УДК 004.896

DEVELOPMENT OF A CREATIVITY BENCHMARK FOR RUSSIAN LANGUAGE WHEN ASSESSING THE QUALITY OF LANGUAGE MODELS

Abdurakhiumov M.A. (ITMO University)

Supervisor – Senior Researcher at National Center for Cognitive Research

Khodorchenko M.A. (ITMO University)

Introduction. The rapid development of Large Language Models (LLMs) has significantly advanced natural language processing (NLP) tasks. However, assessing the creativity of LLMs, particularly in the Russian language, remains a challenge. Traditional evaluation metrics, such as BLEU scores and classification-based metrics, fail to capture qualitative aspects such as coherence, creativity, and relevance. Additionally, the lack of dedicated NLP tasks and datasets for Russian further complicates the assessment process. This study aims to address these challenges by adapting a creativity benchmark for LLMs in the Russian language.

Main Section. The research focuses on developing a Russian Creativity Benchmark by systematically adapting existing evaluation frameworks. To achieve this, a literature review was conducted on creativity assessment in NLP, identifying gaps in current methodologies. The adaptation process involved analysing the structure and evaluation metrics of SimulBench, a framework designed for assessing creativity in English-language models. Since many linguistic features and cultural references in English do not directly translate into Russian, careful modifications were necessary to ensure the benchmark's relevance. Russian equivalents were prepared for English-specific elements, maintaining linguistic and contextual integrity.

The experimental framework was designed to evaluate the creative capabilities of Russian-language models by adapting existing datasets and prompts. These prompts were not only translated but also adjusted for cultural and linguistic differences. The adaptation process involved replacing culturally specific references, such as Western literature and historical figures, with appropriate Russian counterparts. Each modification was validated by native Russian speakers and linguistic experts to ensure accuracy. The tasks covered various domains of creativity, including travel guides, poetry, journalism, and interactive storytelling, allowing for a comprehensive evaluation of the models' creative abilities.

To assess the models, a diverse selection of Russian-language LLMs was tested, including saiga_llama3_8b, Vikhr-Nemo-12B-Instruct-R-21-09-24, TinyLlama-1.1B-32k-Instruct, Mistral-Nemo-Instruct-2407, and Qwen2.5-7B-Instruct. These models were chosen based on their performance in the Vikhr Benchmark Arena Hard for Russian-language tasks. Each model was evaluated using adapted metrics that focused on linguistic coherence, cultural relevance, and creativity.

One of the key challenges in evaluating creativity lies in the limitations of traditional NLP metrics, which often fail to capture the nuanced aspects of creative expression. To address this, Google's Gemini Flash model was used as an automated evaluator. This choice was based on its advanced cross-lingual understanding and ability to apply consistent evaluation criteria across diverse creative outputs. The evaluation process was designed to measure three primary aspects: creativity, coherence, and diversity. The creativity score assessed the originality and novelty of generated content, particularly focusing on how well the models produced unique ideas and narratives. The coherence score evaluated the logical flow, narrative consistency, and adherence to Russian linguistic conventions. The diversity score quantified lexical variety, and the use of different stylistic devices commonly found in Russian literature.

Conclusion. The study highlights the critical gaps in assessing the creativity of LLMs in the Russian language. By developing a specialized benchmark, we provide a foundation for evaluating and improving model performance beyond conventional metrics. The findings contribute to enhancing LLM adaptability to complex linguistic structures, promoting more meaningful and creative AI-generated text in Russian.

References:

1. Cobbe, K., et al. (2021). Evaluating Large Language Models Trained on Code. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 1–20).

2. Hendrycks, D., et al. (2020). Measuring Massive Multitask Language Understanding. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 1–15).

3. Du, X., Liu, Y., & Han, Y. (2023). Collaborative Creativity in Large Language Models: A Multi-Model Approach to Enhancing Creative Outputs. Artificial Intelligence Review, 56(2), 123–145.

4. Gómez-Rodríguez, C., & Williams, H. (2023). English Creative Writing with LLMs: Analyzing Outputs for Novelty and Value. Journal of Artificial Intelligence Research, 74(1), 1–30.

Abdurakhimov M.A. (author)

Khodorchenko M.A. (Supervisor)