RuSimulBench a Benchmark for Assessing the Quality of Language Models in Russian

Muslimbek Abdurakhimov
ITMO University
St Petersburg, Russia
abdurahimov.muslimbek@gmail.com

Maria Khodorchenko
ITMO University
St Petersburg, Russia
mariyaxod@yandex.ru

***Abstract*—The rapid advancement of Large Language Models (LLMs) has significantly improved natural language processing capabilities across multiple languages. However, evaluating their performance in languages with limited benchmarking resources, such as Russian, remains a challenge. In this work, we introduce a novel benchmarking framework designed to assess Russian LLMs along two critical dimensions: creativity and stability. Creativity is essential for generating diverse, original, and contextually appropriate responses, while stability ensures that models provide consistent outputs when faced with slight prompt variations.**

Our benchmark makes use of a unique and comprehensive approach, combining automated and human evaluated systems. We propose the stability coefficient for measuring the value of novel model's prompts rewording and demonstrate the creative score, which claims value based on originality, diversity and coherence in reactions of silenced models. For thoroughness, we include several LLM architectures that were adapted to Russian language to be assessed against structured test cases derived from the MERA benchmark.

The experimental results provide deep insights into the trade-offs between stability and creativity in Russian language models, highlighting strengths and areas for improvement in existing architectures. By offering a standardized evaluation approach, our benchmark contributes to the development of more reliable and effective Russian-language AI systems. Additionally, our findings can inform future research on enhancing LLM adaptability and robustness in low-resource languages.

I. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating human-like text, performing complex reasoning tasks, and adapting to various domains. However, assessing their quality remains a significant challenge, particularly for languages with limited benchmarking resources, such as Russian [12]. Although there are extensive evaluation frameworks for English, there is a lack of standardized methods to measure the performance of Russian language models in different linguistic and cognitive dimensions.

In this study, we introduce a benchmark designed to evaluate the quality of Russian LLMs, focusing on two critical aspects: creativity and stability. Creativity is essential for tasks that require diverse, original, and contextually appropriate responses, while stability ensures that a model maintains consistency when given slight variations of the same prompt. The ability to balance these qualities is crucial for the development of reliable and effective AI systems, particularly in applications involving content generation, dialogue systems, and decision support.

Our benchmark employs a systematic evaluation framework that includes task-specific test cases, automated and human-assessed metrics, and comparative analysis across different model architectures. By analyzing creativity and stability simultaneously, we aim to provide deeper insights into how Russian LLMs generate and process language. This work contributes to the broader goal of improving AI transparency and usability, offering a standardized tool for researchers and developers working with Russian-language models.

The contribution of our work can be summarized as follows:

- Novel benchmark for Russian LLMs: Unlike existing evaluations that focus on general language understanding, our framework is specifically designed to assess both creativity and stability in Russian LLMs, filling a critical gap in benchmarking resources.
- Introduction of a Stability Coefficient: We introduce a novel quantitative metric to measure how consistently Russian LLMs respond to slight variations in input prompts, providing a new way to evaluate model robustness.
- Comprehensive Analysis of Russian LLM performance: Our study conducts an extensive evaluation across multiple LLM architectures adapted for Russian, offering comparative insights that were previously unavailable in the field.

II. Literature review

In recent years, the evaluation of Large Language Models (LLMs) has garnered significant attention, particularly regarding their performance in non-English languages such as Russian [1], [2]. Although there are extensive benchmarks for English LLMs, for Russian language has seen comparatively fewer comprehensive evaluation frameworks. For evaluating LLMs on Russian language there has been done several works. For example, The MERA benchmark [3] is a notable initiative aimed at evaluating Russian-language LLMs. MERA encompasses 21 tasks across 10 skill domains, providing a structured platform for assessing generative models in Russian. It employs a fixed experimental pipeline to ensure consistency and reliability in evaluations, addressing the need for standardized assessment tools in the Russian context. Another

significant contribution is the RussianSuperGLUE benchmark [4], designed to test general language understanding in Russian. Modeled after the SuperGLUE benchmark for English, RussianSuperGLUE includes nine tasks that evaluate various linguistic and cognitive abilities, such as natural language inference and common sense reasoning. This benchmark provides baselines and human-level evaluations, serving as a valuable resource for researchers developing Russian-language models [5].

However, there is a gap in the evaluation of the assessment of creativity and stability of LLM [6]. Evaluation of creativity in LLM presents unique challenges, as creativity is a multifaceted and often subjective quality [7], [8]. Currently growing study of creativity of LLMs and AI is approacing the automation of the evaluation creativity solutions [9]. Recent studies have sought to establish frameworks for assessing creativity in LLMs [1] [4]. One approach adapts the modified Torrance Tests of Creative Thinking to evaluate models across tasks emphasizing fluency, fl exibility, or iginality, an d elaboration [10] [11]. This methodology includes the development of comprehensive datasets and LLM-based evaluation methods to quantify creative performance [12]. Another study investigates the creative thinking of LLMs through the Divergent Association Task (DAT) [6], an objective measure that asks models to generate unrelated words and calculates the semantic distance between them [13]. Findings suggest that advanced LLMs, such as GPT-4, can exhibit divergent semantic associations, a fundamental aspect of creativity [14]. Also, SimulBench, a benchmark designed to evaluate creativity in large language models (LLMs), focuses on open-ended tasks that require models to generate imaginative, coherent, and contextually appropriate outputs [15]. These tasks often involve storytelling, hypothetical scenario generation, and creative problem-solving, providing a framework to assess models' abilities to exhibit human-like creativity. SimulBench also incorporates evaluation metrics such as fluency, c oherence, a nd novelty, which are central to understanding the creative capabilities of LLMs [16]. Most importantly, balancing creativity and stability in the evaluation of LLMs is a complex undertaking. While creativity involves generating novel and diverse outputs, stability requires consistent and reliable responses. A mathematical abstraction has been proposed to balance this trade-off, suggesting that models can be trained on specific loss functions that account for both creativity and reality. This approach aims to fine-tune LLMs to achieve an optimal balance between generating innovative content and maintaining factual accuracy [17].

III. RUSIMULBENCH DESCRIPTION

In this section, we introduce our new evaluation benchmark. RuSimulBench is a benchmark designed to evaluate Russian language models across two key aspects: stability and creativity. It provides a structured framework for testing how well models handle slight variations in prompts while maintaining consistency (stability) and how effectively they generate diverse and original responses (creativity). Using standardized test cases and evaluation metrics, RuSimulBench aims to offer a comprehensive assessment of Russian LLMs.

A. Tasks

1) Creativity Tasks: The evaluation of language model outputs, particularly in creative tasks and especially for the Russian language with complex grammar and a rich cultural heritage, presents unique challenges that traditional metrics often do not adequately address.

Developing the part of the benchmark for creativity has created the systematic approach. The initial step involved studying the structure, methodology, and evaluation metrics of SimulBench to understand its core principles. This included analyzing the datasets, tasks, and metrics utilized for assessing creativity in English-language models. Particular attention was given to identifying components specific to English, such as linguistic features and cultural references, and preparing corresponding Russian equivalents that reflect the nuances of the Russian language.

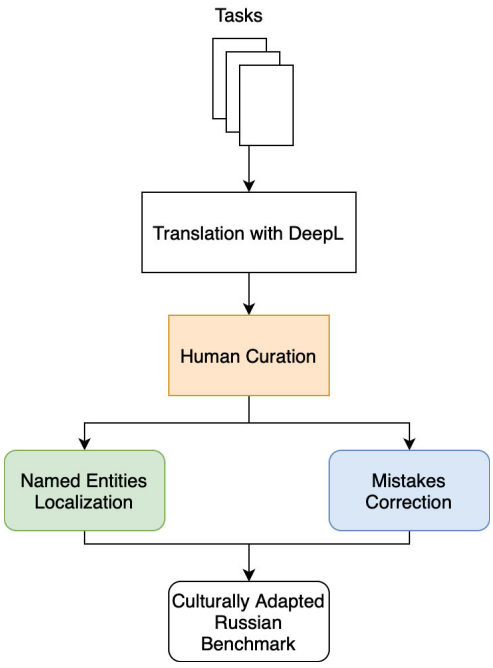

Fig. 1. Process of the rusification of prompts into Russian

To ensure cultural and linguistic relevance, existing datasets and prompts were translated and adapted for Russian. You can see it from the fig. 1. DeepL was used for initial translation, as it performs better in managing phraseological and contextual differences [17], which is particularly important for evaluating the Russian language. However, the replacement of culturally specific elements, such as references to Western literature or historical figures, with appropriate Russian counterparts was carried out separately. Each translation and adaptation was validated by native Russian speakers to ensure accuracy and maintain the integrity of the tasks. The prompts selected for russification covered a variety of creative domains, including

travel guides, poetry, journalism, and text-based adventure games, ensuring a comprehensive evaluation of creativity.

In the next Fig. 2, the process of cross-lingual prompt adap-tation for specialized task instructions is shown, specifically focusing on English to Russian translation examples.

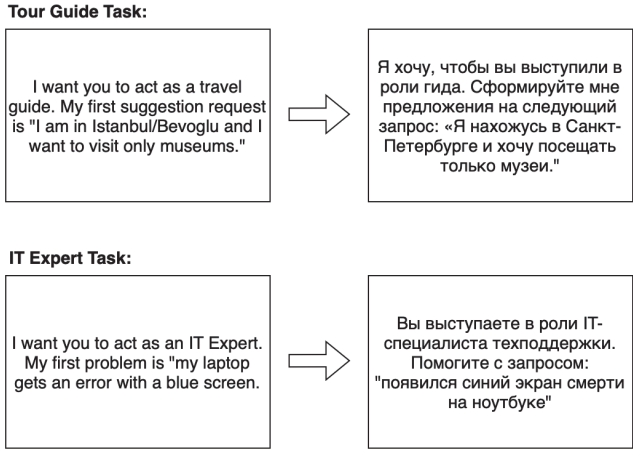

Fig. 2. An example of prompts after adaptation into Russian

Two distinct scenarios are presented, each illustrating how role-specific prompts can be effectively localized while maintaining their functional purpose. In the first example, labeled "Tour Guide task," the English prompt "I am in Istanbul/Beyoğlu and I want to visit only museums". Notably, the localization involves not just linguistic translation but also cultural adaptation, replacing Istanbul with Saint Petersburg to maintain relevance for the target audience while preserving the core intent of the museum-focused tourism request. The second example, labeled "IT Expert Task," demonstrates the adaptation of technical support prompting. The English prompt "my laptop gets an error with a blue screen". This translation is particularly interesting as it employs the colloquial Russian term blue screen of death, showing how technical terminology is locally adapted while maintaining the essential meaning of the computer error scenario.

2) Stability Tasks: Stability evaluation measures how consistently a model responds when given slightly different versions of the same prompt. A stable model should provide logically equivalent answers regardless of minor changes in wording. This is tested by rephrasing prompts, introducing synonyms, or making small modifications while checking if the model maintains consistency in meaning and reasoning.

We have used MERA benchmark tasks for evaluation of the stability of the LLMs.

CheGeKa is an open-domain question-answering dataset in Russian, consisting of QA pairs collected from the official Russian quiz database ChGK. It involves open-ended questions and is evaluated using the F1 score and Exact Match (EM) metrics. This task assesses a model's ability to provide accurate and complete answers to questions that require general knowledge about the world [1].

LCS challenges language models to find the longest common subsequence between pairs of input strings. As a classic dynamic programming problem, LCS tests a model's ability to identify and apply efficient algorithmic approaches. The model's performance is evaluated based on accurately predicting the length of the longest subsequence shared by the two strings [1].

ruDetox is a diagnostic dataset for evaluating models' ability to detoxify offensive Russian language while preserving meaning and fluency. The task involves transforming input sentences containing toxic or abusive content into more neutral and polite rephrasing's. Models are assessed on their detoxification effectiveness, semantic preservation, and output fluency [1].

ruOpenBookQA is a Russian multiple-choice question-answering dataset consisting of elementary-level science questions. The questions test understanding and application of core scientific facts, as well as related common-sense reasoning. To simulate an open-book exam, models are provided with relevant textual resources and must select the correct answer from several choices [1].

B. Evaluation

RuSimulBench employs both automatic and human evaluation metrics. Stability is measured using consistency metrics such as cosine similarity and logical equivalence scoring.

1) Stability Coefficient: We presented the proposed methodology to test the hypothesis of the stability coefficient of language models to prompt changes and evaluated its potential implications for practical use and understanding of the issue. To do this, we conducted a series of experiments on tasks from the MERA benchmark.

We selected tasks from the MERA benchmark typically 4 tasks (Chegeka, LCS, ruDetox, ruOpenBookQA). Sampling of task data sets to collect approximately 10 examples for initial hypothesis testing. After that generation of 10 variations of task texts for each selected task, ranging from brief to extensive. To evaluate model performance, we used the Vikhr and TinyLlama, and etc. models with a temperature of 0 to eliminate random responses and focus on their generation results on different prompts. After that, we calculated the stability coefficient.

For evaluation of stability, which was our first work we used following formula:

$$S = \left(\frac{\sum_{i \neq j} \text{similarity}_{ij} - n}{n \cdot (n-1)} \right) \cdot P \tag{1}$$

where:

- similarity_{ij} represents the cosine similarity between the embeddings of response R_i and response R_j.
- n denotes the total number of prompt variations per question ($n = 10$ in our experiments).
- P is the answer probability, calculated based on token probabilities in the generated response.

2) Creativity Evaluation: For evaluation of creativity we employed Google's Gemini Flash model as an evaluator for assessing the creative outputs of various Russian language models. This choice was motivated by several key considerations and implemented with specific methodological safeguards. Gemini Flash was selected as the evaluator due to its demonstrated capabilities in cross-lingual understanding and its ability to apply consistent evaluation criteria across diverse creative outputs. The model's strong performance in both English and Russian language tasks made it particularly suitable for our evaluation framework, where cultural and linguistic nuances play crucial roles [18].

To ensure robustness and address potential limitations of automated evaluation, we supplemented Gemini's assessments with manual human evaluation. For this, selected outputs from Russian language models were translated into English and then independently reviewed by bilingual annotators with backgrounds in linguistics and creative writing. The human raters were tasked with scoring creativity along similar dimensions (originality, coherence, relevance) used by the automated system, allowing us to cross-check alignment between human judgment and model-based scores. While the primary benchmark results rely on automated metrics, this human evaluation served as a validation layer.

The evaluation process was guided by the following structured prompt (translated to English):

> *Evaluation Prompt: Rate the following response on a scale from 0 to 10 with detailed justification:*
> *Original request:* {*original_prompt*} *Response:* {*response*}
> *Evaluation criteria:*
> *1. Creativity: How unique and original is the response?*
> - *0-3: Low quality (template-based, unoriginal response, lacks creative approach)*
> - *4-6: Medium quality (partially original response with minimal creativity)*
> - *7-10: High quality (response is unique, contains non-standard ideas and creative approach)*
>
> *2. Diversity: Are different linguistic means and stylistic devices used?*
> - *0-3: Low quality (monotonous style, lack of variations in linguistic means)*
> - *4-6: Medium quality (some diversity present, but in limited volume)*
> - *7-10: High quality (wide range of linguistic means used, diversity in style and presentation)*
>
> *3. Relevance: How accurately does the response correspond to the original request?*
> - *0-3: Low quality (response is not related or weakly corresponds to the request)*
> - *4-6: Medium quality (response generally corresponds to the request, but contains inaccuracies)*
> - *7-10: High quality (response fully corresponds to the request, covers all its aspects)*
>
> *Requirements for your answer:*
> - *Provide a numerical score for each criterion (on a scale from 0 to 10)*
> - *Explain your evaluation for each criterion in detail, including specific examples from the text*
> - *Suggest possible improvements to enhance the quality of the response.*

Which means, Creativity measures the uniqueness and originality of a response. A score between 0 and 3 indicates a generic and unoriginal answer with little to no creative elements. Responses scoring 4 to 6 demonstrate some originality but remain relatively conventional. High-scoring responses (7 to 10) exhibit significant creativity, incorporating unique ideas and unconventional approaches.

Diversity evaluates the richness of language and stylistic variation. Responses rated 0 to 3 are monotonous and lack variation in word choice or sentence structure. A score of 4 to 6 suggests moderate diversity, with some stylistic variation but limited complexity. The highest scores (7 to 10) are awarded to responses that showcase a broad range of linguistic techniques, diverse vocabulary, and stylistic creativity.

Relevance assesses how well the response aligns with the given prompt. Low-scoring responses (0 to 3) are largely irrelevant or only weakly related to the original query. Those scoring 4 to 6 are somewhat relevant but may contain inaccuracies or fail to fully address the prompt. Responses rated 7 to 10 are highly relevant, accurately answering the query while maintaining logical coherence and completeness.

To ensure objective and meaningful evaluation, reviewers were required to assign a numerical score for each criterion, justify their assessment with specific examples from the text, and suggest potential improvements for enhancing response quality.

To quantitatively assess the creative abilities of models, we introduce the **Creative Score**, a weighted combination of three fundamental components: *Creativity*, *Diversity*, and *Coherence*. For evaluation we used following formula:

$$\text{CS} = \alpha \cdot \text{Creativity} + \beta \cdot \text{Diversity} + \gamma \cdot \text{Coherence} \quad (2)$$

where:

- α, β, γ are weighting coefficients such that $\alpha+\beta+\gamma = 1.0$.
- Each component score (Creativity, Diversity, Coherence) is normalized to a common scale of [0,10].

Component Definitions: Each component is further broken down into measurable sub-components:

- **Creativity** = originality, novelty, unexpected connections
- **Diversity** = lexical variety, stylistic range, narrative approaches
- **Coherence** = logical flow, linguistic correctness, narrative consistency

The weighting factors (α, β, γ) were set as follows to align with Russian literary and cultural norms, ensuring that the evaluation respects linguistic nuances and traditional storytelling structures:

- α = 0.4 (Creativity was prioritized as a key aspect of originality and novelty).
- β = 0.3 (Diversity was weighted to capture stylistic range and linguistic richness).
- γ = 0.3 (Coherence ensured logical consistency and readability of responses).

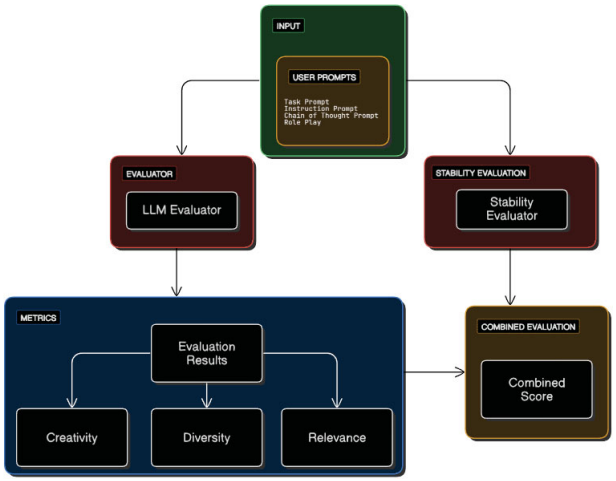

Fig. 3. Process of the Evaluation of Metrics

In the Fig. 3 we can see the experimental process of our project. The experimental framework was designed to evaluate the creative capabilities of various Russian-language models in handling various types of prompts. Our evaluation uses Gemini through its API interface with the prepared prompts designed and evaluates 3 main metrics: Creativity, Diversity, and Relevancy. Creativity score measures the originality and inventiveness of the generated content, with particular attention to novel combinations of ideas and unexpected but coher-ent narrative elements. The coherance score, on the other hand, evaluates the logical flow and structural integrity of the outputs, including the assessment of narrative consistency and adherence to Russian linguistic conventions. Finally, the Diversity Score, checks the range and variety of creative expressions, including lexical diversity and the use of different stylistic devices common in Russian literature.

3) Combined Evaluation: To further refine our evaluation methodology, we propose a Combined Stability-Creativity Score that integrates stability assessment with a creativity constraint. This ensures that while responses remain diverse, they also adhere to creative expectations. For this, specificially we use temperature at 0.6 or higher, and check if the answers meet the criterias of Diversity, Instruction following and Creativity Retention. This adaptation ensures that the model is not only stable under controlled prompt perturbations but also capable of generating diverse, yet high-quality and creative responses.

To provide a holistic assessment, we introduce the Combined Evaluation Score (CES), which is calculated as the averaged sum of the Creativity Score (normalized to a [0,1] scale) and the Stability Coefficient:

$$\text{CES} = \frac{\text{Normalized Creativity Score} + \text{Stability Coefficient}}{2} \tag{3}$$

This approach ensures that models are not only creatively expressive but also consistently reliable across prompt variations.

IV. Results

A. Benchmarking Settings

To ensure a fair and reproducible evaluation, we standardized the benchmarking conditions across all tested models. This section outlines the selected models, generation parameters, and trial settings applied during the assessment.

For stability evaluation, we selected four tasks from the MERA benchmark: Chegeka, LCS, ruDetox, and ruOpenBookQA. Each task consisted of ten variations of prompt texts, ranging from brief to extensive, to analyze the stability of language models when faced with different formulations of the same question. The generation setting was configured with a temperature of 0 to minimize randomness and focus on model consistency in responses.

For both stability and creativity assessments, we benchmarked the following language models:

- Vikhr-Nemo-12B[1] – Vikhr 12B, fine-tuned using NVIDIA's NeMo framework.
- Saiga-LLaMA3-8B[2] – -
- Mistral-Nemo[3] – based on Mistral 7B, fine-tuned by NVIDIA's NeMo framework.
- Qwen2.5-7B[4], which is Qwen 2.5 series, a 7B-parameter model from Alibaba.
- TinyLLaMA-1.1B[5] – A lightweight TinyLLaMA model with 1.1B parameters.

To ensure a comprehensive evaluation of language model stability and creativity, we selected a diverse set of open-source models with varying architectures and parameter sizes. The primary focus was on models that demonstrated strong performance in multilingual tasks, including Russian, given the research context.

However, we acknowledge that our selection does not include proprietary models such as OpenAI's GPT-4 (used in ChatGPT) or Google's Gemini, which could have provided additional insights. The exclusion of these models was due to several factors, including accessibility constraints, reproducibility concerns, and the need for transparency in benchmarking. Future work could explore a broader comparison by incorporating these proprietary models to further validate the hypotheses and provide a more comprehensive assessment of model performance across different linguistic and creative benchmarks.

[1]https://huggingface.co/Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24
[2]https://huggingface.co/IlyaGusev/saiga_llama3_8b
[3]https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407
[4]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[5]https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0

Each model underwent ten independent trials per task to ensure statistical robustness in our evaluation. For stability, the similarity scores between generated responses were computed using cosine similarity, and the stability coefficient was calculated accordingly.

For creativity evaluation, we employed Google's Gemini Flash model as an evaluator. The evaluation procedure involved generating three independent assessments per task to account for variations in Gemini's interpretation. The generated outputs were analyzed based on the Creative Score formula, incorporating three key factors: Creativity, Diversity, and Coherence. To balance stability and creativity, the generation setting for creativity evaluation was configured with a temperature of 0.4, allowing for some degree of variability while maintaining structured responses.

These standardized benchmarking settings enabled a consistent and comparable analysis of model performance across different linguistic tasks, ensuring reliability in both stability and creativity assessments.

B. Comparison between models

Creativity evaluation assesses the model's ability to generate diverse and innovative responses. This is done by presenting open-ended prompts that encourage original thinking, such as storytelling, idea generation, or problem-solving tasks. The model's responses are then analyzed for uniqueness, fluency, and contextual relevance, helping to determine its creative capabilities.

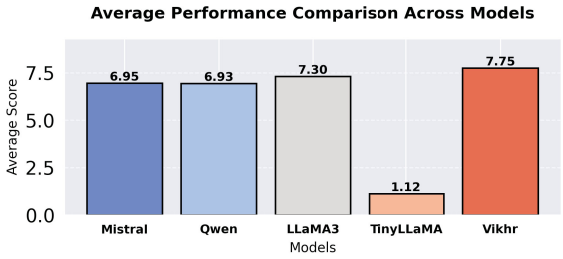

Fig. 4. Average Performance Comparison Across Models

The Fig. 4 presents the average performance comparison across all models, revealing a clear hierarchy in overall capa-bilities. *Vikhr* emerged as the top performer with an average score of 7.75, followed by *Llama3* (7.30) and *Mistral* (6.95). Notably, there is a substantial performance gap between these leading models and the lower-performing ones, with *Qwen2.5- 7B-Instruct* achieving 6.93 and *TinyLLaMA* scoring 1.12.

This stratification in the fig. 5 suggests that model size and architecture significantly influence creative generation capabilities in Russian language tasks. For example, our findings indicate a pronounced correlation between model size and performance. TinyLLaMA consistently performs poorly in all tasks, with scores predominantly in the range of 0-2, and complete failures (0.0) in culturally nuanced tasks such as poetry, culinary experience, and cultural history. This suggests

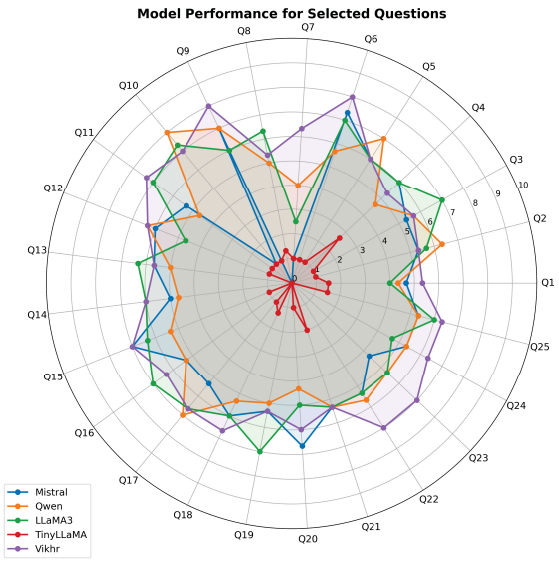

Fig. 5. Performance of Models for all the questions

a minimum parameter threshold required for effective Russian language generation tasks.

Among larger models, a more nuanced picture emerges. Vikhr demonstrates superior capabilities in technical domains (Tech Reviewer: 8.0, Tech Writer: 8.0), while LLaMA3 excels in creative writing (Essay Writer: 7.0, Adventure Game: 7.0). Qwen shows particular strength in communication-focused tasks (Public Speaking Coach: 8.0).

Our analysis reveals interesting task-specific failure patterns. Mistral, despite strong performance in technical reviews, completely fails at ASCII art generation (0.0). Similarly, most models struggle with emergency response scenarios, with the exception of the Russian-adapted Vikhr. These disparities suggest that certain generative tasks require specialized training or cultural context that general language models may lack.

Furthermore, the varying performance across different creative domains indicates that model architecture and training methodology should be tailored to intended use cases, with specialized models potentially outperforming general-purpose ones in specific creative generation contexts.

Evaluation of the Stability coefficient has also done for the models.

The fig. 6 shows the overall Stability Coefficient Across models for the task. Llama3.2 model produces the most variable answers, with the highest score in answer variability, followed by Vikhr. This suggests that Vikhr may generate a wide range of responses, potentially capturing diverse aspects of the LCS problem but also indicating less consistency. Mistral, on the other hand, shows the least variability, suggesting more consistent but potentially less diverse answers. Tinyllama fall in between, showing moderate variability

And finally, the important evaluation CES (Combined Evaluation Score) also calculated for the models. From the fig. 7, we can see the results. Among the evaluated models, Vikhr demonstrates the best balance between stability and

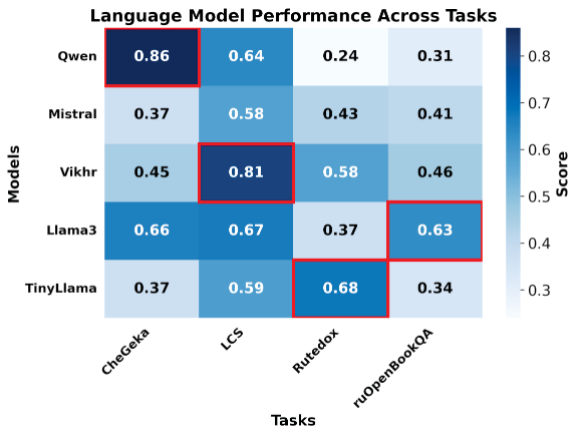

Fig. 6. Stability Coefficient Across models

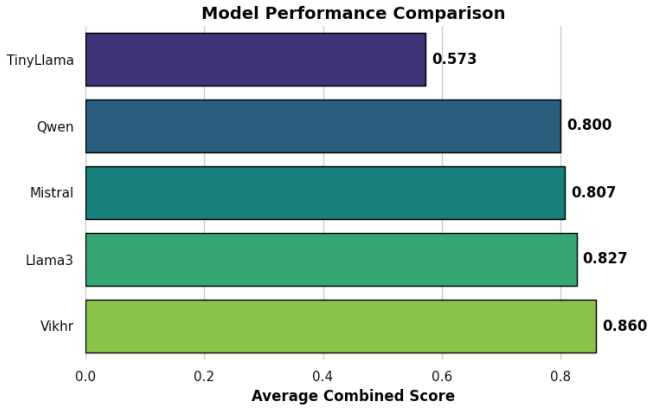

Fig. 7. Performance of Models for all the questions

creativity, achieving a high CES by maintaining both diverse and coherent outputs. Mistral, while exhibiting strong stability, struggles with creative flexibility, often producing repetitive responses. In contrast, Qwen achieves high creativity but suffers from inconsistencies across prompts, leading to lower stability. These insights highlight the trade-offs between stability and creativity, guiding future improvements in model fine-tuning.

V. Discussion

A benchmarking system for Russian Large Language Models (LLMs) requires unique solutions because it must handle the distinct conditions of both Russian linguistic features and computational and cultural aspects. The benchmarking dataset shortage in Russian together with pretraining corpora inconsistencies along with domain-specific biases creates challenges for Russian Large Language Models because English has extensive benchmarking resources available.

Creativity assessment represents a core component of our benchmark because the Russian language features complex morphological structures and adaptable word order patterns and contextual meaning usage. The successful demonstration of creativity by Russian LLMs requires mastery of word inflections alongside mastery of idiomatic expressions as well as effective handling of diverse stylistic registers because Russian differs syntactically from English. Human evaluators must supplement automated metrics because creative assessment becomes complicated by the nature of Russian language.

Stability functions as an essential major dimension for reliability when it comes to AI-generated responses. Minor alterations in prompts to Russian texts lead to large structural changes in the sentence because this language displays substantial syntactic flexibility. The stability coefficient we implemented responds to phrasing variations in inputs, so models prevent output inconsistencies or errors when input text is slightly changed.

Fairness, together with free-from-bias, needs special attention in this process. The Russian-language models bring in biases when they are trained from their foundational datasets in the same manner as English-language models do.

The evaluation of Russian LLMs requires the implementation of elaborate linguistic-based assessment frameworks. Our benchmark serves as a base assessment instrument for examining creativity alongside stability but future development will broaden its analysis areas through additional language elements and cognitive patterns and it will improve assessment methods based on practical applications.

VI. Conclusion

This paper presents a novel benchmarking framework that evaluates the creativity and stability of Russian Large Language Models as well as attempts to solve the puzzle of methodical evaluation of Russian lenguage models.

With the help of imposing a cretivity framework on a model, we ensure that the generative model is not only stable, but can also verbl and consistently rich in context.

Furthermore, our set of metrics provides deep insights into the existing strengths and weaknesses of the Russian Large Language Models. It allows us to trace the interplay between the two most notable qualities of the models and highlights the stability versus creativity tradeoffs.

Moving forward, our work can serve as a foundation for expanding evaluation methodologies to additional linguistic and cognitive dimensions, such as reasoning, factual consistency, and ethical considerations. Further research could also explore how fine-tuning or retrieval-augmented methods impact model performance in both creativity and stability.

Ultimately, our benchmark contributes to the broader goal of improving AI transparency and usability for Russian-language applications. We hope that this work will encourage more rigorous, systematic evaluation efforts, leading to the development of more reliable and effective AI systems in diverse real-world scenarios.

Acknowledgment

This work was supported by the Russian Science Foundation, agreement no. 24-71-00115, https://rscf.ru/en/project/24-71-00115/.

References

[1] J. Blanchet, P. Cui, J. Li, and J. Liu, "Stability evaluation via distributional perturbation analysis," 2024. [Online]. Available: https://arxiv.org/abs/2405.03198

[2] R. Sinha, Z. Song, and T. Zhou, "A mathematical abstraction for balancing the trade-off between creativity and reality in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2306.02295

[3] A. Fenogenova and et al., "MERA: A comprehensive LLM evaluation in Russian," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9920–9948. [Online]. Available: https://aclanthology.org/2024.acl-long.534/

[4] W. Orwig, E. Edenbaum, J. Greene, and D. Schacter, "The language of creativity: Evidence from humans and large language models," *The Journal of Creative Behavior*, vol. 58, 01 2024.

[5] T. Shavrina and et al., "RussianSuperGLUE: A Russian language understanding evaluation benchmark," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4717–4726. [Online]. Available: https://aclanthology.org/2020.emnlp-main.381/

[6] I. Ul Haq and M. Pifarré, "Dynamics of automatized measures of creativity: mapping the landscape to quantify creative ideation," *Frontiers in Education*, vol. 8, 2023. [Online]. Available: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2023.1240962

[7] Anonymous, "Automated creativity evaluation for large language models: A reference-based approach," in *Submitted to ACL Rolling Review - December 2024*, 2025, under review. [Online]. Available: https://openreview.net/forum?id=O8X7VQmSBn

[8] T. Chakrabarty, V. Padmakumar, F. Brahman, and S. Muresan, "Creativity support in the age of large language models: An empirical study involving emerging writers," *ArXiv*, vol. abs/2309.12570, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262217523

[9] G. Franceschelli and M. Musolesi, "On the creativity of large language models," *AI SOCIETY*, pp. 1–11, 11 2024.

[10] B. Atil and et al., "Non-determinism of "deterministic" llm settings," 2025. [Online]. Available: https://arxiv.org/abs/2408.04667

[11] J. Patterson, B. Barbot, J. Lloyd-Cox, and R. Beaty, "Audra: An automated drawing assessment platform for evaluating creativity," *Behavior Research Methods*, vol. 56, 11 2023.

[12] Y. Zhao and et al., "Assessing and understanding creativity in large language models," *CoRR*, vol. abs/2401.12491, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.12491

[13] M. Kamaluddin, M. Rasyid, F. Abqoriyyah, and A. Saehu, "Accuracy analysis of deepl: Breakthroughs in machine translation technology," *Journal of English Education Forum (JEEF)*, vol. 4, pp. 122–126, 06 2024.

[14] W.-L. Chiang and et al., "Chatbot arena: An open platform for evaluating llms by human preference," 2024. [Online]. Available: https://arxiv.org/abs/2403.04132

[15] Q. Jia, X. Yue, T. Zheng, J. Huang, and B. Y. Lin, "Simulbench: Evaluating language models with creative simulation tasks," 2024. [Online]. Available: https://arxiv.org/abs/2409.07641

[16] H. Chen and N. Ding, "Probing the "creativity" of large language models: Can models produce divergent semantic association?" in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 881–12 888. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.858/

[17] I. Churin and et al., "Long input benchmark for russian analysis," 2024. [Online]. Available: https://arxiv.org/abs/2408.02439

[18] G. Team and et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024. [Online]. Available: https://arxiv.org/abs/2403.05530