительное обучение или настройка данных моделей на материале курса «Лексикология современного русского языка» не проводились. В исследовании использовались модели «в чистом виде», чтобы оценить их базовые возможности по решению лингвистических задач без специальной подготовки.

Для обеспечения одинаковых условий тестирования использовался следующий промпт для всех моделей — «Вопрос: … Варианты ответов: а) … б) … в) … г) … д) …. Ответь только одной буквой (а, б, в, г или д), соответствующей правильному варианту ответа.», где вместо многоточий соответственно подставлялись вопросы и варианты ответа.

Рассмотренные модели показали следующие результаты (табл. 2).

Таблица 2. Процент правильных ответов больших языковых моделей для решения лингвистических задач (на материале тестовых заданий по курсу «Лексикология современного русского языка»)

Раздел	Claude 3 Opus	GPT-4	GigaChat Pro	Qwen-1.5-72B	GigaChat Lite	GPT-3.5
Лексикография	88,00	76,00	72,00	64,00	64,00	48,00
Слово как единица языка. Лексическое значение слова	73,33	60,00	60,00	46,67	43,33	46,67
Системные отношения в лексике	83,33	66,67	50,00	46,67	40,00	53,33
Активная и пассивная лексика	76,67	63,33	56,67	46,67	50,00	33,33
Происхождение слов	93,33	70,00	43,33	33,33	26,67	40,00
Фразеология	77,50	52,50	60,00	42,50	27,50	17,50
Всего	81,62	63,78	56,76	45,95	40,54	38,38

Для наглядности отобразим процент правильных ответов на все вопросы из всех блоков вместе на гистограмме (рисунок).

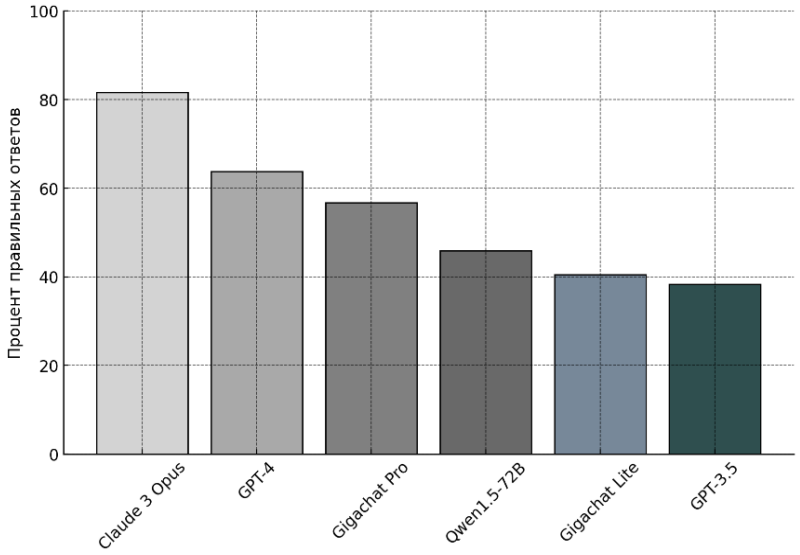

Рисунок. Процент правильных ответов больших языковых моделей для решения лингвистических задач (на материале тестовых заданий по курсу «Лексикология современного русского языка»)

Таким образом, в результате проведенного тестирования качества работы больших языковых моделей и последующего сравнительного анализа полученных данных можно сделать следующие выводы.

Наилучшие результаты как суммарно, так и по отдельным блокам вопросов, продемонстрировала проприетарная большая языковая модель Claude 3 Opus от компании Anthropic, выпущенная позже остальных рассмотренных моделей (4 марта 2024 г.). Хотя на текущий момент в рейтинге [10] данная модель делит первое место с GPT-4, в задачах филологии и лингвистики на русском языке Claude 3 Opus значительно ее превзошла.

Второе место в тестировании как суммарно, так и по отдельным блокам вопросов, заняла модель GPT-4 от компании OpenAI, лидирующая на текущий момент в рейтинге [10]. Лишь в блоке вопросов по фразеологии GPT-4 уступила отечественной модели GigaChat Pro от компании Сбер.

Лучшая отечественная модель — GigaChat Pro от компании Сбер — заняла лишь третье место, хотя вопросы тестов были весьма специализированными и касались основного языка модели (русского языка). Таким образом, флагманской отечественной модели GigaChat Pro требуется дальнейшая оптимизация, если предполагается реализовывать ее основное конкурентное преимущество перед мировыми моделями — более тонкое понимание русского языка.

Четвертое место в тестировании заняла open source модель Qwen-1.5-72B от компании Alibaba, которая смогла опередить такие проприетарные модели GigaChat Lite и GPT-3.5. Это указывает на высокий уровень лучшей (согласно рейтингу [10]) на сегодняшний день open source модели и большой потенциал ее использования при разворачивании на собственных серверах и работе с конфиденциальными данными, которые не должны выйти за пределы контура компании.

Анализ результатов по отдельным тематическим блокам вопросов показывает, что наибольшие трудности у моделей вызвали задания по фразеологии (средний процент правильных ответов — 46,25 %), что может быть связано с высокой идиоматичностью и непрозрачностью внутренней формы фразеологизмов. Наилучшие результаты были достигнуты в блоке вопросов по лексикографии (средний процент правильных ответов — 68,67 %), что свидетельствует о хорошей способности моделей работать со словарными дефинициями и различать типы словарей.

ЛИТЕРАТУРА

1. Chang Y., Wang X., Wang J., Wu Y., Yang L., Zhu K., Chen H., Yi X., Wang C., Wang Y., Ye W., Zhang Y., Chang Y., Philip S. Yu, Yang Q., Xie X. A survey on evaluation of large language models // ACM Transactions on Intelligent Systems and Technology. 2024. Т. 15, № 3. P. 1–45. DOI: 10.1145/3641289.
2. Guo Z., Jin R., Liu C., Huang Y., Shi D., Supryadi, Yu L., Liu Y., Li J., Xiong B., Xiong D. Evaluating large language models: A comprehensive survey // arXiv preprint arXiv:2310.19736. 2023. DOI: 10.48550/arXiv.2310.19736.
3. Тимаков К. А. Сравнение актуальных языковых моделей Google Bard и ChatGPT // Международная научно-практическая конференция «Современные стратегии и цифровые трансформации устойчивого развития общества, образования и науки». М., 2023. С. 168–171. DOI: 10.34755/IROK.2023.93.45.076.

4. Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., Steinhardt J. Measuring massive multitask language understanding // arXiv preprint arXiv:2009.03300. 2020. DOI: 10.48550/ arXiv:2009.03300.
5. Zheng L., Chiang W., Sheng Y., Zhuang W., Wu Z., Zhuang Y., Lin Z., Li Z., Li D., Xing E., Zhang H., Gonsalez J., Stoica I. Judging llm-as-a-judge with mt-bench and chatbot arena // Advances in Neural Information Processing Systems. 2024. Vol. 36. DOI: 10.48550/arXiv.2306.05685.
6. Fenogenova A., Chervyakov A., Martynov N., Kozlova A., Tikhonova M., Akhmetgareeva A., Emelyanov A., Shevelev D., Lebedev P., Sinev L., Isaeva U., Kolomeytseva K., Moskovskiy D., Goncharova E., Savushkin N., Mikhailova P., Dimitrov D., Panchenko A., Markov S. MERA: A Comprehensive LLM Evaluation in Russian // arXiv preprint arXiv:2401.04531. 2024. DOI: 10.48550/arXiv.2401.04531.
7. Cobbe K., Kosaraju V., Bavarian M., Chen M., Jun H., Kaiser L., Plappert M., Tworek J., Hilton J., Nakano R., Hesse C., Schulman J. Training verifiers to solve math word problems // arXiv preprint arXiv:2110.14168. 2021. DOI: 10.48550/arXiv.2110.14168.
8. Lewkowycz A., Andreassen A., Dohan D., Dyer E., Michalewski H., Ramasesh V., Slone A., Anil C., Schlag I., Gutman-Solo T., Wu Y., Neyshabur B., Gir-Ari G., Misra V. Solving quantitative reasoning problems with language models // Advances in Neural Information Processing Systems. 2022. T. 35. C. 3843–3857. DOI: 10.48550/arXiv.2206.14858
9. Drori I., Zhang S., Shuttleworth R., Tang L., Lu A., Ke E., Liu K., Chen L., Tran S., Cheng N., Wang R., Singh N., Patti T., Lynch J., Shporer A., Verma N., Wu E., Strang G. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level // Proceedings of the National Academy of Sciences. 2022. T. 119, № 32. DOI: 10.1073/pnas.2123433119.
10. LMSYS Chatbot Arena Leaderboard // Hugging Face. URL: https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard (дата обращения: 23.06.2024).