# НАЗВАНИЕ ТЕЗИСОВ (НЕ БОЛЕЕ 3 СТРОК).

# ШРИФТ TIMES NEW ROMAN 14 PT, ЖИРНЫЙ,

# ЗАГЛАВНЫЕ БУКВЫ (стиль «Заголовок 1»)

Buzykanov S1, Usmanov A1, Gavrikov P1, Revayev D1

1. MIPT, Department of Multimedia Technnologies and Telecommunications

In recent years, large language models (LLMs) have shown remarkable capabilities in natural language processing tasks, but their training and inference costs remain prohibitively high, especially when applied to complex, real-world data. This research presents a novel benchmark designed to evaluate LLMs' ability to reason over facts distributed across extremely long documents and assess the computational resources required for their training and inference. The benchmark includes synthetic tasks that emulate real-world challenges, pushing models to handle long-context reasoning, memory management, and inference speed.

Our study evaluates and compares various LLMs based on transformer architectures and state-space models (SSMs), which offer potential improvements in context processing and computational efficiency. We will present metrics that capture model response quality, effective length of processed context, memory requirements, and training time for different architectures. By providing insights into the trade-offs between model performance and resource consumption, this research contributes to addressing the challenges of scaling LLMs in resource-constrained environments.

The complexity and high costs of training and deploying large language models highlight the importance of efficient evaluation methods. This research is especially relevant as the demand for deploying LLMs in diverse industries continues to grow, making cost-effective solutions crucial. Our findings will benefit both academic researchers and industry practitioners looking for optimized solutions in intelligent data analysis systems, AI-driven applications, and computational resource management.

**Keywords:**  
Large Language Models, Transformers, State-Space Models, Benchmarking, Model Evaluation, Computational Efficiency, Long-Context Processing, Neural Networks

References

1. Y. Kuratov, A. Bulatov, P. Anokhin, I. Rodkin, D. Sorokin, A. Sorokin, M. Burtsev: BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack, arXiv:2406.10149