Long Context Benchmark for the Russian Language

Igor Churin[1], Murat Apishev[1,2], Maria Tikhonova[1,3], Denis Shevelev[1], Aydar Bulatov[4,5], Yuri Kuratov[4,5], Sergej Averkiev[1], Alena Fenogenova[1,3]

[1]SberAI, [2]Yandex, [3]HSE University, [4]AIRI, [5]MIPT

Correspondence: alenush93@gmail.com

Abstract

Recent progress in Natural Language Processing (NLP) has driven the creation of Large Language Models (LLMs) capable of tackling a vast range of tasks. A critical property of these models is their ability to handle large documents and process long token sequences, which has fostered the need for a robust evaluation methodology for long-text scenarios. To meet this requirement in the context of the Russian language, we present our benchmark consisting of 18 datasets designed to assess LLM performance in tasks such as information retrieval, knowledge extraction, machine reading, question answering, and reasoning. These datasets are categorized into four levels of complexity, enabling model evaluation across context lengths up to 128k tokens. To facilitate further research, we provide open-source datasets, a codebase, and a public leaderboard associated with the benchmark.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive abilities in many NLP applications. Interacting with people through free-form text instructions, they serve as versatile tools for multiple scenarios, transforming the landscape of AI systems. One direction where LLM usage is developing rapidly includes tasks requiring long text processing, such as information retrieval (IR) and summarization, where their applications alleviate the handling of long texts for humans.

However, until recently, most LLMs had difficulties handling long sequences of tokens and were only able to work with a limited context length of several thousand tokens. In recent years, new methods have enabled the models to increase their context significantly, empowering them to solve a new variety of tasks. This, in turn, and the community's demand for automatic systems solving such tasks at a good level has created a need for a thorough evaluation of LLM long context understanding.

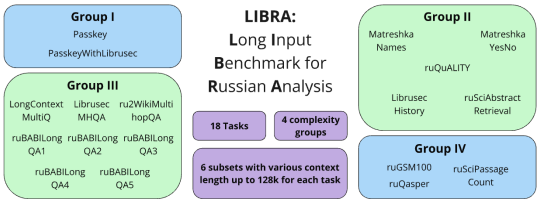

Figure 1: **The LIBRA benchmark** is a set of 18 long-context tasks ranging in length from 4k to 128k tokens, grouped into four categories based on the complexity of required skills.

To address this demand in English, several long context understanding benchmarks have been created recently with LongBench (Bai et al., 2023)[1] and L-Eval (An et al., 2023)[2] heading the list. However, the Russian language, at this point, lacks a fair instrument for transparent evaluation of long context understanding.

Our work addresses this problem and presents a new benchmark, which we call **L**ong **I**nput **B**enchmark for **R**ussian **A**nalysis, or LIBRA, for the evaluation of LLM long context understanding abilities in Russian (see Figure 1) including such aspects as IR, machine reading, question answering (QA), and reasoning. The contribution of our work can be summarized as follows:

- we present a methodology for the evaluation of long-context abilities of LLMs for the Russian language;
- we publicly release a set of 18 datasets of various skills and complexities in Russian, which form the LIBRA benchmark;

[1]https://huggingface.co/datasets/THUDM/LongBench
[2]https://huggingface.co/datasets/L4NLP/LEval

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 1–13
November 9, 2025

- we release a codebase [3], a public leaderboard [4] and a set of baseline solutions.

2 Related Work

Long Context Large Language Models. One of the crucial tasks in the development of LLMs is to increase the length of the context that the model can understand. This problem has two key points: the complexity of calculations for long sequences and the ability of the model to extract important data in a long context. The solution of the first problem can be attributed to research on the effective processing of the self-attention as in Longformer (Beltagy et al., 2020), LongNet (Ding et al., 2023) and FlashAttention (Dao et al., 2022; Dao, 2023), using caches for previously calculated outputs such as Transformer-XL (Dai et al., 2019), Unlimiformer (Bertsch et al., 2024) and LongLLaMA (Tworkowski et al., 2024) or replacing it with another mechanism with more effective inference as in RetNet (Sun et al., 2023) and Mamba (Gu and Dao, 2023). The solution to the second problem is to improve positional encoding techniques such as ALiBi (Press et al., 2021) and RoPE-based approaches (Sun et al., 2022; Peng et al., 2023).

Long Context Benchmarks. Until recently, most LMs had relatively small context lengths limited by a few thousand tokens. Thus, standard Natural Language Understanding (NLU) benchmarks (Wang et al., 2018, 2019; Shavrina et al., 2020) contained tasks within this size. Even today, benchmarks created recently, such as HELM (Bommasani et al., 2023), MT-Bench (Zheng et al., 2023), and Russian-oriented benchmark MERA (Fenogenova et al., 2024) follow this pattern, limiting their tasks by relatively small context window size to simplify the evaluation procedure and reducing its cost.

The pioneers of long context processing benchmarks have been ZeroSCROLLS (Shaham et al., 2023)[5], designed to test zero-shot model capabilities for NLU over long texts; L-eval (An et al., 2023)[6], focused on a standardized evaluation methodology for long context LMs addressing two key aspects: dataset construction and evaluation metrics; Loong (Wang et al., 2024b), which aligns with realistic scenarios through extended multi-document QA; and LongBench (Bai et al., 2023), the bilingual multi-task benchmark for long context understanding, comprising 21 tasks in English and Chinese. Finally, Goldman et al. (2024) categorizes the existing long-context datasets and positions them with respect to their difficulty, which they define by the dispersion and the scope.

The concept of the *needle-in-a-haystack* (Kamradt, 2023) is frequently employed in long-context benchmarks, involving the insertion of sentence-level information at varying depths within a document to create tasks of differing complexity. In addition to categorizing tasks by type, many benchmarks classify tasks based on their complexity using various criteria. For instance, tasks may be grouped by the number of facts required for reasoning (Kuratov et al., 2024), the type of reasoning QA (e.g., single-hop vs. multi-hop, as shown by Wang et al. (2024a)), or the complexity and depth of the needle. In the latter case, deeper or more abstract needles challenge models more significantly, testing their ability to locate and reason over critical details in long documents (Karpinska et al., 2024).

However, the limitation of the benchmarks mentioned above is that they are mainly English-oriented (or Chinese). As for the Russian language, there is an urgent need for a reliable system able to evaluate LLM long context understanding abilities. To address this problem, we propose LIBRA, which brings a methodology and 18 tasks for a long context understanding evaluation in Russian.

3 LIBRA

3.1 Benchmark Overview

In this section, we introduce LIBRA (Long Input Benchmark for Russian Analysis), a new benchmark for long context understanding, which includes 18 tasks for LLM evaluation created specifically for Russian. LIBRA aims to evaluate a large scope of LLMs, including pretrain models and models with supervised finetuning (SFT) with any system prompt that can be picked up.

The main purpose of the benchmark is to create a reliable instrument for the long context understanding evaluation, enabling the study of the model's ability to solve various tasks of different complexity with respect to the input context length. For this purpose, all tasks in the LIBRA benchmark are divided into 4 complexity groups, and the datasets have several subsets of various context lengths rang-

[3] https://github.com/ai-forever/LIBRA

[4] https://huggingface.co/spaces/ai-forever/LIBRA-Leaderboard

[5] https://www.zero.scrolls-benchmark.com/

[6] https://huggingface.co/papers/2307.11088

	Task Name	Data Origin	Skills	Metric	Dataset Size
I	Passkey	New	Reasoning	EM	1200
	PasskeyWithLibrusec	New	Reasoning	EM	1200
II	MatreshkaNames	New	Dialogue Context, Reasoning	EM	900
	MatreshkaYesNo	New	Dialogue Context, Reasoning	EM	1799
	LibrusecHistory	New	Reasoning	EM	128
	ruSciAbstractRetrieval	New	Reasoning	EM	1240
	ruQuALITY	Translated	Reasoning	EM	202
III	LongContextMultiQ	New	Reasoning	EM	1200
	LibrusecMHQA	New	Reasoning	EM	384
	ru2WikiMultihopQA	Translated	Reasoning	EM	300
	ruBABILongQA1	New	Reasoning	EM	600
	ruBABILongQA2	New	Reasoning	EM	600
	ruBABILongQA3	New	Reasoning	EM	600
	ruBABILongQA4	New	Reasoning	EM	600
	ruBABILongQA5	New	Reasoning	EM	600
IV	ruSciPassageCount	New	Reasoning	EM	600
	ruQasper	Translated	Reasoning	F1	203
	ruGSM100	Translated	Math, Logic	EM	100

Table 1: The LIBRA tasks outline. The numbers **I**, **II**, **III**, and **IV** in the left column indicate the complexity group of the tasks described in Subsection 3.2. The **Skills** column defines the skills to be tested on a specific task. **Data Origin** discloses the source of the dataset. The **Dataset Size** column shows the number of items in the whole dataset.

ing from 4k up to 128k tokens[7]. The latter makes it possible to explore the influence of the context length on the model results.

3.2 Complexity group description

We describe each complexity group of tasks using criteria inspired by other benchmarks that classify tasks by complexity. Specifically, we considered the depth of the needle, the complexity of reasoning, and the difficulty of the domain.

The first complexity group (I) consists of tasks that require finding a short text fragment in long textual paragraphs containing irrelevant information. This group includes Passkey and PasskeyWithLibrusec datasets.

The second complexity group (II) includes tasks that require answering the question based on a relevant context. The following types of tasks are related to this group: QA such as MatreshkaNames, MatreshkaYesNo, LibrusecHistory, ruSciAbstractRetrieval and multiple choice QA, such as ruQuALITY.

The natural development of tasks from the second class of complexity are tasks with questions, the answers to which are not explicitly contained in the text but require the analysis of fragments of input data and the generation of an answer based on it. Such tasks in our classification belong to **the third complexity group (III)** and represent a multi-hop QA (MHQA) type. This group includes the following tasks: ruBABILongQA1, ruBABILongQA2, ruBABILongQA3, ruBABILongQA4, ruBABILongQA5, LongContextMultiQ, LibrusecMHQA and ru2WikiMultihopQA.

Finally, to **the fourth complexity group (IV)** belongs to the tasks that require understanding the whole context, solving mathematical problems, and QA tasks within complex domains. This group includes ruSciPassageCount, ruGSM100 and ruQasper datasets.

We do not include code generation and analysis tasks in LIBRA as most of the software code in the world is written in languages based on English.

3.3 Context Length Estimation

We divide all datasets into subsets of various context lengths. The latter, however, may vary across different models and tokenizers. In our work, we used the fertility of tokenizer to distribute samples across different context lengths, which indicates the average number of tokens in which one word is tokenized. Thus, the average length in tokens for the text can be approximated by the number of words multiplied by the fertility number.

For the fertility approximation, we calculate the average fertility of the classic LLM tokenizers, which are further evaluated as baselines (see Appendix C for model description) on a complete list of datasets, by computing it as the total number of tokens divided by the total number of words. The

[7] See explanation on token length calculation in Section 3.3

Model Name	Fertility
GLM4-9B-Chat	2.15
T-lite-instruct-0.1	2.34
Saiga-LLaMA-3-8B	2.40
LLaMA-3-8B	2.40
LLaMA-3-8B-Instruct	2.40
LLaMA-3.1-8B-Instruct	2.40
LLaMA-3.1-8B	2.40
Phi-3-mini-128k-instruct	2.74
LLaMA-2-7B-32K	2.83
LongAlpaca-7B	2.83
LongChat-7B-v1.5-32k	2.83
Mistral-7B-v0.1	3.08
Mistral-7B-v0.3	3.08
Mistral-7B-Instruct-v0.3	3.08
Mistral-Nemo-Instruct-2407	3.08
ChatGLM2-6B-32k	3.50

Table 2: The average fertility of tokenizers, where fertility is defined as the average number of tokens per word.

fertility of each model is shown in Table 2. The average fertility is 2.7. However, we decided to choose it with a margin so that the multilingual model with the highest fertility can be tested on the entire benchmark. As a result, we set the standard fertility to 3.

Finally, using the selected fertility value, we divided all datasets into subsets of various context lengths ranging from 4k to 128k tokens. Table 3 gives the resulting dataset sizes and average sample context lengths.

3.4 Datasets

This section describes the datasets and data collection process in detail. The benchmark datasets originate from the following sources: 1) entirely new datasets based on open data in Russian (14 datasets out of 18) and 2) translation of English datasets using Google translator API[8] followed by manual verification and correction. We do not generate samples with LLMs and use annotators markup. This helps reduce bias from using models like GPT-4, which are also part of the assessment. However, it has some drawbacks, as full annotation can be costly and time-consuming in certain cases. The exact dataset format can be found in Appendix B.

Passkey The Passkey is a synthetic QA dataset based on the idea of the original passkey dataset from LongLLaMA's GitHub repository[9]. The main idea of the task is to extract a relevant piece of code number from a long text fragment that was created by repeating short sentence template containing noise. The model must find this code among the irrelevant information.

PasskeyWithLibrusec The PasskeyWithLibrusec is a more complicated version of Passkey QA dataset, in which we use randomly selected texts from the Librusec dataset[10] as noise to make this dataset more difficult for LLMs.

MatreshkaNames This dataset is based on Russian names[11] and Matreshka[12] . The Matreshka dataset comprises brief interactions involving "user" and "bot" roles, along with a brief description of the topic being discussed by each participant. To form longer contextual samples, we combined multiple interactions and replaced the names "user" and "bot" with the pull of names taken from the dataset of Russian names. Subsequently, we randomly selected a topic from the combined interactions and the name of the person discussing that topic. The dataset requires the model to identify the individual who discussed the selected topic.

MatreshkaYesNo The MatreshkaYesNo is a binary classification dataset based on Matreshka and Russian names sets. It is similar to the MatreshkaNames dataset but instead of predicting names, the model is supposed to indicate whether this topic was mentioned in the dialog. The dataset is balanced across Yes/No answers.

LibrusecHistory This dataset was created in QA format using Librusec. Each sample comprises a text paragraph and a corresponding question. To create tasks with different input lengths, we selected large texts from books in different domains and styles, divided them into fragments of several thousand tokens, and created the annotation (see Appendix A). These fragments became the dataset's samples. Longer samples, with lengths up to 64,000 tokens, were created by supplementing these fragments with neighboring paragraphs from the original large text on both sides.

ruSciAbstractRetrieval The ruSciAbstractRetrieval is a QA dataset ideologically similiar to the PassageRetrieval (Bai et al., 2023)[13] dataset from LongBench, that aims to evaluate model's reasoning skills. Each element of the dataset consists of a summary description of the topic and a set text para-

[8] https://pypi.org/project/googletrans/

[9] https://github.com/CStanKonrad/long_llama/blob/main/examples/passkey.py

[10] https://huggingface.co/datasets/IlyaGusev/librusec

[11] https://www.kaggle.com/datasets/rai220/russian-cyrillic-names-and-sex/data

[12] https://huggingface.co/datasets/zjkarina/matreshka

[13] https://huggingface.co/datasets/THUDM/LongBench/viewer/passage_retrieval_en

	Dataset Name	4k size / avg len	8k size / avg len	16k size / avg len	32k size / avg len	64k size / avg len	128k size / avg len
I	Passkey	200 / 2790	200 / 5450	200 / 10996	200 / 21730	200 / 43391	200 / 87974
	PasskeyWithLibrusec	200 / 2705	200 / 5563	200 / 10835	200 / 22215	200 / 44682	200 / 88189
II	MatreshkaNames	150 / 3190	150 / 6314	150 / 12128	150 / 24168	150 / 48184	150 / 96135
	MatreshkaYesNo	299 / 3200	300 / 6317	300 / 12134	300 / 24173	300 / 48189	300 / 96142
	LibrusecHistory	-	32 / 4515	32 / 9003	32 / 17976	32 / 35924	-
	ruSciAbstractRetrieval	210 / 3264	210 / 7260	210 / 15245	210 / 31231	200 / 63594	200 / 127777
	ruQuALITY	-	41 / 6380	161 / 12387	-	-	-
III	LongContextMultiQ	200 / 2940	200 / 6360	200 / 12240	200 / 26572	200 / 37482	200 / 68239
	LibrusecMHQA	-	384 / 4574	-	-	-	-
	ru2WikiMultihopQA	-	49 / 6378	128 / 11633	123 / 25523	-	-
	ruBABILongQA1	100 / 4002	100 / 8001	100 / 16002	100 / 32001	100 / 64002	100 / 128001
	ruBABILongQA2	100 / 4002	100 / 8001	100 / 16002	100 / 32001	100 / 64002	100 / 128001
	ruBABILongQA3	100 / 4011	100 / 8010	100 / 16011	100 / 32010	100 / 64011	100 / 128010
	ruBABILongQA4	100 / 4014	100 / 8013	100 / 16014	100 / 32013	100 / 64014	100 / 128013
	ruBABILongQA5	100 / 4006	100 / 8005	100 / 16006	100 / 32005	100 / 64006	100 / 128005
IV	ruSciPassageCount	100 / 3528	100 / 7128	100 / 13616	100 / 27160	100 / 53108	100 / 105949
	ruQasper	-	48 / 5768	134 / 11071	21 / 25185	-	-
	ruGSM100	-	-	100 / 9083	-	-	-

Table 3: Sizes and average sample lengths for the task subsets of various context lengths. **Dataset Name** shows the name of the dataset. The columns **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show the number of samples and average sample lengths in tokens for the corresponding context length.

graphs created from abstracts of scientific articles from ruSciBench[14]. The goal is to identify the paragraph where the specified topic is discussed. To create this dataset, we randomly choose some abstracts from ruSciBench and generate descriptions of their topics using human annotators to acquire targets.

ruQuALITY The ruQuALITY dataset was created as a translation of the original QuALITY[15] from L-Eval, which consists of selected samples with a long context from the original multiple choice QA dataset called QuALITY (Pang et al., 2021). The model must find relevant information in the text and answer by choosing one of the four suggested options.

LongContextMultiQ The LongContextMultiQ is a multi-hop QA long context dataset for Russian that is based on data used for the MultiQ (Taktasheva et al., 2022)[16] dataset creation. The original MultiQ dataset is created by multi-hop dataset generation based on Wikidata[17] and Wikipedia, and consists of samples with different length. We selected 200 samples from these generated sources with a long context for each context length.

LibrusecMHQA This dataset was created in MHQA format, also using Librusec as a LibrusecHistory. The main difference between these datasets is that in the LibrusecMHQA dataset, the necessary information for the answer is distributed in several parts of the context, making the task more difficult and allowing us to evaluate the model's reasoning skills better. The generation procedure for samples of different lengths remains the same.

ru2WikiMultihopQA The ru2WikiMultihopQA was created by translating the dataset 2WikiMultihopQA[18] from LongBench, which consists of selected samples with a long context from the original multi-hop QA dataset 2WikiMultihopQA (Ho et al., 2020). This Wikipedia-based dataset tests reasoning skills by requiring a model to combine information from multiple texts to answer a question. The format of this dataset, which consists of up to 5-hop questions, makes it difficult for LLMs.

ruBABILong We created a new methodology based on the idea from Kuratov et al. (2024) to create the Russian Benchmark for Artificial Intelligence for Long (ruBABILong)-context evaluation. It contains five long-context QA reasoning tasks using facts hidden among distractor facts and long books. The **ruBABILongQA1** task requires answering a question about a person's location using a single supporting fact. The **ruBABILongQA2** and **ruBABILongQA3** tasks introduce the challenge of differentiating subjects and objects, utilizing

[14] https://huggingface.co/datasets/mlsa-iai-msu-lab/ru_sci_bench

[15] https://huggingface.co/datasets/L4NLP/LEval/viewer/quality

[16] https://huggingface.co/datasets/ai-forever/MERA/viewer/multiq

[17] https://www.wikidata.org/wiki/Wikidata:Introduction

[18] https://huggingface.co/datasets/THUDM/LongBench/viewer/2wikimqa_e

two and three supporting facts, respectively. The **ruBABILongQA4** task tackles spatial reasoning through two-argument relations, while the **ruBABILongQA5** task involves tracking multiple objects to solve the three-argument relation problem. Each task contains 100 samples, scaled to six sequence lengths from 4k to 128k. We constructed the task facts for Russian according to the methodology of the original data from the bAbI dataset (Weston et al., 2016); no translation was performed, and facts were created directly in Russian. The background texts were sampled from Russian Librusec books.

ruSciPassageCount The dataset ruSciPassageCount uses the basic idea of the original PassageCount[19] dataset. This QA dataset requires the model to use the full context to solve the problem. To generate the data, we randomly select abstracts from ruSciBench, choose a number of repeats and an ID for the paragraph to repeat. Next, we add the remaining non-repeated paragraphs to the repeated paragraph until up to the desired context length. The resulting sequence of paragraphs is randomly shuffled. The ground truth for each sample is the number of unique paragraphs.

ruQasper The ruQasper was created by translating the Qasper[20] dataset from LongBench, which consists of selected samples with a long context from the original QA dataset over academic research papers (Dasigi et al., 2021). The goal of the task is to find the answer to the question in one of the parts of the article. The context for samples is drawn from scientific articles.

ruGSM100 The ruGSM100 dataset is a translation of gsm100[21] one from L-Eval. It contains 100 math problems to be solved using Chain-of-Thought in a few-shot mode. This dataset aims to evaluate the model's reasoning and logical skills in maths. The context for all tasks is a prompt of 16 examples with problem descriptions and answers.

Datasets BABILong, MHQA, and Passkey serve as examples of needle-in-a-haystack tasks.

3.5 Submission

To make a submission to the leaderboard, users first create a configuration file by adapting configs/template.ini (e.g., llama_3.1.ini) from the project's repo to specify the model parameters. Once the config is ready, they generate predictions and run the evaluation script from the repository. Both predictions and evaluation results are saved locally. Finally, users submit their results by creating a pull request to the repository. Upon approval, the model name and its evaluation are integrated into the system, with results made available on the leaderboard.

4 Experimental setup

We evaluate 17 popular LLMs that feature long context capability, including GPT-4o[22] (see Appendix C for the baseline details).

In order not to go beyond the context window we use zero-shot evaluation for all tasks, except for ruGSM100 in which the few-shot examples provided as a part of long context input. When the input length of the sample surpasses the maximum model context length, we truncate the input sequence from the right. For reproducibility, the baselines were evaluated with greedy decoding (temperature = 1.0, num_beams = 1, do_sample = False). We select the best result for each model from the two supported formats: with/without the chat template.

In addition, for each task, we fixed a natural language prompt unified for all the models (see Appendix B for the exact prompt formulation). The prompts were estimated from an empirical analysis of the tasks through a series of experiments. However, it should be noted that the benchmark methodology does not rigidly fix the prompts. Users can use their own prompts for evaluation. The choice of effective prompts requires additional research, which we leave for future work. We run all the experiments on two NVIDIA A100 GPU.

5 Results

The baseline results with respect to context length are given in Table 7 and with respect to tasks are in Tables 4, 5, and 6. Model-wise detailed results are provided in the benchmark repository. Analyzing baseline performance, we can draw the following conclusions.

Group I The tasks from this group are relatively simple, and most models pass them well within

[19] https://huggingface.co/datasets/THUDM/LongBench/viewer/passage_count

[20] https://huggingface.co/datasets/THUDM/LongBench/viewer/qasper_e

[21] https://huggingface.co/datasets/L4NLP/LEval/viewer/gsm100

[22] GPT-4o was included via API access as the state-of-the-art model representing the upper bound for long-context capabilities.

Model Name	Passkey	MatreshkaYesNo	MatreshkaNames	PasskeyWithLibrusec	LibrusecHistory	ruGSM100	ruSciPassageCount	ru2WikiMultihopQA
Complexity group	**I**	**II**	**II**	**I**	**II**	**IV**	**IV**	**III**
GPT-4o	**100.0**	**79.9**	**58.7**	**100.0**	**99.2**	**84.0**	**37.2**	**58.5**
GLM4-9B-Chat	100.0	68.0	47.3	100.0	82.0	8.0	7.5	48.8
LLaMA-3.1-8B-Instruct	100.0	69.5	39.8	100.0	64.8	23.0	5.6	27.8
LLaMA-3.1-8B	100.0	39.9	22.4	100.0	95.3	20.0	4.1	33.4
Mistral-Nemo-Instruct-2407	97.8	53.2	32.2	99.4	53.1	0.0	12.8	27.9
Mistral-7B-Instruct-v0.3	66.7	35.3	16.3	66.6	50.8	11.0	8.2	43.2
Phi-3-mini-128k-instruct	84.7	70.7	18.8	85.5	41.4	24.0	6.2	18.9
Mistral-7B-v0.3	66.7	32.0	10.0	66.7	68.0	9.0	0.0	41.0
LLaMA-2-7B-32K	66.7	33.4	3.4	65.5	40.6	7.0	4.7	37.2
LongChat-7B-v1.5-32k	66.5	33.4	5.9	66.0	26.6	5.0	4.8	35.2
LLaMA-3-8B-Instruct	33.3	27.3	16.6	33.3	22.7	0.0	6.5	17.7
T-lite-instruct-0.1	33.3	25.7	14.0	33.3	22.7	0.0	5.1	12.9
Saiga-LLaMA-3-8B	33.3	28.0	15.6	33.2	24.2	0.0	3.8	17.7
LLaMA-3-8B	33.3	20.2	10.0	33.3	22.7	0.0	3.3	18.4
Mistral-7B-v0.1	35.0	16.8	8.1	38.3	23.4	13.0	1.3	23.0
ChatGLM2-6B-32k	63.7	33.4	1.3	65.0	8.6	5.0	3.7	17.5
LongAlpaca	42.4	30.5	0.4	40.6	13.3	2.0	3.8	30.3

Table 4: The table presents the evaluation results. **Model Name** shows the name of the model. **Complexity group** indicates the complexity groups into which the tasks were divided in Table 1. The score for each task is averaged by the context length. The best score is put in bold, the second best is underlined.

Model Name	LongContextMultiQ	ruSciAbstractRetrieval	LibrusecMHQA	ruBABILongQA1	ruBABILongQA2	ruBABILongQA3
Complexity group	**III**	**II**	**III**	**III**	**III**	**III**
GPT-4o	7.8	**78.0**	**52.9**	**77.3**	**53.3**	27.2
GLM4-9B-Chat	7.8	77.8	44.5	54.1	29.8	22.3
LLaMA-3.1-8B-Instruct	**7.9**	77.4	31.8	55.8	24.0	23.9
LLaMA-3.1-8B	6.0	73.6	45.3	53.9	25.4	**29.6**
Mistral-Nemo-Instruct-2407	5.2	65.3	29.9	54.7	17.3	16.0
Mistral-7B-Instruct-v0.3	4.8	43.6	33.6	14.3	2.8	6.0
Phi-3-mini-128k-instruct	5.2	29.3	13.8	30.5	8.8	9.0
Mistral-7B-v0.3	5.2	30.5	39.1	37.3	16.7	15.7
LLaMA-2-7B-32K	7.9	39.1	27.6	40.3	16.6	16.3
LongChat-7B-v1.5-32k	3.2	41.1	24.7	17.5	7.2	4.0
LLaMA-3-8B-Instruct	4.9	31.4	46.1	23.7	4.1	4.5
T-lite-instruct-0.1	5.2	31.2	48.4	21.7	14.5	8.2
Saiga-LLaMA-3-8B	4.8	31.7	45.1	25.4	4.4	6.1
LLaMA-3-8B	7.0	30.9	41.4	20.8	7.7	9.1
Mistral-7B-v0.1	4.4	28.5	34.1	21.0	7.7	9.0
ChatGLM2-6B-32k	1.2	13.6	6.8	12.2	1.5	2.5
LongAlpaca	0.8	23.5	7.8	3.8	0.3	3.5

Table 5: The table presents the evaluation results. **Model Name** shows the name of the model. **Complexity group** indicates the complexity groups into which the tasks were divided in Table 1. The score for each task is averaged by the context length. The best score is bold, the second best is underlined.

Model Name	ruBABILongQA4	ruBABILongQA5	ruQuALITY	ruQasper	Overall
Complexity group	**III**	**III**	**II**	**IV**	**Overall**
GPT-4o	**66.0**	**84.7**	**89.5**	23.0	**65.4**
GLM4-9B-Chat	52.8	70.3	74.1	5.0	50.0
LLaMA-3.1-8B-Instruct	14.0	59.2	42.1	6.5	43.0
LLaMA-3.1-8B	52.1	67.9	12.0	4.3	43.6
Mistral-Nemo-Instruct-2407	12.4	45.9	67.0	**24.5**	39.7
Mistral-7B-Instruct-v0.3	27.6	37.6	30.6	5.4	28.0
Phi-3-mini-128k-instruct	1.0	44.1	38.8	3.5	29.7
Mistral-7B-v0.3	23.6	47.1	15.2	5.8	29.4
LLaMA-2-7B-32K	16.7	43.0	15.5	4.7	27.0
LongChat-7B-v1.5-32k	12.7	33.3	23.1	5.0	23.1
LLaMA-3-8B-Instruct	19.6	25.3	34.6	2.2	19.6
T-lite-instruct-0.1	22.3	24.4	11.0	2.7	18.7
Saiga-LLaMA-3-8B	20.3	25.2	17.9	2.5	18.8
LLaMA-3-8B	19.1	22.6	8.5	2.2	17.3
Mistral-7B-v0.1	12.4	23.2	17.3	2.5	17.7
ChatGLM2-6B-32k	0.6	8.8	49.2	2.6	16.5
LongAlpaca	0.2	29.4	44.0	2.0	15.5

Table 6: The table presents the evaluation results. **Model Name** shows the name of the model. **Complexity group** indicates the complexity groups into which the tasks were divided in Table 1. The score for each task is averaged by the context length. The **Overall** score is obtained by averaging the results over each task. The best score is put in bold, the second best is underlined.

Model Name	4k	8k	16k	32k	64k	128k
GPT-4o	**76.0**	**70.6**	**67.6**	**61.2**	**55.5**	**53.6**
GLM4-9B-Chat	61.9	57.6	52.1	49.6	49.1	43.8
LLaMA-3.1-8B-Instruct	56.1	48.5	44.7	43.8	44.5	39.1
LLaMA-3.1-8B	56.6	51.1	45.4	45.2	48.2	34.1
Mistral-Nemo-Instruct-2407	56.1	49.8	43.2	39.5	34.3	26.3
Mistral-7B-Instruct-v0.3	47.6	43.8	37.1	32.3	-	-
Phi-3-mini-128k-instruct	18.5	36.2	34.5	33.1	34.3	28.6
Mistral-7B-v0.3	50.5	45.2	39.8	36.6	-	-
LLaMA-2-7B-32K	47.0	44.6	37.3	34.7	-	-
LongChat-7B-v1.5-32k	41.5	37.3	31.7	26.9	-	-
LLaMA-3-8B-Instruct	58.0	56.1	-	-	-	-
T-lite-instruct-0.1	61.4	53.2	-	-	-	-
Saiga-LLaMA-3-8B	59.3	53.9	-	-	-	-
LLaMA-3-8B	56.1	49.5	-	-	-	-
Mistral-7B-v0.1	51.0	44.9	-	-	-	-
ChatGLM2-6B-32k	30.5	25.9	23.6	16.2	-	-
LongAlpaca	28.1	24.4	19.9	15.5	-	-

Table 7: The evaluation scores of various models across different context lengths. The columns **4k**, **8k**, **16k**, **32k**, **64k**, **128k** present evaluation scores averaged over all tasks. The best score is put in bold, the second best is underlined.

their maximum input length.

Group II MatreshkaYesNo, turns out to be the most straightforward task in the group, which all models cope with naturally. The ruQuALITY task is of medium complexity; several models achieved good scores on them. The classic QA task LibrusecHistory is effectively handled by modern models. The most complex task in this group is MatreshkaNames. For it several models (e.g., ChatGLM2-6B-32k, LLaMA-2-7B-32K) show low results for any input length.

Group III For tasks from ruBABILong, an increase in context leads to worse results. ruBABILongQA2 and ruBABILongQA3 turn out to be significantly more complex than others, which coincides with BABILong results from Kuratov et al. (2024). The length of the context plays a significant role; as it grows, the quality immediately begins to decline for all but the strongest models.

LibrusecMHQA turns out to be a complex dataset; the maximum quality of the models for solving this problem is only 52.9.

Group IV ruSciPassageCount is the most difficult task created from scratch, which all models except GPT-4o handle poorly; the result's sensitivity to the context's size is high. Most models fail to cope with ruQasper for complex tasks and domains and with mathematical problems from ruGSM100.

Overall, GPT-4o stands out among others, significantly exceeding its closest competitor GLM4-9B-Chat. SFT models generally perform better than the pretrained onces. In most cases, an increase in the input length negatively affects the model results on the task. In general, the results indicate that our prior division of tasks into groups is highly correlated with their complexity.

We also compared model results in English and Russian for the 4 translated datasets. The analysis and the detailed comparison can be found in our repository due to the page limit.

6 Conclusion

The rapid development of LLM has posed new challenges in evaluating their ability to process long texts. To address this problem, we have introduced LIBRA. This benchmark evaluates LLM long context understanding abilities through 18 long-context textual tasks and enables model evaluation across various context lengths ranging from 4k to 128k.

Our contribution encompasses a benchmark methodology with open-sourced datasets of different lengths and domains, a codebase for model evaluation, and baseline solution scoring. The datasets are published under the MIT license, and the leaderboard is available on HuggingFace [23].

Limitations

Data Representation. The texts included in the benchmark are gathered from specific domains, which might not cover the full range of Russian language usage. As a result, models may excel in benchmark tasks but struggle with texts outside these domains, limiting their generalization ability. Several datasets were created using automatic translation followed by manual adaptation. This approach was mainly chosen due to the high cost of manual data creation.

Methodology limitations. When creating the datasets, we hypothesized that synthetic augmentation of the context length of the datasets, such as LibrusecHistory, would not affect the results. Our experiments show that these tasks are pretty challenging for many models. We made this methodological assumption due to the limitations of human data annotation; it is difficult for people to read large texts and concentrate enough to create questions and search for information within them. This data creation method may result in task errors, particularly when a newly extended text fragment contains conflicting information that could impact the answer. However, we found this approach acceptable due to the increased speed and cost-effectiveness.

[23] https://huggingface.co/spaces/ai-forever/LIBRA-Leaderboard

Long context. The benchmark focuses on evaluating long contexts, but the definition of "long context" can differ based on the application and the model. The chosen context lengths may not be ideal for all usage scenarios, and models could exhibit varying performance. In this paper, we have measured the average fertility of baseline model tokenizers on a full list of datasets from our benchmark to sample different contexts and analyzed the models' results on our datasets across various context lengths. LMs with more parameters may inherently perform better, but this does not necessarily reflect improvements in long context understanding.

Additionally, in the present work we focused exclusively on evaluating performance with respect to context length, without considering the relative position of important information within the context. Future work should include performance evaluation on needle-in-a-haystack tasks with respect to the position of the needle along with an in-depth error analysis.

Data leakage is a critical concern for modern benchmarks because current models are trained on a significant amount of text from the Internet. Long context benchmarks are particularly risky, as their texts are based on web sources and books. This could potentially lead to data leakage. However, creating original long texts from scratch not found on the web is exceptionally costly. As a result, we use open sources to develop the benchmark, acknowledging the potential risks. Nevertheless, we firmly believe this will make a valuable contribution to the Russian community, as no long context datasets are currently available.

Ethical Considerations. The data used in the benchmark was created from open data sources. When annotating the data, we obtained transparent permission from all users and made efforts to maintain the confidentiality and anonymity of participants. As the benchmark develops, ongoing efforts are required to identify and minimize biases in the benchmark datasets and evaluation metrics. The benchmark does not currently contain the datasets covering the ethical or AI safety skill evaluation, but this is a space for future work.

References

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. 2024. Unlimiformer: Long-range transformers with unlimited length input. *Advances in Neural Information Processing Systems*, 36.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. 2023. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024. MERA: A comprehensive LLM evaluation in Russian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9920–9948, Bangkok, Thailand. Association for Computational Linguistics.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. 2024. Is it really long context if all you need is retrieval? towards genuinely difficult long context nlp. *arXiv preprint arXiv:2407.00402*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing llms.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A" novel" challenge for long-context language models. *arXiv preprint arXiv:2406.16264*.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *arXiv preprint arXiv:2406.10149*.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and 1 others. 2021. Quality: Question answering with long input texts, yes! *arXiv preprint arXiv:2112.08608*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. Russiansuperglue: A russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*.

Ekaterina Taktasheva, Tatiana Shavrina, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia Bashmakova, Svetlana Iordanskaia, and 1 others. 2022. Tape: Assessing few-shot russian language understanding. *arXiv preprint arXiv:2210.12813*.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2024. Focused transformer: Contrastive training for context scaling. *Advances in Neural Information Processing Systems*, 36.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Qian Wang, and Yue Zhang. 2024a. Novelqa: A benchmark for long-range novel question answering. *arXiv preprint arXiv:2403.12766*.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, and 1 others. 2024b. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Datasets and Benchmarks Track*.

A Data Annotation Details

The datasets LibrusecHistory, LibrusecMHQA, and ruSciAbstractRetrieval were created via the crowdsourced platform.

In the LibrusecHistory, annotators were instructed to read a lengthy text and generate four questions based on the text and answer them. Guidelines were provided regarding the type of questions to ask: 1) questions should be answerable using information present in the text 2) the questions must not be about widely known information but should be related to the text 3) questions can cover various aspects such as character actions, appearance, thoughts, events, and scene descriptions 4) logical deductions are not required to answer the questions 5) Each question should have a single, clear, unambiguous answer from the text.

The design of the dataset LibrusecMHQA project follows a similar structure to LibrusecHistory, but the question criteria were more complex. In this dataset, the questions were answered by expert editors rather than through crowd-sourcing. The main distinction in the criteria for annotators is the multi-hop questions, where simply reading the sentence containing the answer is insufficient. Instead, reading at least a paragraph of 2-5 sentences, or the entire relevant fragment, is necessary to gather information and generate a complete answer.

The ruSciAbstractRetrieval was collected by crowd-sourced annotators. These annotators were asked to read a long text annotation and briefly describe the contents. The criteria for the description were as follows: 1) The description must start with the word "Describes". 2) It must be a single sentence, which can be complex. 3) The description should not exceed 30 words, including conjunctions, particles, and prepositions. 4) It should include the main general ideas identified in the abstract but should not include details.

Training examples were available for all projects. The contributions of human annotators are amassed and stored in a manner that ensures anonymity. The average hourly compensation exceeds the minimum wage per hour in Russia. Each annotator is informed about topics that may be sensitive in the data, such as politics, societal minorities, and religion. Table 8 summarizes general details concerning the creation of the datasets via crowd-source on ABC[24] data labeling platform.

B Dataset Examples

This section provides examples of the task format for the benchmark datasets. The exact prompts for the benchmark are not fixed. Here, we provide prompts used in our experiments[25].

Passkey: *You are provided with a long text that contains the access key. Just remember the access key.*
Context: {*context*}
You only need to specify the access key in the response.
Question: {*input*}
Answer:

PasskeyWithLibrusec: *You are provided with a long text that contains the access key. Just remember the access key.*
Context: {*context*}
You only need to specify the access key in the response.
Question: {*input*}
Answer:

MatreshkaNames: *You are provided with several dialogues. Remember the names of the people and the topics they talked about.*
Context: {*context*}
In the answer, specify only the name of the interlocutor who spoke on the topic from the next question.
Question: {*input*}
Answer:

MatreshkaYesNo: *You are provided with several dialogues. Remember the names of the topics that the interlocutors talked about.*
Context: {*context*}
In the answer, you only need to specify 'Yes' if there was such a topic and 'No' if there was no such topic in the dialogues.
Question: {*input*}
Answer:

LibrusecHistory: *You are given a long text in which you need to find the answer to the question.*
Context: {*context*}
Find the answer in the text to the following question.
Question: {*input*}
Answer:

ruSciAbstractRetrieval: *Below are a few paragraphs. Determine which paragraph the short description corresponds to.*
Context: {*context*}
Determine which paragraph the short description corresponds to. The response must contain the paragraph number.
Question: {*input*}
Answer:

ruQuALITY: *You are given a long text in which you need to find the answer to the question.*
Context: {*context*}
You will be given several answers to the question in the text; choose only one correct one.
Question: {*input*}
Answer:

LongContextMultiQ: *You are given a long text where you need to find the answer to the question.*
Context: {*context*}
Find the answer in the text to the following question.
Question: {*input*}
Answer:

LibrusecMHQA: *You are given a long text where you need to find the answer.*
Context: {*context*}
Find the answer in the text to the following question.

[24] https://elementary.activebc.ru

[25] All examples are presented in English for transparency and are given for illustrative purposes only to clarify the idea of a given task. The examples are not necessarily a direct translation of specific examples from the dataset. The exact prompts in their original formulation in Russian can be found in our repository https://github.com/ai-forever/LIBRA.

Task Name	Total	Pay Rate	Example Number	Overlap
LibrusecHistory	84$	6.25$/hr	32	1
LibrusecMHQA	458$	6.25$/hr	40	3
ruSciAbstractRetrieval	290$	6.25$/hr	100	3

Table 8: The details of datasets collection. **Total** is the budget spent to annotate the tasks employed for metric evaluation. **Pay Rate** is the hourly rate computed as a simple average of pay rates based on time spent annotating one row and the reward for this row. **Example Number** refers to the total number of samples processed while collecting or verifying the dataset. **Overlap** is the median number of votes per dataset sample averaged across all annotation tasks for the same dataset (if more than 1 task is provided).

Model Name	Model Type	Parameters	Max Length	Model HuggingFace link
GPT-4o	SFT	-	128k	-
GLM4-9B-Chat	SFT	9B	128k	THUDM/glm-4-9b-chat
LLaMA-3.1-8B-Instruct	SFT	8B	128k	meta-llama/Meta-Llama-3.1-8B-Instruct
LLaMA-3.1-8B	Pretrain	8B	128k	meta-llama/Meta-Llama-3.1-8B
Mistral-Nemo-Instruct-2407	SFT	12B	128k	mistralai/Mistral-Nemo-Instruct-2407
Phi-3-mini-128k-instruct	SFT	3.8B	128k	microsoft/Phi-3-mini-128k-instruct
Mistral-7B-Instruct-v0.3	SFT	7B	32k	mistralai/Mistral-7B-Instruct-v0.3
Mistral-7B-v0.3	Pretrain	7B	32k	mistralai/Mistral-7B-v0.3
LLaMA-2-7B-32K	Pretrain	7B	32k	togethercomputer/LLaMA-2-7B-32K
LongChat-7B-v1.5-32k	SFT	7B	32k	lmsys/longchat-7b-v1.5-32k
ChatGLM2-6B-32k	SFT	6B	32k	THUDM/chatglm2-6b-32k
LongAlpaca-7B	Pretrain	7B	32k	Yukang/LongAlpaca-7B
LLaMA-3-8B-Instruct	SFT	8B	8k	meta-llama/Meta-Llama-3-8B-Instruct
T-lite-instruct-0.1	SFT	8B	8k	AnatoliiPotapov/T-lite-instruct-0.1
Saiga-LLaMA-3-8B	SFT	8B	8k	IlyaGusev/saiga_llama3_8b
LLaMA-3-8B	Pretrain	8B	8k	meta-llama/Meta-Llama-3-8B
Mistral-7B-v0.1	Pretrain	7B	8k	mistralai/Mistral-7B-v0.1

Table 9: The models evaluated as baselines. **Model Type** shows whether the model is a pretrain or an SFT. **Parameters** indicate the number of model parameters in Billions. **Max Context Length** shows maximal context lengths in tokens. **Model HuggingFace Link** provides the model link on HuggingFace Hub for the open-source models.

Question: {*input*}
Answer:

ru2WikiMultihopQA: *The answer to the question is based on the above excerpts.*
Context: {*context*}
Answer the question briefly, based on the above excerpts.
Question: {*input*}
Answer:

ruBABILongQA1: *I'm giving you a context with facts about the location of different people. You need to answer the question based only on information obtained from the facts. If the person was in different places, use the last location to answer the question.*
Context: {*context*}
Answer the question as briefly as possible.
Question: {*input*}
Answer:

ruBABILongQA2: *I'm giving you a context with facts about the location and actions of different people. You need to answer the question based only on factual information. If a person took an item in one place and went to another, that item is also in the second place. If a person leaves an item in the first place and moves to the second place, the item remains in the first place.*
Context: {*context*}
Answer the question as briefly as possible.
Question: {*input*}
Answer:

ruBABILongQA3: *I'm giving you a context with facts about the location and actions of different people. You need to answer the question based only on factual information. If a person took an item in one place and went to another, that item is also in the second place. If a person leaves an item in the first place and moves to the second place, the item remains in the first place.*
Context: {*context*}
Answer the question as briefly as possible.
Question: {*input*}
Answer:

ruBABILongQA4: *I'm giving you a context with facts about the location and actions of different people. You need to answer the question based only on factual information.*
Context: {*context*}
Answer the question as briefly as possible.
Question: {*input*}
Answer:

ruBABILongQA5: *I'm giving you a context with facts about the location and actions of different people. You need to answer the question based only on factual information.*
Context: {*context*}
Answer the question as briefly as possible.
Question: {*input*}
Answer:

ruSciPassageCount: *Below are a few paragraphs. Read them and determine the number of unique paragraphs.*
Context: {*context*}
Determine the number of unique paragraphs. The answer must contain only one number.
Question: {*input*}

Answer:

ruQasper: *You are provided with a scientific article and a question.*
Context: {*context*}
Answer the question as briefly as possible, using a single phrase or sentence if possible. Don't give any explanations.
Question: {*input*}
Answer:

ruGSM100: *Examples of mathematical problems are given below. Think step by step and answer the question.*

Context: {*context*}

Think step by step and answer the question.

Question: {*input*}

Answer:

C Detailed Model Information

We evaluate 17 popular LLMs, including GPT-4o[26]. All models except for GPT-4o are open-source. The baseline models and their specifics are presented in Table 9.

[26] https://chatgpt.com/