Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2025”

April 23–25, 2025

Cultural Evaluation of LLMs in Russian: Catchphrases and Cultural Types

Elizaveta Gromenko
HSE
Moscow, Russia
el.gromenko@gmail.com

Daria Kalacheva
MIPT
Moscow, Russia
TheDariaK@yandex.ru

Ksenia Klokova
MIPT
Moscow, Russia
klokova.ks@mipt.ru

Maxim Krongauz
MIPT, RSUH
Moscow, Russia
krongauz.ma@mipt.ru

Oksana Moroz
BHSAD
Moscow, Russia
omoroz@britishdesign.ru

Valery Shulginov
MIPT
Moscow, Russia
shulginov.va@mipt.ru

Tatiana Yudina
MIPT
Moscow, Russia
yudina.tatiana.a@mipt.ru

Abstract

This study addresses the gap in evaluating large language models' (LLMs) cultural awareness and alignment within the Russian sociocultural context by introducing a structured framework comprising 8 Cultural Types (e.g., Spiritual Practitioner, Soviet Intellectual) and 5 catchphrase groups (e.g., memes, proverbs). A 400-question evaluation dataset was developed to probe 10 multilingual LLMs, including GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro, across fact-based cultural knowledge and nuanced linguacultural understanding in a zero-shot setting. Results show that closed-source models GPT-4o and Claude 3.5 Sonnet outperform other models, with one of the smallest models (Mistral NeMo 12B) achieving the lowest result. Performance disparities were noticed in separate evaluation on Cultural Type tasks and catchphrases. Model-specific skews emerged, with lower-ranked models showing inclination toward specific cultural types. Qualitative analysis revealed common errors, such as selecting synonymous but incorrect answers or failing to grasp culturally specific logic. The contribution outlines the limitations of LLMs in interpreting cultural context and lays the groundwork for further research in assessing the cultural-linguistic alignment of LLMs.

Keywords: LLM; Russian language; catchphrases; cultural type; LLM evaluation; QA tasks
DOI: 10.28995/2075-7182-2025-23-XX-XX

Культурные замеры больших языковых моделей на русском языке: речевые клише и культурные типы

Елизавета Громенко
НИУ ВШЭ
Москва, Россия
el.gromenko@gmail.com

Дарья Калачева
МФТИ
Москва, Россия
TheDariaK@yandex.ru

Ксения Клокова
МФТИ
Москва, Россия
klokova.ks@mipt.ru

Максим Кронгауз
МФТИ, РГГУ
Москва, Россия
krongauz.ma@mipt.ru

Оксана Мороз
БВШД
Москва, Россия
omoroz@britishdesign.ru

Валерий Шульгинов
МФТИ
Москва, Россия
shulginov.va@mipt.ru

Татьяна Юдина
МФТИ
Москва, Россия
yudina.tatiana.a@mipt.ru

Аннотация

Наше исследование посвящено изучению культурной осведомленности больших языковых моделей о современном русскоязычном социокультурном контексте. Для этого предложена структурированная система, включающая 8 культурных типов (например, Духовный практик, Советский интеллигент) и 5 групп речевых клише. На основе этой системы был разработан набор данных из 400 вопросов различных форматов для оценки фактологического знания культурных особенностей и более тонкого лигвокультурного понимания в условиях zero-shot. Результаты тестирования 10 мультиязычных LLM (включая GPT-4o, Claude 3.5 Sonnet и Gemini 1.5 Pro) демонстрируют превосходство закрытых моделей GPT-4o и Claude 3.5, тогда как наименьший результат показала компактная модель Mistral NeMo 12B. Выявлены различия в результатах моделей при раздельной оценке заданий на культурные типы и речевые клише. Обнаружены специфические смещения: менее эффективные модели демонстрировали склонность к определенным культурным типам. Качественный анализ выявил типичные ошибки, включая выбор синонимичных, но некорректных, ответов или неспособность распознать культурно-специфичную логику. Исследование подчеркивает ограничения LLM в интерпретации культурного контекста и формирует основу для дальнейших исследований оценки культурно-лингвистической согласованности языковых моделей.

Ключевые слова: большие языковые модели; русский язык, речевые клише; культурный тип; оценка больших языковых моделей; вопросно-ответные задачи

1 Introduction

The rapid advancement of contemporary generative technologies raises questions about the level of integration of large language models (LLMs) into the cultural environment and their ability to consider users' axiological orientations. In this context, tools for assessing the cultural competencies of the LLMs are being actively developed. One of the main lines of research is the evaluation of cultural alignment and biases in the models (Cao et al., 2023; AlKhamissi et al., 2024; Naous et al., 2024; Rao et al., 2024). Another research direction focuses on probing and benchmarking LLMs for their cultural knowledge in specific domains and languages. Those include the evaluation of commonsense knowledge in diverse cultures (Myung et al., 2024; Shen et al., 2024; Koto et al., 2024); evaluation against cultural dimensions (Son et al., 2023; Yin et al., 2024; Wang et al., 2024; Kim et al., 2024); and probing for cultural norms (Fung et al., 2023) and values (Arora et al., 2023; Zhao et al., 2024; Karinshak et al., 2024).

Several benchmarks have been developed for the Russian language to assess general language proficiency, as well as ethical, logical, and other competencies. The largest task sets include RussianSuperGLUE (Shavrina et al., 2020), MERA (Fenogenova et al., 2024), and TAPE (Taktasheva et al., 2022).

Despite the existence of such comprehensive general assessment tools, methods for the evaluation of cultural awareness in LLMs for Russian are still missing.

The present study attempts to fill this gap by developing the methodology which allows to probe LLMs for the cultural awareness and alignment across diverse aspects of contemporary Russian culture. We operationalize this through the development of (1) a taxonomy of 8 distilled Cultural Types and (2) 5 distinct catchphrase groups. These constructs form the basis of a structured evaluation framework, which we employ to construct an evaluation dataset and conduct evaluation of 10 large language models.

The rest of this paper is structured as follows. First, we describe the methodology of identifying the cultural types, provide their brief description, and describe the groups of catchphrases. Section 4 reviews the evaluation dataset. In Section 5 we describe the evaluation setup, which includes models, prompts and metrics. We report on the evaluation results in Section 6, and conclude the paper in Section 7.

2 Cultural Types

The present section explicates the analytical framework developed to study the interaction of LLMs with the Russian national cultural context. By synthesizing methodological principles derived from interpretive sociology, cultural anthropology, and linguacultural studies, we establish and operationalize the conceptual category of "cultural type". A cultural type is defined by specific traits:

- Each cultural type is characterized by shared social norms, values, cultural activities, and historical backgrounds that shape a specific collective identity and perspectives on the cultural landscape (Tönnies, 2001). Representatives of a particular cultural type can articulate their sense of belonging as part of an in-group, which helps them distinguish themselves from "others";
- For the purposes of analysis each cultural type can also be seen as a distinct linguacultural type (LCT), possessing a unique speech repertoire, specific linguistic strategies for expressing cultural preferences, and distinctive behaviors in cultural consumption (Lutovinova, 2009);
- Hence cultural types exhibit specific colloquial usages that reflect unique cultural values, these can be explored through Hofstede's cultural dimensions theory (Hofstede, 2001). For instance, cultural types may represent either communities of practice, where identities are articulated through shared actions and goals, or communities of interest, where identities form around common interests and values (Lave and Wenger, 1991; Eckert, 2006);
- Following Hofstede's theory, it was concluded that the cultural code (Lotman, 2000) governing language behavior within these types is somewhat transgenerational, allowing different generations to coexist within the same cultural type.

Following interviews with the participants of the research team (conducted through a method similar to focus groups) and consultations grounded in empirical ethnographic observation (Zubarevich and Zubarevich, 2010), eight distinct cultural types were identified. Their short descriptions are provided below.

Basic Type encompasses individuals with a fundamental level of knowledge essential for effective integration into the cultural landscape of Russian society. A quintessential example of this type is a high school graduate. Those within this category are diverse and may hold various social statuses, occupations, interests, and preferences.

Careerist-Achiever covers people who stick to a pragmatic approach to life and career. They may be involved in common activities but rather represent a community of interest, sharing the values of productivity, efficiency, pragmatism, and objectivity.

IT Visionary indicates innovators and technoptimists who are deeply involved in the production and dissemination of technologies — such as IT specialists, engineers, and analysts. They tend to be open-minded and at the same time prioritize thought-provoking cultural content that reflects their interests (science fiction, dystopias, and so on).

Modern Intellectual indicates representatives of contemporary creative professions who are actively engaged in various forms of intellectual labor. This may, but do not always, result in the creation and consumption of commercialized and widely sought-after "creative products," (i.e., intellectual property: books, films, plays, etc.).

Nonconformist as a type is united by hyperconsumption of certain ideas and cultural texts, which are not currently mainstream. This category includes individuals who identify with countercultural values, as well as those who temporarily align themselves with specific subcultural movements.

Spiritual Practitioner embraces those who undertake different spiritual practices for the purpose of cultivating spiritual development. They may be deeply involved in religion, philosophy, or some pseudoscientific theories like astrology and tarot cards. Thus, on the verbal level, they may be indicated through certain rituals, actions, and the use of "magical" spells and words.

Soviet Intellectual as a cultural type is deeply engaged with both Soviet ideological culture and dissident narratives. As a custodian of "domestic" modernist perspectives, this type embodies the complexities of the Soviet cultural landscape, but probably excluded from the cultural context generated in the digital environment.

Trend Watcher is characterized by active consumption of mass culture products, heightened awareness of changes within the cultural landscape, strong focus on contemporary agendas, and active digital socialization. Its representatives' communication often includes slang, references to cultural phenomena, and memes, reflecting their active participation in today's cultural discourse.

Each cultural type has its unique background, encompassing specific cultural knowledge and preferences. To ensure transparency in this research, the background was organized by referencing various social and cultural domains, including arts, media, science, politics, religion, and sports, as well as concepts related to everyday life. Specific artifacts from these areas, such as cultural texts, quotes, and names, were used to create thematic maps for each type. Validation of thematic map content was conducted through cross-checking within the research group. The identified spectrum of cultural types allows us to focus on significant cultural groups that can be considered essential to Russian society and enables granular analysis. However, the spectrum can be extended and specified in further studies by other methods such as focus groups, autoethnographic research, and so on.

3 Catchphrases

The cultural commonsense knowledge in our research is represented in two parts: the Basic Type described above and a separate block that contains catchphrases. Catchphrases reflect elements of culture and represent a kind of cultural language code that refers to something well-known and generally significant (Krongauz, 1995). Such speech formulas, due to frequent use and regular involvement in the processes of language games, lose their identification, if they had one, and are reproduced as an element of cultural information regardless of the author and source (Linguistic Encyclopedic Dictionary, 1990). It is extremely significant that such speech formulas are traditional for a certain society; therefore, they can be used as signs of any culture as a whole, as well as of some special group within it (Nikitina, 1995; Nikolaeva, 1995). In other words, being verbal stereotypes in society, catchphrases play an important role in the social and cultural definition of the community, and their decoding is natural for any carrier of such cultural information.

For the current research, we considered catchphrases from a wide range of sources: the most popular memes (*Eto fiasko, bratan* (lit. *This is fiasco, bro*)); quotes from fiction, movies, songs, and advertising slogans (*Rukopisi ne goryat* (lit. *manuscripts don't burn*), (*Nado, Fedya, nado* (lit. *we must, Fedya, we must*)); phraseological units (*yazyk bez kostei* (lit. *tongue without bones*)); proverbs and idioms (*nogi volka kormyat* (lit. *the wolf's legs feed him*)).

In addition, the catchphrases feature the theme of childhood. This group represents the active cultural vocabulary of a child, mainly of preschool age, which, nevertheless, is also used and reproduced in adulthood. The children's catchphrases include: riddles; counting rhymes (*vyshel mesyats iz tumana* [...] (lit. *the moon came out of the fog* [...])); tongue twisters (*Karl u Klary ukral korally* [...] (lit. *Karl stole corals from Clara* [...])); quotes from fairy tales and children's literature (*Vot kakoi rasseyannyi s ulitsy Basseinoi* (lit. *Such an absent-minded one from Basseynaya street*)); songs; teasers and jokes (*obmanuli duraka na chetyre kulaka* (lit. *fooled a fool for four kulaks*)); proverbs (*s kem povedesh'sya, ot togo i naberesh'sya* (lit. *You will learn from who you hang out with*)).

4 Dataset and Tasks

The cultural types, thematic maps, and catchphrase groups outlined in Sections 2 and 3 formed the foundational framework for constructing a 400-question evaluation dataset. A few examples of the questions are included in Appendix A.

Tasks in the dataset take the form of 5 different question types: multiple-choice with one correct answer (MCQ (1)), multiple-choice with several correct answers (MCQ (N)), gap-filling (GF) with one correct answer, one-to-one matching (M), and one correct answer extraction from a given text (AE). For the distribution of the number of questions and question types across cultural types and catchphrase groups, see Table 1. Details on the number of potential answers for each question type are provided in Appendix B.

Cultural type / catchphrase group	**MCQ (1)**	**MCQ (N)**	**GF**	**Ma**	**AE**
each cultural type (40)	22	5	3	7	3
childhood (30)	8	-	22	-	-
memes (10)	5	-	5	-	-
phraseological units (10)	-	-	10	-	-
proverbs (12)	8	-	4	-	-
quotations (18)	5	1	7	5	-

Table 1: Distribution of number of questions and question types per cultural type and catchphrase group

For the purpose of consistent experimentation, we define a unified interface for all categories (cultural types and catchphrase groups) in the dataset. Each category is represented in JSON format, where each record contains five fields: *id*, *question type*, *prompt*, *question,* and *answer*.

5 Evaluation Setup

5.1 Models

We evaluate several LLMs on their knowledge of the defined cultural types and catchphrase groups. All tested models have multilingual capacity and support the Russian language. Other selection criteria include the support of general questions and high performance on Russian benchmarks[1]. The LLM pool includes both open- and closed-source models and is listed in Table 2.

Model	**No. params**
GPT-4o	-
GPT-3.5 Turbo 16K	-
Claude 3.5 Sonnet	-
Gemini 1.5 Pro	>200B
Gemma 2 27B	27B
Qwen2.5 72B Instruct	72B
Llama 3.1 405B Instruct	405B
Command R+	104B
Mistral NeMo 12B	12B

Table 2: Models for evaluation with their respective number of parameters

The GPT-4o model was included in two versions (released in May and August, 2024) in order to assess the consistency in this model family. We also included GPT-3.5 Turbo to compare its performance against the next generation models. Additionally, the compact Mistral NeMo model was added to evaluate its performance relative to significantly larger counterparts.

[1] Being in the top 10 on llmarena.ru as of 20.11.2024.

5.2 Prompts

LLMs are known to be sensitive to how prompts are formulated (Si et al., 2023; Zhuo et al., 2024). To develop prompts that the selected models understand and respond to in a specified manner, we created a separate set of 20 questions, which contained all 5 types of questions. Each type of question has its own prompt, and each of the resultant prompts consists of four parts: task explanation, information about a number of correct answers, specification of the output format, and a requirement not to explain the reasoning. The latter is needed for the unification of the assessment. All prompts used for the evaluation were in Russian and are included in Appendix A, along with the examples of output formats.

5.3 Metrics

The evaluation dataset consists of various question types, which need to be evaluated differently. We use Accuracy for the MCQ (1) and GF and Exact Match (EM) for the AE questions. For the question types that feature a possibility for partially correct answers — MCQ (N) and Ma — we use 1-Hamming Loss and the Jaccard index, respectively.

The overall evaluation metric is based on the weighted aggregation of individual category metrics by question types. We first compute the weighted average per category based on the number of questions of a certain type in that category. Then the category metrics are averaged, weighted by the total number of questions in the respective category, to compute the overall one. Its modification to evaluate separately on the cultural type and the catchphrase blocks is calculated in a similar manner.

5.4 Evaluation

We evaluate LLMs in a zero-shot setting with task prompting and prompt each model one time. The temperature is set to 0 for all models, with other sampling parameters left to their default values. For the output parsing, we employ a series of heuristic rules to extract an answer, formatted according to the prompt specification. Then, we conduct a manual check, aimed at catching any inconsistencies, and correct the formatting if needed. Lastly, we calculate and report the category and overall metrics.

6 Results

As mentioned in the previous section, we evaluate 10 large language models on their awareness and alignment with Russian culture across 8 cultural types and 5 groups of catchphrases and compute the overall metric to obtain a single-value qualitative measure of their performance. The results, shown in Fig. 1, indicate that the top-scoring models are both versions of the GPT-4o and Claude 3.5 Sonnet, with the lowest score achieved by the smallest of the tested models: Mistral NeMo 12B. The rest of the models (middle-scorers) form two clusters: the ones that achieved almost 80% in the performance measure (Llama 3.1 405B Instruct and Gemini 1.5 Pro) and those in the 65-75% range (Qwen2.5 72B Instruct, Gemma 2 27B, and Command R+).

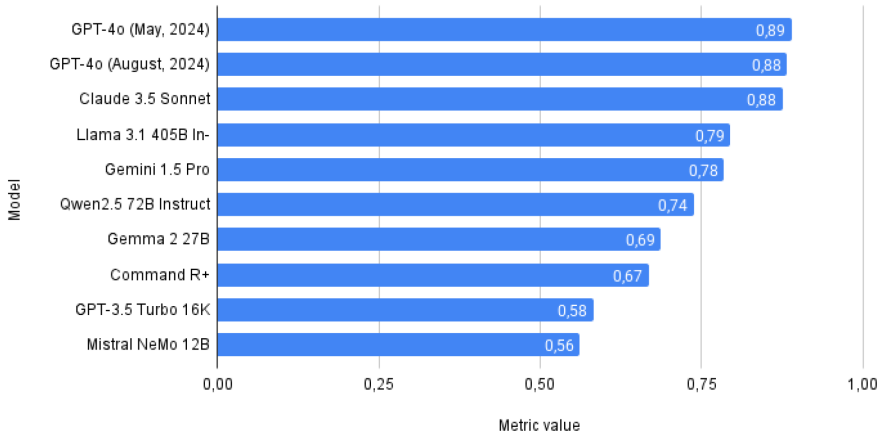

Figure 1: Overall performance of the selected models on the evaluation dataset, sorted in descending order

The performance in respect to the separate evaluation against Cultural Types and catchphrases is shown in Fig. 2. The distribution of the models' performance against the Cultural Type blocks mimics the overall performance since this block contains the larger number of questions. However, all models performed significantly worse against the catchphrases. This divergence likely stems from fundamental differences in task design: Cultural Type questions primarily assess factual knowledge (including foreign cultural influences), while catchphrases evaluate nuanced cultural knowledge that is a product of internalized collective cultural experience.

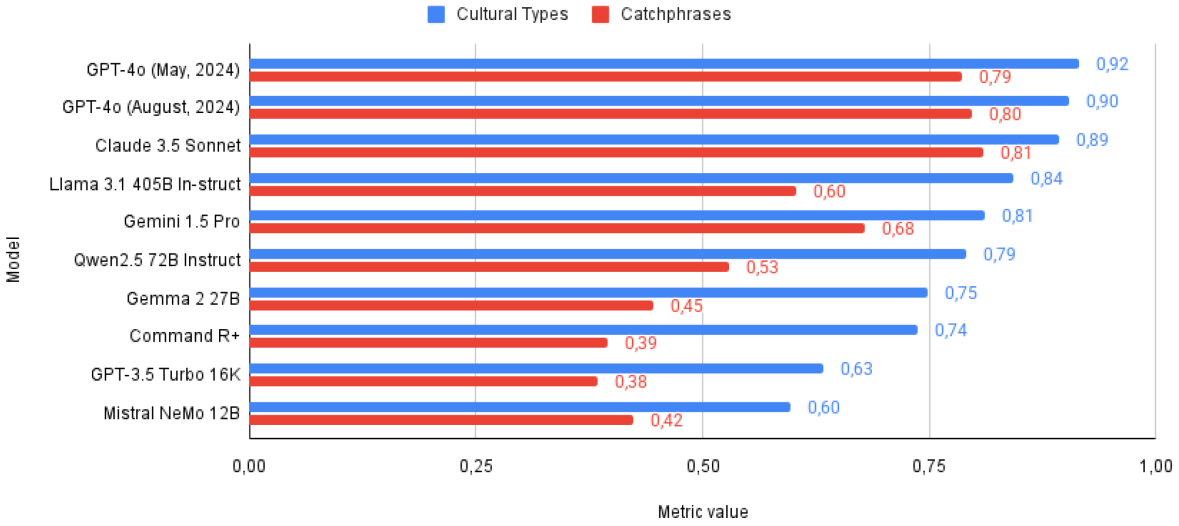

Figure 2: Results of the separate evaluation on the Cultural Type and the catchphrase blocks

The analysis of model performance across individual Cultural Types (Fig. 3) revealed that the top six models demonstrated comparable performance across all types, with minor deviations. However, the bottom four models exhibited significant skews toward specific types. Notably, GPT-3.5 Turbo and Mistral NeMo displayed alignment with the Spiritual Practitioner category, while GPT-3.5 Turbo additionally achieved peak performance in the Nonconformist type.

Evaluation by catchphrase group (Fig. 4) showed greater variability, with no model achieving parity across all categories. GPT-4o and Claude 3.5 Sonnet underperformed most prominently in the Meme group, whereas Llama 3.1 405B Instruct and Gemini 1.5 Pro exhibited weaker results in Quotes. Performance disparities across catchphrase groups were model-specific, indicating individual strengths and limitations. These findings suggest no singular problematic Cultural Type or catchphrase group which challenges the models' performance. Instead, cultural awareness evaluations reveal intrinsic model-specific traits, emphasizing the necessity for developers and users to account for these idiosyncrasies.

Comparing the two versions of the GPT-4o models, there are slight differences in the IT Visionary, Careerist-Achiever, and Nonconformist Cultural Types, with the May version achieving higher scores. Both models performed identically on the catchphrase block, apart from the Childhood category, where the August version made one less mistake. A closer look at the mistakes made revealed that the models almost always make the same ones. This slight difference in performance in some categories could potentially be attributed to the sensitivity to prompts or the ordering of potential answers and needs further investigation. Additionally, both the GPT-4o versions seem to be well-rounded across all categories (with the exception for Memes), while the GPT-3.5 Turbo, a previous generation in this model family, shows an evident skew as mentioned above.

During the qualitative analysis of the errors made by the models, we observed that many models struggled with questions where a correct answer is seemingly illogical, compared to the straightforward factual knowledge (names, places, etc.). For example, identifying a cosmic phenomenon that could negatively affect human life from the astrological perspective, or esoteric ideas of what should be done in order to become wealthier. In both the cultural type and the catchphrase block, another common kind of mistake was concerned with choosing an answer synonymous with the correct one (as in choosing the word *friend* instead of *bratan* in *This is fiasco, bratan*, mentioned in Section 3, or choosing the words

sosedki (eng. *neighbors*) and *devchonki* (eng. *girls*) in the famous quote from Pushkin's 'The Tale of Tsar Saltan' *Tri devitsy pod oknom pryali pozdno vecherkom* (eng. *Three fair maidens, late one night, sat and spun by candlelight.).*

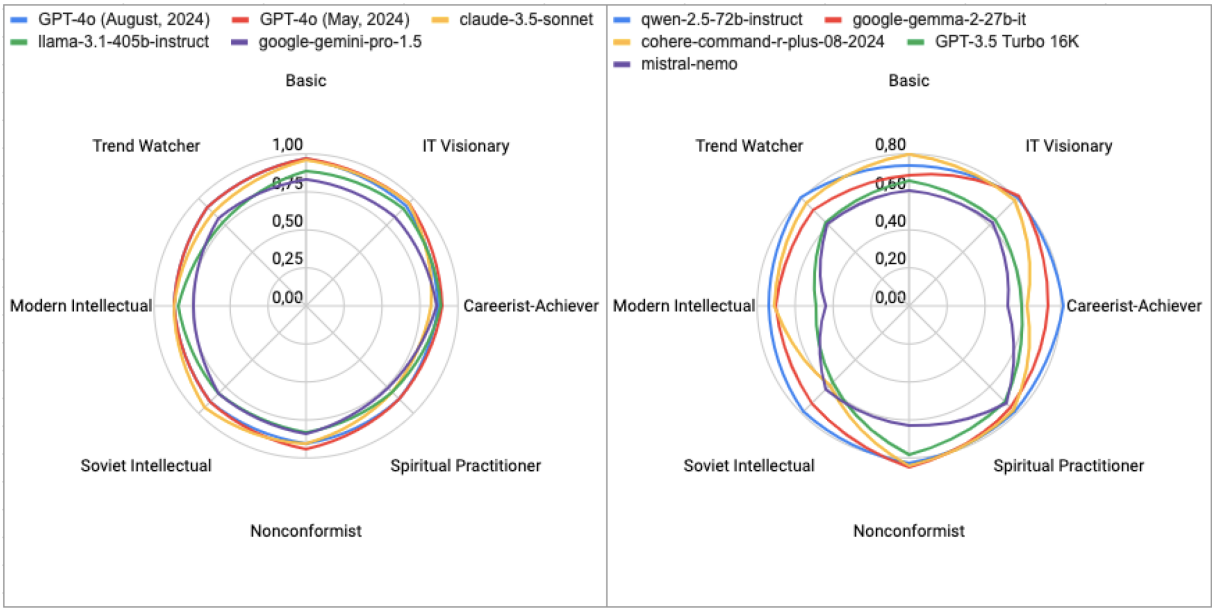

Figure 3: Results of the separate evaluation on individual Cultural Types
(the first five models to the left side of the graph, the last five models — to the right)

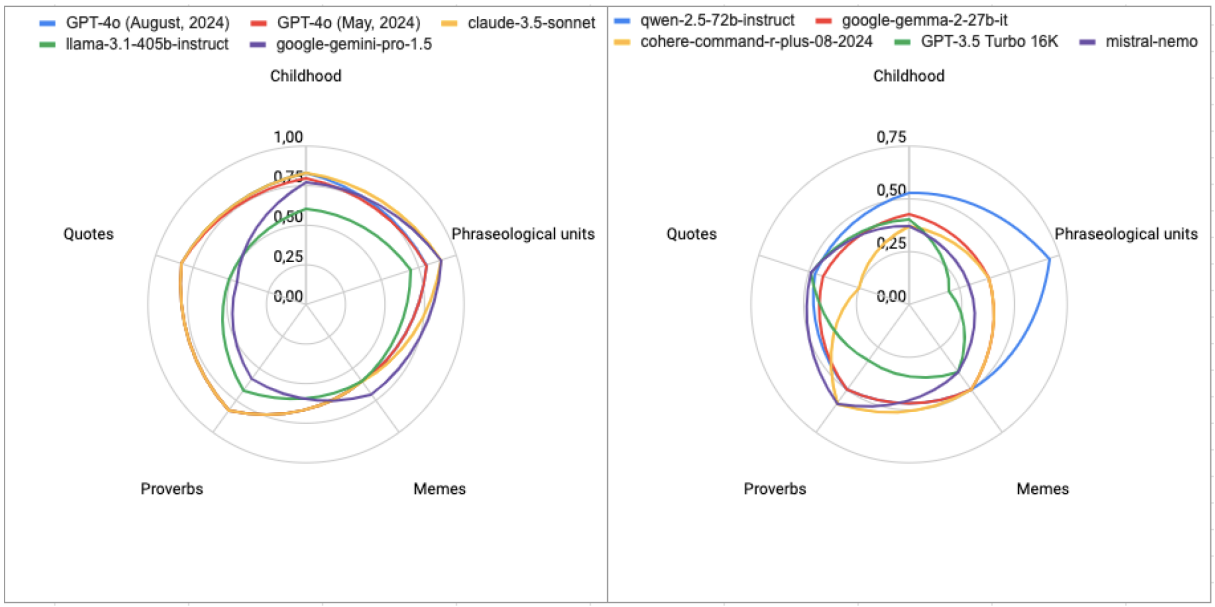

Figure 4: Results of the separate evaluation on individual catchphrase groups
(the first five models to the left side of the graph, the last five models — to the right)

7 Conclusion

This study focuses on the development of the evaluation methodology which allows to assess LLMs for the cultural awareness and alignment to the contemporary Russian sociocultural environment. We developed a structural evaluation framework which operationalizes through 8 Cultural Types and 5 groups of well-known catchphrases. This framework served as basis for the development of an evaluation dataset of 400 questions and probing 10 multilingual LLMs for their ability to understand and reflect cultural nuances.

The results reveal significant variability in performance. Overall top performance was achieved by closed-source models GPT-4o and Claude 3.5 Sonnet, while the lowest scores were achieved by GPT-3.5 Turbo and the smallest of the evaluated models Mistral NeMo 12B. Performance diverged significantly between tasks: models excelled at Cultural Type questions, which often evaluated fact-based knowledge, but underperformed on catchphrase evaluations, reflecting gaps in nuanced linguacultural knowledge.

Analysis revealed model-specific biases: lower-ranked models skewed toward specific cultural types (e.g., GPT-3.5 Turbo favored Spiritual Practitioner and Nonconformist), while catchphrase performance varied idiosyncratically — GPT-4o and Claude 3.5 Sonnet struggled with Memes, whereas others faltered on Quotes. GPT-4o versions showed minor differences in Cultural Type scores but identical catchphrase performance, suggesting prompt sensitivity. Error analysis highlighted two failure patterns: (1) difficulty resolving culturally illogical but correct answers, and (2) synonym confusion. These findings underscore the need for further research into improving LLMs' cultural alignment. Future work could expand the dataset, refine evaluation methodologies, and explore ways to enhance models' cultural awareness.

References

[1] Ahmad I., Dudy S. et al. Are Generative Language Models Multicultural? A Study on Hausa Culture and Emotions using ChatGPT // Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP. — Bangkok, Thailand, 2024. — P. 98–106.

[2] AlKhamissi B., ElNokrashy M. et al. Investigating Cultural Alignment of Large Language Models // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. — Bangkok, Thailand, 2024. — P. 12404–12422.

[3] Arora A., Kaffee L., Augenstein I. Probing Pre-Trained Language Models for Cross-Cultural Differences in Values // Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP). — Dubrovnik, Croatia, 2023. — P. 114–130.

[4] Cao Y., Zhou L. et al. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study // Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP). — Dubrovnik, Croatia, 2023. — P. 53–67.

[5] Eckert P. Communities of practice // Concise encyclopedia of pragmatics, 2nd edition. Oxford: Elsevier, 2006. — P. 109–112.

[6] Fenogenova A., Chervyakov A. et al. MERA: A Comprehensive LLM Evaluation in Russian // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. — Bangkok, Thailand, 2024. — P. 9920–9948.

[7] Fung Y., Chakrabarty T. et al. NORMSAGE: Multi-Lingual Multi-Cultural Norm Discovery from Conversations On-the-Fly // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. — Singapore, 2023. — P. 15217–15230.

[8] Hofstede G. Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations. —Thousand Oaks, CA: Sage Publications, 2001.

[9] Karinshak E., Hu A. et al. LLM-GLOBE: A Benchmark Evaluating the Cultural Values Embedded in LLM Output // Computing Research Repository. — 2024. — Vol. arXiv:2411.06032. — version 1. Access mode: https://arxiv.org/abs/2411.06032.

[10] Kim E., Suk J. et al. CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). — Torino, Italia, 2024. — P. 3335–3346.

[11] Koto F., Mahendra R. et al. IndoCulture: Exploring Geographically Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces // Transactions of the Association for Computational Linguistics. — Vol. 12, 2024. — P. 1703–1719.

[12] Krongauz M. A. The energy of cliched forms [Energiya klishirovannykh form], Speech and mental stereotypes in synchrony and diachrony. Conference abstracts [Rechevye i mental`nye stereotipy v sinkhronii i diakhronii. Tezisy konferencii.]. — Moscow, Russia: Institute of Slavic and Balkan Studies, Russian Academy of Sciences [Institut slavyanovedeniya i balkanistiki RAN], 1995. — P. 50–52.

[13] Lave J., Wenger E. Situated learning: Legitimate peripheral participation. — Cambridge: Cambridge university press, 1991. — P.45–52.

[14] Linguistic Encyclopedic Dictionary [Lingvisticheskij entsiklopedicheskij slovar'], Yartseva V.N. (editor), — Moscow, Russia:Soviet Encyclopedia [Sovetckaya entsiklopediya], 1990. — P. 683.

[15] Lotman Yu.M. Culture and Explosion [Kul`tura i vzryv], Semiosphere [Semiosfera]. — Saint Petersburg, Russia: Iskusstvo, 2000. — P. 12–149.

[16] Lutovinova O.V. "Linguocultural type" in a series of related concepts used to study linguistic personality [«Lingvokul`turnyj tipazh» v ryadu smezhnykh ponyatij, ispol`zuemykh dlya issledovaniya yazykovoj lichnosti], Scientific notes of the Transbaikal State University. Series: Philology, history, oriental studies [Uchenye zapiski Zabajkal`skogo gosudarstvennogo universiteta. Seriya: Filologiya, istoriya, vostokovedenie]. — Vol. 3, 2009. — P. 225–228.
[17] Myung J., Lee N. et al. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages // Computing Research Repository. — 2024. — Vol. arXiv:2406.09948. — version 1. Access mode: https://arxiv.org/abs/2406.09948.
[18] Naous T., Ryan M. J. et al. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. — Bangkok, Thailand, 2024. — P. 16366–16393.
[19] Nikolaeva T.M. The swing of freedom/unfreedom: tragedy or salvation? [Kacheli svobody`/nesvobody`: tragediya ili spasenie?], Speech and mental stereotypes in synchrony and diachrony. Conference abstracts [Rechevye i mental`nye stereotipy v sinkhronii i diakhronii. Tezisy konferencii.]. — Moscow, Russia: Institute of Slavic and Balkan Studies, Russian Academy of Sciences [Institut slavyanovedeniya i balkanistiki RAN], 1995. — P. 83–88.
[20] Nikitina S.E. Stereotypical judgments and speech cliches as cultural barriers [Stereotipnye suzhdeniya i rechevye klishe kak kul`turnye bar`ery], Speech and mental stereotypes in synchrony and diachrony. Conference abstracts [Rechevye i mental`nye stereotipy v sinkhronii i diakhronii. Tezisy konferencii.]. — Moscow, Russia: Institute of Slavic and Balkan Studies, Russian Academy of Sciences [Institut slavyanovedeniya i balkanistiki RAN], 1995. — P. 81–83.
[21] Rao A., Yerukola A. et al. Normad: A benchmark for measuring the cultural adaptability of large language models // Computing Research Repository. — 2024. — Vol. arXiv:2404.12464. — version 7. Access mode: https://arxiv.org/abs/2404.12464.
[22] Shavrina T., Fenogenova A. et al. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Online, 2020. — P. 4717–4726.
[23] Shen S., Logeswaran L. et al. Understanding the Capabilities and Limitations of Large Language Models for Cultural Commonsense // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Mexico City, Mexico, 2024. — P. 5668–5680.
[24] Si C., Gan Z. et al. Prompting GPT-3 To Be Reliable // Computing Research Repository. — 2023. — Vol. arXiv:2210.09150. — version 2. Access mode: https://arxiv.org/abs/2210.09150.
[25] Son G., Lee H. et al. HAE-RAE Bench: Evaluation of Korean Knowledge in Language Models // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). — Torino, Italia, 2024. — P. 7993–8007.
[26] Taktasheva E., Fenogenova A. et al. TAPE: Assessing Few-shot Russian Language Understanding // Findings of the Association for Computational Linguistics: EMNLP 2022. — Abu Dhabi, United Arab Emirates, 2022. — P. 2472–2497.
[27] Tönnies F. Community and Civil Society. — Cambridge: Cambridge University Press, 2001. — P. 52–93.
[28] Wang Y., Zhu Y. et al. CDEval: A Benchmark for Measuring the Cultural Dimensions of Large Language Models // Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP. — Bangkok, Thailand, 2024. — P. 1–16.
[29] Yao B., Jiang M. et al. Benchmarking Machine Translation with Cultural Awareness. // Findings of the Association for Computational Linguistics: EMNLP 2024. — Miami, Florida, USA, 2024. — P. 13078–13096.
[30] Yin D., Bansal H. et al. GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. — Abu Dhabi, United Arab Emirates, 2022. — P. 2039–2055.
[31] Zhao W., Mondal D. et al. WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). — Torino, Italia, 2024. — P. 17696–17706.
[32] Zhuo J., Zhang S. et al. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs // Findings of the Association for Computational Linguistics: EMNLP 2024. — Miami, Florida, USA, 2024. — P. 1950–1976.
[33] Zubarevich N.V. Regions of Russia: Inequality, Crisis, Modernization [Regiony Rossii: neravenstvo, krizis, modernizaciya]. — Moscow, Russia: Independent Institute for Social Policy [Nezavi-simy`j institut social`noj politiki], 2010.

Appendix A. Examples of Questions in Cultural Types

Basic Type

(1) *На фантике какой конфеты изображена картина Ивана Шишкина «Утро в сосновом бору»? А) Белочка Б) Гулливер В) Мишка на Севере Г) Кара-Кум Д) Красная Шапочка Е) Мишка косолапый.*
On the wrapper of which candy is Ivan Shishkin's painting "Morning in a Pine Forest" depicted? A) Belochka B) Gulliver C) Mishka na severe D) Kara-Kum E) Krasnaya Shapochka E) Mishka Kosolapyj.

Nonconformist

(2) *Дополните фразу: «Пепси, пейджер, _» А) ВГТРК Б) вечеринка В) MTV Г) молодость Д) кока-кола Е) смартфон.*
Complete the phrase: "Pepsi, pager, _" A) VGTRK B) party C) MTV D) youth E) Coca-Cola E) smartphone.

Trend Watcher

(3) *Выберите резидентов Comedy Club: А) Павел Воля Б) Гарик Харламов В) Сергей Орлов Г) Семен Слепаков.*
Select Comedy Club residents: A) Pavel Volya B) Garik Kharlamov C) Sergey Orlov D) Semyon Slepakov.

Appendix B. Question Types and Prompts

Question type	Number answers	Output format	Prompt (Rus)	Prompt (translated to English)
Multiple choice (1)	1 correct out of 6	Letter: A	На вход подаются инструкции для ответа на вопрос. В них будет несколько вариантов ответа. Определите верный ответ. В ответе укажите только одну букву правильного ответа. Объяснять свой выбор не нужно.	The input contains instructions for answering a question. They will contain several answer options. Determine the correct answer. In the answer, indicate only one letter of the correct answer. There is no need to explain your choice.
Multiple choice (N)	N correct out of 4 or 6	One or more letters in alphabetical order: AB	На вход подаются инструкции для ответа на вопрос. В них будет несколько вариантов ответа. Определите верный ответ. Учтите, что их может быть несколько. В ответе укажите только букву ответа или последовательность из букв ответа в алфавитном порядке без пробела. Объяснять свой выбор не нужно	The input contains instructions for answering a question. They will contain several answer options. Determine the correct answer. Note that there may be several. In the answer, indicate only the letter of the answer or a sequence of letters of the answer in alphabetical order without a space. There is no need to explain your choice.
Gap-filling	1 correct out of 6	Letter: A	На вход подаются вопросы, в которых нужно заполнить пропуски. Пропуски обозначены знаком "_". Заполните пропуск предложенными вариантами. В ответе укажите только одну букву правильного ответа. Объяснять свой выбор не нужно.	The input contains a question, in which you need to fill in the gaps. Gaps are marked with the sign "_". Fill in the gap with the suggested options. In the answer, indicate only one letter of the correct answer. You do not need to explain your choice.
One-to-one matching	2 lists of 4 elements each	Sequence of numbers and letters: 1B 2V 3A 4G	На вход подаются инструкции для ответа на вопросы. Для каждого элемента из первого списка выбери наиболее подходящий элемент из второго списка. Буквы не должны повторяться. Нужно постараться составить 4 пары. В ответе укажите только последовательность из цифр-букв этих списков через пробел в формате 1Б 2А 3Г 4В.	The input contains instructions for answering questions. For each element from the first list, choose the most suitable element from the second list. The letters should not be repeated. You need to try to make 4 pairs. In your answer, indicate only the sequence of numbers and letters from these lists separated by a space in the format 1B 2A 3D 4C.
Answer extraction	1 correct out of 6	A noun the in nominative case	На вход подаются инструкции для ответа на вопрос. Найдите в тексте правильный ответ. Укажите ответ в именительном падеже без знаков препинания. Объяснять свой выбор не нужно.	The input contains instructions for answering a question. Find the correct answer in the text. Provide the answer in the nominative case without punctuation. There is no need to explain your choice.

Table 3: Question types used in the evaluation dataset, number of correct and potential answers, along with prompts and output format