

**Ministry of Science and Higher
Education of the Russian Federation
ITMO University**

Faculty of Digital Transformation

Educational program: Big Data and Machine Learning

REPORT
on research internship

Task topic: Benchmark for assessing the quality of language models when working
with the Russian language

Student: Abdurakhimov Muslimbek Abdulboqi ogli, J1433C

Head of Practice from ITMO University: Khodorchenko Maria Andreevna, Senior
Researcher at research center "Strong artificial intelligence in industry"

Practice completed with grade _____

Commission member signatures:

fullname
(signature)

fullname
(signature)

fullname
(signature)

Date _____

St. Petersburg

2023

ESSAY

Abdurakhimov M.A. Benchmark for assessing the quality of language models when working with the Russian language

Report contains 30 pages, 9 figures, 3 table, 29 sources.

Keywords: Language Models, Evaluation Metrics, Model evaluation, natural language understanding, benchmarks, NLP models, language modelling, general language understanding evaluation.

The purpose of the work of benchmarking work for assessing the quality of language models when working with the Russian language is to establish a standardized and comprehensive framework for evaluating the performance and effectiveness of language models in the context of Russian linguistic nuances. This endeavor aims to provide a systematic and reliable method for comparing different language models, identifying their strengths and weaknesses, and fostering improvements in their applicability to the Russian language.

Structure: The thesis consists of an introduction, three chapters, a conclusion, and a list of sources used.

TABLE OF CONTENTS

ESSAY	2
TERMS AND DEFINITIONS.....	4
1. INTRODUCTION	5
1.1. Background.....	5
1.2. Problem Statement and motivation	5
1 LITERATURE REVIEW	7
1.1. Significance of Benchmarking	7
1.2. Selected benchmarks to analyze.....	8
2 TASK CLASSIFICATION AND COMPARATIVE ANALYSIS.....	14
2.1. Task classification of the benchmarks.....	14
2.1.1 HELM:	14
2.1.2. BIG- bench.....	14
2.1.3 MERA	15
2.1.4. Im – evaluation- harness	16
2.1.5. HuggingFace LLM Leaderboard	16
2.1.6. Russian SuperGLUE.....	17
2.1.7. RuSentEval.....	18
2.2. Comparative analysis.....	19
2.3. Results and discussion	20
3 EVALUATION METRICS	22
3.1. Accuracy, Recall, Precision, F1 Score	22
3.2. BLEU:.....	23
3.3. ROUGE:	24
3.4. METEOR:.....	24
CONCLUSION.....	26
REFERENCES	28

TERMS AND DEFINITIONS

In this research work report, the following terms are used with their respective definitions:

Benchmarking: The process of evaluating and comparing the performance of language models against established criteria and standards, often involving the use of specific tasks or datasets.

Language Models (LMs): Computational models designed to understand and generate human-like language. In the context of this research, language models refer to algorithms and systems capable of processing and generating text in the Russian language.

Fine-tuning: a strategy in machine learning, where pre-trained model is further trained on a specific task or domain with a smaller task-specific dataset.

Semantic analysis: process of understanding and extracting meaning from text beyond its surface structure. It involves interpreting the semantics, or meaning, of words, phrases, and sentences.

Few-Shot Learning: It involves training a model to generalize from a small set of examples, often as few as one or a few shots.

Transfer Learning: machine learning technique where a model trained on one task is adapted or transferred to a different but related task.

1. INTRODUCTION

In recent years, within the NLP community, there has been a notable increase in research endeavors focusing on enhancing machines capacity for profound language comprehension. In this role benchmarks play a crucial role. In the field of Machine Learning, a benchmark entails a collection of datasets linked to one or more metrics, along with a mechanism for aggregating the performance of various systems. Benchmarks play a crucial role in evaluating the advancements of new methodologies across diverse dimensions and aiding in the selection of optimal systems for practical applications [1].

1.1. Background

The use of language models in tasks such as machine translation, sentiment analysis, and information retrieval has become increasingly prevalent. However, the effectiveness of these models in capturing the intricacies of the Russian language remains a focal point of investigation. As language models are often trained on vast and diverse datasets, their adaptation and performance in the context of Russian linguistic structures and semantics demand specialized attention.

Why benchmarks are so important? Reliable evaluation procedures, as highlighted in numerous studies, are essential for fair comparisons of new methods and systems. Typically, a carefully selected metric is employed to gauge the performance on a given task—such as accuracy for classification or mean-squared error for regression. Benchmarks serve as crucial tools for assessing and benchmarking progress in the ever-evolving field of ML [2].

1.2. Problem Statement and motivation

Despite remarkable strides in natural language processing over the past few decades, machines continue to face challenges in achieving proficiency in aspects such as reasoning and comprehensive understanding. Specifically, in the context of the Russian language, the absence of well-defined benchmarks poses a notable gap in evaluating models. This prompts the exploration of the possibility of introducing new and more intricate tasks to stimulate the development of innovative

modifications in Large Language Models (LLMs). Therefore, this reason serves as a compelling motivation for the report, aiming to conduct a thorough analysis of existing benchmarks for the Russian language.

To successfully analyze and complete the report we have following objectives:

- Conduct a detailed comparison and analysis of existing benchmarks relevant to the Russian language.
- Analyze the evaluation metrics used to measure the performance of language models.
- Drawing meaningful conclusions will help in forming a comprehensive understanding of the current landscape of language models in the Russian language domain.
- Based on the insights gained from the analysis, propose strategic directions and recommendations for the enhancement of language models in the Russian language.

1 LITERATURE REVIEW

1.1. Significance of Benchmarking

Over the past decade, the field of natural language processing (NLP) has witnessed a paradigm shift, largely attributed to the remarkable progress in language models. This literature review aims to explore the evolution of language models, focusing on key advancements, challenges, and emerging trends.

Presently, language models (LMs) and the process of benchmarking play pivotal roles in shaping a wide array of language technologies. However, despite their growing prominence, a clear comprehension of their capabilities, constraints, and associated risks remains incomplete [13]. Moreover, the dynamic evolution of language models and benchmarking practices underscores the urgency of gaining a nuanced understanding of their capabilities, limitations, and potential risks in the ever-changing landscape of language technologies [12]. But, in the last few years The NLP community has consistently expanded its scope of evaluation, transitioning from singular datasets like SQuAD [12] to compact dataset compilations such as SuperGLUE (Wang et al., 2019b), and further to extensive collections like the GPT-3 evaluation suite [25] Eleuther AI LM Harness [25] and BIG-Bench[13].

However, despite growing interest in the field, English remains the main point of research [24]. Specifically, for Russian language for the first time, a benchmark with 9 tasks analogically to the SuperGLUE methodology (Want et al., 2019) was developed from scratch for the Russian language, marking the initiation of the benchmarking era for the Russian Language [15]. Following these developments, several benchmarks have been introduced to the Russian Language NLP community. Notable examples include Russian SUPERGLUE [1], RuSentEval [23], MERA [24]. Despite these advancements, the necessity to continue developing new benchmarks persists following reasons.

Why do we need develop new benchmarks for Russian language? Developing new benchmarks for the Russian language is imperative due to the predominant focus of recent improvements in NLP benchmarks on English. This emphasis leaves the nuances and linguistic characteristics of the Russian language relatively

unaddressed [24]. Furthermore, existing benchmarks, often tailored to specific contexts, may not sufficiently cover a diverse range of domains and NLP tasks. This limitation becomes especially problematic as language understanding tasks grow in complexity [14] [16]. Therefore, the need for new benchmarks arises to ensure a more comprehensive and nuanced evaluation of language models in the Russian language across various domains and contexts.

1.2. Selected benchmarks to analyze

From the array of benchmarks across various fields for the Russian language, we have selected the most current and pivotal benchmarks for our analysis.

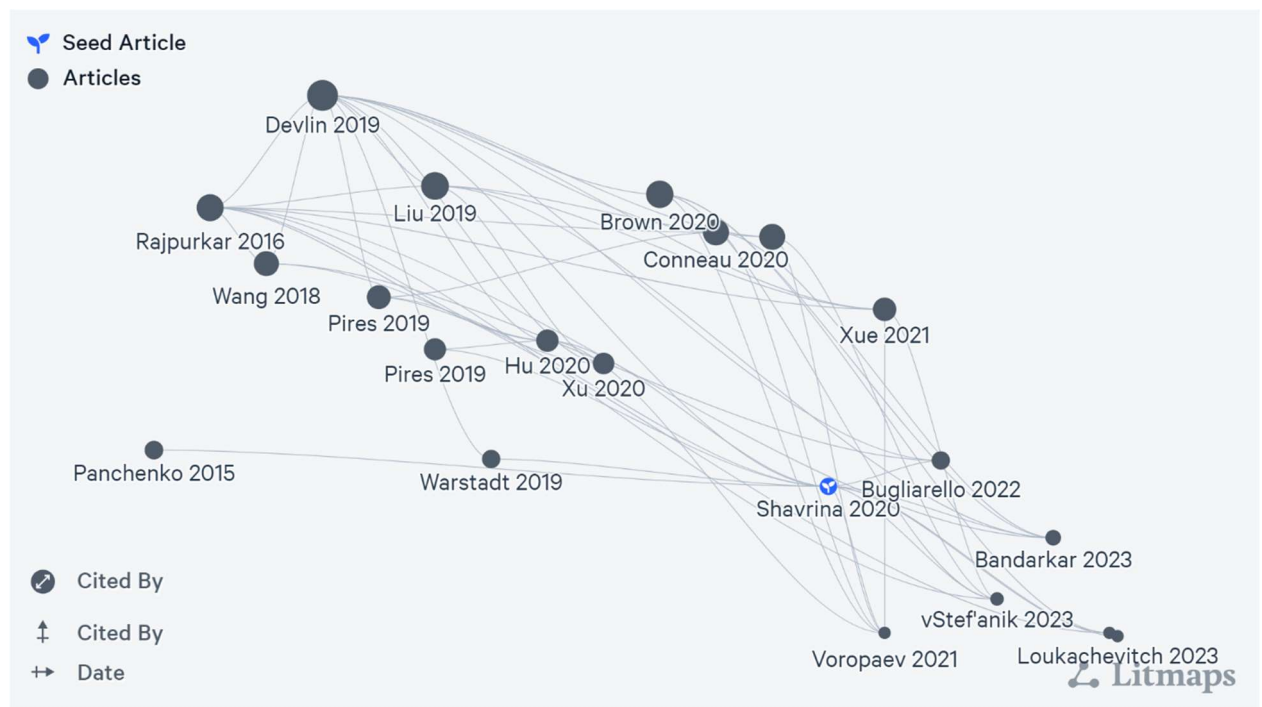


Figure 1. Litmaps of the research paper on Russian SUPERGLUE [7]

The Litmaps visualization (Figure 1) provides a comprehensive overview of the interconnected themes within the existing literature on language model benchmarks. The main point is the Russian SuperGLUE article by Shavrina, T., et al. (2020) [13] and each node represents a benchmark, and the lines indicate relationships and influences. The visualization likely depicts the interconnectedness of RussianSuperGLUE with other benchmarks, highlighting its importance in evaluating language models for the Russian language.

The chosen benchmarks represent a diverse array of linguistic tasks, each contributing uniquely to our comprehensive assessment of language models in the

context of the Russian language. Here, we provide an overview of the selected benchmarks, emphasizing their significance and relevance to our study.

1. HELM (Holistic Evaluation of Language Models):

HELM offers a nuanced approach to evaluating language models beyond traditional metrics. HELM is a versatile language model architecture designed for general-purpose use across a wide spectrum of NLP tasks [12].

One of the distinctive features of HELM lies in its hierarchical evaluation strategy. Rather than relying solely on task-specific assessments, HELM considers the broader applicability and adaptability of language models across various NLP tasks. This approach acknowledges the multifaceted nature of language understanding and generation, emphasizing the importance of a model's ability to generalize its knowledge and skills [12] [17].

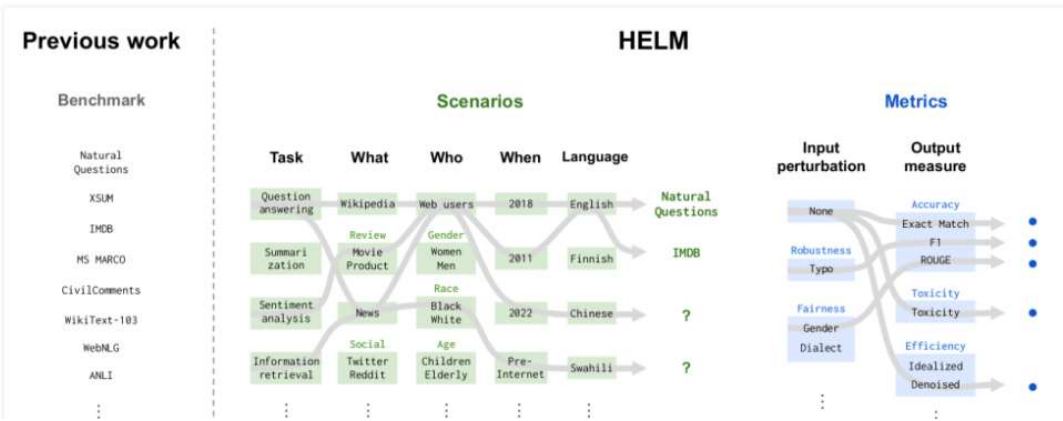


Figure 2. Broad Coverage of HELM [14]

2. Beyond the Imitation Game Benchmark (BIG-bench):

The benchmark was collaboratively developed in an open-source manner on GitHub, where contributors submitted tasks through GitHub pull requests. The proposed tasks underwent peer review through discussions within the pull requests. To encourage contributions, individuals whose tasks were accepted had the opportunity to become co-authors of this paper, serving as an introduction to BIG-bench [15].

The benchmark's objective goes beyond merely quantifying a model's performance in a specific task. Instead, it aspires to achieve more ambitious and meaningful goals by predicting the future capabilities of Large Language Models

(LLMs). The authors emphasize a particular interest in understanding the correlation between the scale of models and their performance, with the overarching aim of anticipating the evolving capabilities of language models [19].

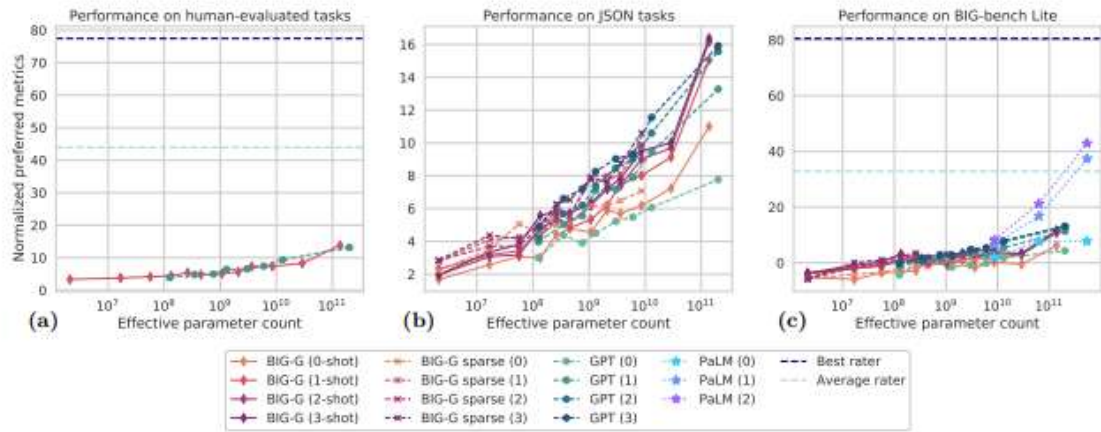


Figure 3. Aggregate performance on BIG- Bench [19]

The given figure is taken from BIG-Bench paper [19] and shows the performance of models. In A subplot, models' combined performance on programmatic and JSON tasks is compared to human rater performance. Despite scale improvements, model performance consistently lags behind human raters. B subplot indicates, models performance on JSON tasks with varying shot values. In C subplot, performance is analyzed across BIG-bench Lite, which is consists of 24 tasks.

3. MERA (Multimodal Evaluation for Russian-language Architectures):

MERA is a comprehensive initiative that spans diverse domains and task types, uniting the collective expertise of industrial companies and academic institutions in Russia. The primary goal is to propel research on large language models [24]. Notably, MERA incorporates human benchmarking, introducing a unique dimension for evaluating models alongside a leaderboard website that fosters healthy competition. Furthermore, the initiative evaluates a spectrum of open models and baselines, such as LLAMA and others, to meticulously gauge their performance. Join us in this collaborative journey that embraces excellence and innovation in language model research.

Лидерборд

Общий скор на лидерборде считается по сумме классов задач + диагностические результаты выдаются отдельно. Внутри каждого сабмита, можно перейти вовнутрь за подробной информацией.

Кликнув на название можно перейти вовнутрь за подробной информацией

Модель, команда	Результат	BPS	CheGeKa	LCS	MathLogicQA	MultiQ	PARus	RCB	ruHumanEval	ruMMLU
1 Human Benchmark MERA	0.887	1	0.719 / 0.645	0.704	0.995	0.928 / 0.91	0.982	0.68 / 0.702	1 / 1 / 1	0.898
2 Mistral-7B-v0.1 Pretrain MERA	0.396	0.391	0.039 / 0	0.1	0.339	0.121 / 0.064	0.516	0.377 / 0.348	0 / 0 / 0	0.68
3 ruGPT3.5_pretrain_baseline MERA	0.208	0.493	0.037 / 0	0.132	0.256	0.114 / 0.036	0.502	0.331 / 0.194	0 / 0 / 0	0.246
4 Random submission MERA	0.203	0.5	0.002 / 0	0.096	0.244	0.014 / 0.001	0.482	0.361 / 0.36	0 / 0 / 0	0.251
5 mGPT 1.3B pretrain MERA	0.198	0.449	0.004 / 0	0.136	0.258	0.055 / 0.014	0.498	0.333 / 0.167	0 / 0 / 0	0.243
6 ruGPT3-large pretrain MERA	0.193	0.415	0.007 / 0	0.122	0.25	0.099 / 0.026	0.498	0.333 / 0.167	0 / 0 / 0	0.245

Figure 4. The Leaderboard of MERA [24]

4. Im – evaluation- harness:

This is an unified framework designed for evaluating generative language models across a diverse spectrum of tasks. Notable features include an extensive collection of over 60 standard academic benchmarks for Large Language Models (LLMs), comprising numerous subtasks and variants [25].

This approach aims to mitigate potential harm when individuals inevitably compare results across various papers, despite discouragement of such practices. Historically, there has also been an emphasis on prioritizing the implementation outlined in the "Language Models are Few Shot Learners" paper [26], aligning with the original goal of directly comparing results with that specific publication.

5. HuggingFace: Open LLM Leaderboard:

In 2018, Hugging Face introduced its groundbreaking Transformers library, making a significant and widely recognized contribution to the AI community. This library showcased influential pre-trained models, such as BERT and GPT, which have since become indispensable tools for a variety of Natural Language Processing (NLP) tasks [21]. Today, this open-source hub serves as a central repository for sharing and accessing pre-trained models in the field of NLP.

In Figure 4, the hub's interface is depicted. Currently, Hugging Face hosts 511 datasets and 1443 models specifically designed for the Russian language. This

positions the Russian language as the 6th most active language on the platform.

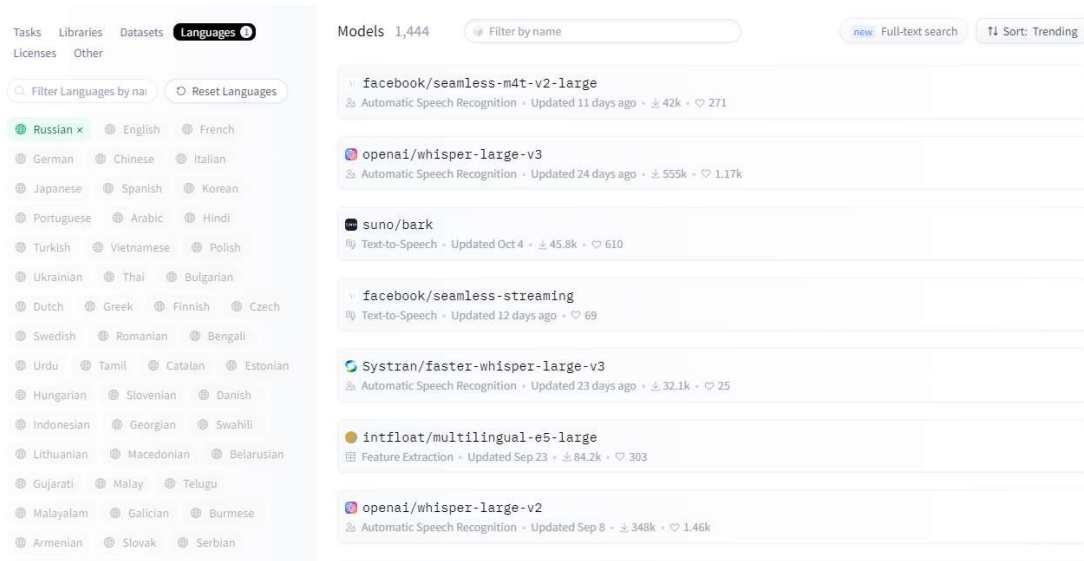


Figure 5. Interface of the hub and models for Russian Language

6. RussianSuperGLUE:

Russian SUPERGLUE is an extension of the well-known SUPERGLUE benchmark adapted for the Russian language. SUPERGLUE, which stands for Stanford Unsupervised and Pre-Trained Ranking GLUE, is a collection of diverse and challenging natural language understanding tasks designed to assess the capabilities of language models comprehensively [13].

While there isn't a direct one-to-one mapping, the corpora utilized in their framework could be deemed as closely related to the specific tasks outlined in the SuperGLUE framework [1].

The adaptation of the SUPERGLUE framework for Russian reflects a commitment to establishing a standardized and challenging evaluation methodology for language understanding across different languages. The extension caters to the unique linguistic characteristics and challenges present in the Russian language, contributing to the broader goal of advancing the field of natural language processing on a global scale [14]. The utilization of corpora that align with the SuperGLUE tasks underscores the intention to maintain consistency in evaluation paradigms while addressing language-specific nuances in Russian SUPERGLUE [1].

7. RuSentEval:

The methodology, adapted from its English counterpart, is tailored to accommodate the nuances specific to Russian. Distinguishing itself from closely related datasets, RuSentEval is meticulously crafted under the guidance of linguistic expertise [24]. This benchmark not only serves as an evaluation framework but also contributes to a deeper understanding of language model performance within the context of Russian linguistic intricacies [21]. Its design ensures a nuanced and comprehensive assessment, making RuSentEval an invaluable resource for advancing natural language understanding in Russian [22].

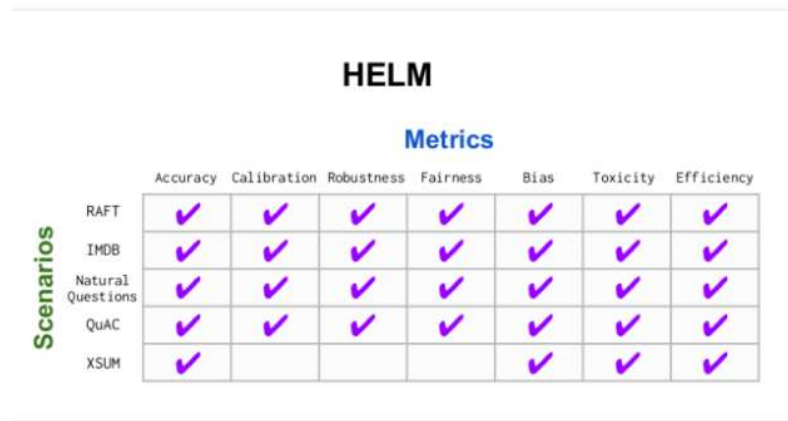
2 TASK CLASSIFICATION AND COMPARATIVE ANALYSIS

2.1. Task classification of the benchmarks

In this section, we delve into the task classification of various benchmarks. Following this classification, we will conduct a comparative analysis to evaluate and contrast these benchmarks.

2.1.1 HELM:

HELM presently incorporates a foundational set of 16 scenarios and encompasses evaluation metrics across 7 distinct categories [12].



The figure shows a table titled 'HELM' with 'Metrics' as the header for the columns and 'Scenarios' as the header for the rows. The metrics are Accuracy, Calibration, Robustness, Fairness, Bias, Toxicity, and Efficiency. The scenarios are RAFT, IMDB, Natural Questions, QuAC, and XSUM. Each cell in the table contains a purple checkmark, indicating that all metrics are evaluated for all scenarios.

	Metrics						
	Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
Scenarios	RAFT	✓	✓	✓	✓	✓	✓
	IMDB	✓	✓	✓	✓	✓	✓
	Natural Questions	✓	✓	✓	✓	✓	✓
	QuAC	✓	✓	✓	✓	✓	✓
	XSUM	✓			✓	✓	✓

Figure 6. Metrics of the HELM [14]

From the figure 3 we can see that HELM adopts multi-metric approach. This prioritization of diverse metrics extends beyond accuracy, enabling and exploration of tradeoffs between different evaluation criteria.

2.1.2. BIG- bench

With over 200 tasks, BIG-bench provides a comprehensive exploration of language models, offering a nuanced understanding of their potential [15].

BIG- bench Lite:

BIG-bench was crafted to encompass an extensive and diverse array of tasks, offering flexibility for arbitrary programmatic tasks. This expansive scope stands out as one of BIG-bench's strengths. However, due to its comprehensive nature, full evaluation can be computationally demanding, especially for programmatic tasks that may entail numerous sequential model calls and pose challenges for certain evaluation pipelines [15].

In response to this challenge, it is curated a subset of 24 tasks to form a streamlined evaluation set known as BIG-bench Lite (BBL). BBL exclusively comprises JSON tasks, providing a more lightweight option for evaluation while maintaining a representative sample of the benchmark's capabilities [15] [16].

auto_debugging	known_unknowns	parsinlu_reading_comprehension
bbq_lite_json	language_identification	play_dialog_same_or_different
code_line_description	linguistics_puzzles	repeat_copy_logic
conceptual_combinations	logic_grid_puzzle	strange_stories
conlang_translation	logical_deduction	strategyqa
emoji_movie	misconceptions_russian	symbol_interpretation
formal_fallacies_...	novel_concepts	vitamin_c_fact_verification
hindu_knowledge	operators	winowhy

Figure 7. The tasks of BIG-bench Lite [15].

2.1.3 MERA

As we talked before the benchmark consists of 21 tasks, which is grouped into four primary categories (some tasks will be considered in the next benchmark task classifications):

1. Semantic Analysis:

CheGeKa: World Knowledge, PARus: Common Sense, ruDetox, ruEthics

2. Reasoning:

MultiQ: Reasoning QA, ruMMLU: Reasoning, ruTiE: Reasoning, Dialogue Context, Memory, RWSD: Reasoning

3. World Knowledge:

ruOpenBookQA: World Knowledge, ruWorldTree: World Knowledge

4. Math and Logic:

BPS: Code, Math, LCS: Code, Math, MathLogicQA: Math + Logic, ruModAr: Math, Logic, ruMultiAr: Math, ruHumanEval: Math, Code, PLP, SimpleAr: Math, USE: Exam

These categorized tasks offer a holistic evaluation framework, covering a spectrum of linguistic, cognitive, and ethical dimensions essential for assessing language models' proficiency.

Name	Modality	Task	Output Format	Class	Metrics	Information
BPS	Code	Code, Math	Open question	Examination	Accuracy	More details
CheGeKa	Text	World Knowledge	Open question	Examination	F1/EM	More details
LCS	Code	Code, Math	Open question	Examination	Accuracy	More details
MathLogicQA	Code	Math+Logic	Select answer	Problem	Accuracy	More details
MultiQ	Text	Reasoning QA	Open question	Problem	F1-score/EM	More details
PARus	Text	Common Sense	Binary classification	Problem	Accuracy	More details
RCB	Text	NLI	Multi-class classification	Problem	Avg. F1/Accuracy	More details
ruDetox	Text	Ethics	Open question	Diagnostic	Toxicity (STA) Content preservation (SIM) Fluency task (FL) $J = J = STA * SIM * FL$	More details
ruEthics	Text	Ethics	Binary classification	Diagnostic	5 MCC	More details
enHateSpeech	Text	Ethics	Binary classification	Diagnostic	Accuracy	More details
ruHHH	Text	Ethics	Binary classification	Diagnostic	Accuracy	More details
ruHumanEval	Text, Code	Math, Code, PLP	Open question	Examination	pass@k	More details
ruMMU	Text	Reasoning	Select answer	Examination	Accuracy	More details
ruModAr	Mathematics	Math, Logic	Open question	Problem	Accuracy	More details
ruMultiAr	Mathematics	Math	Open question	Problem	Accuracy	More details
ruOpenBookQA	Text	World Knowledge	Select answer	Problem	Avg. F1/Accuracy	More details
ruTIE	Text	Reasoning, Dialogue Context, Memory	Binary classification	Problem	Accuracy	More details
enWorldTree	Text	World Knowledge	Select answer	Problem	Avg. F1/Accuracy	More details
RWSD	Text	Reasoning	Binary classification	Problem	Accuracy	More details
SimpleAr	Mathematics	Math	Open question	Problem	Accuracy	More details
USE	Text	Exam	Multi-class classification, Open question, Multiple choice	Examination	Grade Norm	More details

Figure 8. The list of Tasks of MERA

2.1.4. Im – evaluation- harness

The IM- Evaluation Harness also features a diverse range of tasks, showcasing its flexibility and adaptability for evaluating language models. Some key tasks include sentiment analysis, named entity recognition, question answering, text summarization, machine translation, paraphrase detection, textual entailment, text classification, coreference resolution, and document clustering. These tasks cover essential aspects of natural language understanding, showcasing the platform's flexibility and adaptability for comprehensive model evaluation.

2.1.5. HuggingFace LLM Leaderboard

The HuggingFace LLM Leaderboard is a significant platform that hosts a variety of benchmarks for evaluating large language models (LLMs). It serves as a central hub for assessing model performance across different tasks and benchmarks submitted by the community. The tasks featured on this leaderboard cover a wide spectrum, including but not limited to sentiment analysis, named entity recognition, question answering, and text classification. The platform is recognized for its accessibility, providing pre-trained models and a collaborative space for researchers and practitioners to contribute and compare their models.

2.1.6. Russian SuperGLUE

Russian SuperGLUE benchmark includes a set of Natural Language Understanding (NLU) tasks [1]. Specifically, the benchmark is organized into 5 groups comprising nine tasks:

- Textual Entailment & NLI: TERRa, RCB, LiDiRus;
- Common Sense: RUSSE, PARus [8];
- World Knowledge: DaNetQA [9];
- Machine Reading: MuSeRC, RuCoS [10];
- Reasoning: RWSD.

Task	Task Type	Task Metric	Train	Val	Test
TERRa	NLI	Accuracy	2616	307	3198
RCB	NLI	Avg. F1 / Accuracy	438	220	438
LiDiRus	NLI & diagnostics	MCC	0	0	1104
RUSSE	Common Sense	Accuracy	19845	8508	18892
PARus	Common Sense	Accuracy	400	100	500
DaNetQA	World Knowledge	Accuracy	1749	821	805
MuSeRC	Machine Reading	F1 / EM	500	100	322
RuCoS	Machine Reading	F1 / EM	72193	7 577	7257
RWSD	Reasoning	Accuracy	606	204	154

Table 1: Russian SuperGLUE task description [1]

In this table given the information about the tasks by the Benchmark Russian SuperGLUE.

The TERRA dataset focuses on binary textual entailment recognition in Russian. It requires determining whether the meaning of a hypothesis is entailed from a given premise [9].

The RCB dataset is a 3-way classification task for textual entailment in Russian, allowing the premise to represent a textual segment rather than a single sentence.

LiDiRus is a diagnostic set evaluating language models on 33 linguistic features, commonsense, and world knowledge [8].

RUSSE is a binary classification task involving word sense disambiguation, requiring models to recognize if an ambiguous word used in the same meaning in two sentences [11].

PARus is a binary classification task for identifying the most plausible alternative for a given premise, translated from the COPA dataset [8].

DaNetQA is a Russian yes/no question answering dataset following the BoolQ design. MuSeRC is a machine reading comprehension task where models choose correct answers for multi-hop questions.

RuCoS [10] is an MRC task involving commonsense reasoning, and RWSD is a binary coreference resolution task translated from the Winograd Schema Challenge.

These diverse datasets aim to assess and enhance the capabilities of language models in various linguistic tasks in Russian language.

2.1.7. RuSentEval

The introduction highlights the success of Transformer language models in NLP tasks, emphasizing their state-of-the-art performance and cross-lingual capabilities. Probing tasks are introduced as a method to analyze the linguistic properties encoded in intermediate representations of these models [22].

The research employed a set of probing tasks categorized into Surface Properties, Syntactic Properties, and Semantic Properties tasks to evaluate language models. In the Surface Properties category, tasks like SentLen involved predicting the number of tokens in a sentence, while WC focused on inferring information about original words through a 1k-way classification. Syntactic Properties tasks included ConjType, classifying sentences based on complex clause connections, and ImpersonalSent, determining if a sentence lacked a grammatical subject [21].

Tasks like TreeDepth probed the knowledge about hierarchical and syntactic structure, and Gapping aimed to detect syntactic gapping in coordinated structures. Moving to Semantic Properties, tasks like SubjNumber and SubjGender focused on probing number and gender features of the subject, while ObjNumber and ObjGender did the same for the direct object. Predicate Voice (PV), Predicate Aspect (PA), and Predicate Tense (PT) probed for morphosyntactic features of the predicate or head of a predicative construction [21]. These tasks collectively provided a comprehensive assessment of the language models across various linguistic

properties [22].

Probing Task	Language	M-BERT	LABSE	XLM-R	MiniLM	M-BART
Nshift	Ru	84.8 [8]	82.6 [5]	86.9 [9]	80.5 [9]	78.6 [12]
	En	81.8 [10]	84.4 [5]	85.7 [10]	79.3 [8]	83.8 [12]
ObjNumber	Ru	82.8 [6]	82.5 [2]	83.7 [10]	77.8 [10]	81.5 [7]
	En	86.2 [6]	85.4 [3]	86.0 [8]	85.2 [6]	85.9 [9]
SentLen	Ru	91.3 [2]	93.3 [1]	94.5 [2]	94.1 [2]	96.2 [4]
	En	96.3 [2]	96.6 [1]	95.8 [2]	96.1 [3]	97.3 [3]
SubjNumber	Ru	90.5 [7]	92.9 [3]	94.9 [11]	94.2 [12]	93.1 [10]
	En	87.8 [7]	90.7 [12]	86.9 [10]	85.6 [6]	87.3 [9]
Tense	Ru	99.5 [8]	99.8 [5]	99.8 [5]	98.2 [7]	99.6 [7]
	En	88.9 [8]	88.8 [6]	88.8 [9]	87.3 [5]	89.1 [9]
TreeDepth	Ru	44.7 [6]	46.1 [4]	46.5 [5]	44.8 [7]	45.8 [11]
	En	41.2 [5]	42.7 [5]	41.8 [7]	40.9 [7]	41.2 [12]
WC	Ru	84.8 [2]	85.8 [1]	82.6 [1]	72.8 [1]	88.0 [1]
	En	92.6 [1]	93.7 [1]	89.8 [1]	82.3 [1]	93.8 [1]

Table 2. Results of Logistic Regression Classifier for the probing tasks [22]

2.2. Comparative analysis

In evaluating language models, diverse benchmarks offer unique strengths and considerations. Helm, can be really useful with diverse metrics system, while BIG bench is useful when we need more task coverage, especially for Russian Language.

After careful consideration and analyze of the given benchmarks from the metrics and task coverage side, we created a comparative table of benchmarks (Table 1).

Benchmark	Task Coverage	Evaluation Metrics	Strength	Weaknesses
Helm	Core set of 16 scenarios, 7 categories of metrics	Precision, Recall	Comprehensive metrics	Lack of specific tasks. Requires deep analysis
BIG- bench	More than 200 tasks	Accuracy, Precision	Diverse Task Coverage	Resource intensive, complexity
MERA	21 tasks (17 base tasks and 4 diagnostic tasks)	Accuracy, f1_macro, human assessment	Open source, leaderboard	Lack of evaluation metrics, and tasks
Im- Evaluation harness	Varying tasks by more than 60 benchmarks	Varying depending on the benchmarks	Task Diversity	-
Russian SUPERGLUE	Textual Entailment Common Sense World Knowledge Machine Reading Reasoning	Precision, Recall	Wide task spectrum	Not cover specialized domains
Open LLM Leaderboard (Hugging Face)	Varying tasks depending on benchmarks submitted by the community	Varying depending on the benchmarks	Easily accessible, Pre- trained models, Hugging Face Hub, Task Diversity	-
RuSentEval	Nshift, ObjNumber, SentLen, SubjNumber, Tense, TreeDepth, WC	f1, accuracy, recall	Standardized Evaluation framework	Limited to Sentiment Analysis

Table 3 – Comparative Table of Benchmarks

2.3. Results and discussion

The diverse set of benchmarks outlined in the evaluation highlights the multifaceted nature of NLP research. Each benchmark brings its unique strengths and weaknesses to the table, catering to different aspects of language understanding and processing. Helm stands out for its comprehensive metrics, providing a holistic view of tasks. However, its weakness in specificity requires researchers to conduct in-depth analyses to draw meaningful insights. This trade-off between comprehensiveness and task granularity is a crucial consideration. BIG-bench impresses with its extensive task coverage, offering a diverse range of challenges [15]. However, its resource-intensive and complex nature poses practical challenges, emphasizing the need for sufficient computational resources and expertise. MERA's open-source nature and the inclusion of a leaderboard enhance accessibility, fostering collaboration and healthy competition among researchers. [21] However, the absence of standardized evaluation metrics and tasks may hinder comparability

and benchmarking across different systems. While Im-Evaluation Harness boasts task diversity, the lack of explicit details limits our understanding. Clarity on the specific tasks and evaluation metrics would be crucial for researchers considering this benchmark.

Russian SUPERGLUE covers a broad spectrum of tasks, offering a comprehensive evaluation of language understanding [1] [13]. However, its limitation in addressing specialized domains might be a drawback for researchers working in niche areas. Hugging Face's Leaderboard stands out for its accessibility, diverse tasks, and the availability of pre-trained models [28]. However, the lack of specificity in the provided details might leave researchers seeking more transparency. RuSentEval's strength lies in its standardized evaluation framework, ensuring consistency in the assessment of sentiment analysis tasks. However, its focus on sentiment analysis might limit its applicability to a broader range of NLP tasks.

Directions for Improvement. The morphological complexity of the Russian language poses a unique challenge. Future benchmarks could include tasks that specifically assess a model's ability to handle morphologically rich languages, addressing issues related to word forms, declensions, and conjugations. In addition, as language models advance, incorporating benchmarks that assess multimodal understanding (text and image or text and audio) could be instrumental. This reflects the evolving nature of communication, where context is often derived from multiple modalities. Also, consideration should be given to expanding benchmarks to include tasks specific to domains like legal, medical, or scientific discourse. Customized benchmarks in these domains would better evaluate language models' applicability across diverse professional contexts.

3 EVALUATION METRICS

In this section, we provide a comprehensive overview of the evaluation metrics employed in our study to assess the performance of language models on various benchmarks. Evaluation metrics are essential tools for quantifying the effectiveness of models across diverse linguistic tasks. We consider a range of metrics to ensure a thorough examination of the models' capabilities. Choice of evaluation metrics is highly task dependent.

3.1. Accuracy, Recall, Precision, F1 Score

Accuracy, as a performance measure, is straightforward and intuitive, representing the ratio of correctly predicted observations to the total number of observations. The common perception is that high accuracy implies an excellent model. While accuracy is indeed a valuable metric, its effectiveness is most pronounced in datasets where false positive and false negative values are approximately equal, creating a symmetrical distribution [8].

$$Accuracy = \frac{\textit{Number of Correct Predictions}}{\textit{Total number of Predictions}}$$

Recall (Sensitivity) reflects the completeness of a model, assessing its capability to identify all relevant instances (True Positive Rate). When the priority is minimizing False Negatives, Recall becomes crucial. A recall of 1.0 signifies the model produces no false negatives [7].

$$Recall = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Negatives}}$$

Precision measures the exactness of a model, emphasizing its ability to return only relevant instances. In scenarios where minimizing False Positives is critical, Precision becomes a key metric. A perfect precision of 1.0 indicates the model produces no false positives [7].

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

The F1 score serves as the harmonic mean of precision and recall, offering a balanced metric when there's an uneven class distribution. It is particularly useful in scenarios where achieving a balance between precision and recall is essential. The F1 score is defined as:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

In summary, precision is favored when avoiding false positives is critical, recall is crucial when minimizing false negatives is essential, F1 score is suitable for balanced evaluation, and accuracy is appropriate when the class distribution is even.

3.2. BLEU

BLEU (Bilingual Evaluation Understudy) is a cost-effective and swift metric to compute, demonstrating a strong correlation with human evaluation and showcasing language independence. It represented on a scale from 0 to 1, but often expressed as percentage. The higher the percentage, the better the result. This is the gradient used for interpretability scale for the BLEU evaluation metrics.



Figure 9. BLEU interpretability scale [17]

The score less than 10 % is almost useless to understand, while mode then 60 % is often better than human. The basic BLEU metric formula can be expressed as following:

$$BLEU = P_B * exp(\sum_{n=0}^n W_n * \log^* p_n),$$

Where p_n is an n-gram precision that uses n-grams up to length N and positive

weights w_n that sum to one, and P_B is the brevity penalty [17]. BLEU is widely used, but it has several limitations:

- **Lack of Consideration for Meaning:** BLEU primarily relies on n-gram overlap, which doesn't directly capture the semantic or contextual meaning of translated text [17].
- **Limited Consideration for Sentence Structure:** The metric doesn't directly account for the overall structure of sentences.
- **Challenge with Morphologically Rich Languages:** Morphologically rich languages, where words can have various forms and inflections, pose a challenge for BLEU. The metric may struggle to appropriately evaluate translations in languages with complex morphological structures [13].
- **Inability to handle synonyms:** Synonyms or alternative expressions that convey similar meanings may not be adequately considered.

Overall, BLEU is valuable for its simplicity and efficiency, but its limitations make it less suitable for capturing the nuanced aspects of translation quality.

3.3. ROUGE

ROUGE (Recall – Oriented Understudy for Gisting Evaluation) is a set of metrics used for the automatic evaluation of machine-generated text, particularly in the context summarization and document summarization. This metric comprises several variants, including ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. ROUGE-N, akin to BLEU-N, assesses the matching of n -grams between the machine-generated hypothesis and human-produced reference summaries [11].

3.4. METEOR

The METEOR (Metric for Evaluation of Translation with Explicit ORdering) metric is an automatic evaluation metric used to assess the quality of machine-generated translations. The basic formula of the METEOR given by Lavie and Agarwal (2007) [18] can be found like this:

$$METEOR = \frac{(1-\beta) * P * R}{R + \beta * P} * (1 - \gamma) [18],$$

Where, P is precision, R is Recall, β is a parameter that balances precision and recall, and γ is a parameter that penalizes word order errors. The actual computation involves additional steps, such as stemming, synonym matching, and considering different levels of n-grams [6]. METEOR excels in evaluating machine-generated translations, offering a thorough assessment. Its strengths lie in a holistic approach that considers unigram precision, recall, stemming, synonymy, and word order [6][9]. But sensitivity to parameter values could result in score variations, and may pose practical challenges.

Building complex metrics to evaluate models. In the context of language model benchmarking, the aggregation of metrics plays a crucial role in determining and ranking on leaderboards. While traditional metrics such as accuracy, precision, and recall provide valuable insights into specific aspects of model performance, their aggregation allows for a more comprehensive evaluation. For example, calculating the average performance across multiple metrics (mean aggregation) provides a balanced assessment. However, it assumes equal importance for each metric, which may not align the priorities of all tasks. On the other hand, converting individual metric scores into ranks (rank aggregation) and then combining these ranks helps mitigate the impact of extreme values. This approach is particularly useful when dealing with datasets of varying complexities.

Moreover, the construction of sophisticated metrics involves a continuous process of refinement and adaptation. Researchers often iterate on these metrics based on evolving insights, task-specific requirements, and advancements in the field. This dynamic nature allows for the continual enhancement of evaluation criteria to keep pace with the evolving landscape of language models and the expanding diversity of NLP tasks.

CONCLUSION

In conclusion, this report embarked on a comprehensive exploration with the primary goal of addressing the gaps in evaluating language models tailored to the Russian language. To achieve this objective, a dual-pronged approach was adopted, involving a detailed comparison and analysis of existing benchmarks alongside an in-depth examination of evaluation metrics for language models.

In the first facet of our investigation, a meticulous analysis of existing benchmarks relevant to the Russian language was conducted. This involved a scrutiny of benchmarks such as Helm, BIG-bench, MERA, Im-Evaluation Harness, Russian SUPERGLUE, Open LLM Leaderboard (Hugging Face), and RuSentEval. Each benchmark was assessed based on its task coverage, evaluation metrics, strengths, and weaknesses. This analysis not only provided a panoramic view of the current benchmarking landscape but also shed light on the specific challenges and strengths associated with evaluating language models in the context of Russian.

The second facet extended our exploration to encompass the analysis of evaluation metrics for language models. This involved a critical examination of the metrics used to assess model performance, such as precision, recall, accuracy, f1_macro, and human assessment. Understanding the intricacies of these metrics is pivotal in comprehensively gauging the effectiveness and proficiency of language models within the Russian language domain.

By successfully achieving these objectives, this report has laid a robust foundation for a nuanced understanding of the Russian language processing landscape. The insights gained from the comparison of benchmarks and analysis of evaluation metrics not only contribute to the current discourse on language models but also offer actionable recommendations for the strategic enhancement of models tailored to the intricacies of the Russian language. As we chart the course for future developments in NLP, this dual-focused approach underscores the significance of both benchmarks and evaluation metrics in advancing the capabilities of language models in the Russian language domain.

While our current efforts have provided a comprehensive analysis of existing benchmarks and evaluation metrics for language models in the Russian language, it's important to acknowledge that there is room for additional research in the future.

REFERENCES

1. Fenogenova, A., Tikhonova, M., Mikhailov, V., et al. (2022). Russian SuperGLUE 1.1: Revising the Lessons Not Learned by Russian NLP models. DOI: 10.28995/2075-7182-2021-20-XX-XX
2. Wang, A., Wang, X., Ji, X., et al. (2023). Assessing and optimizing large language models on spondyloarthritis multi-choice question answering (SpAMCQA): study protocol for a bilingual evaluation benchmark. DOI: 10.21203/rs.3.rs-3625354/v1
3. Panchenko, A., et al. (2018). RUSSE'2018: a shared task on word sense induction for the Russian language. arXiv preprint arXiv:1803.05795
4. Ruder, S. (2021). Challenges and Opportunities in NLP Benchmarking. URL: <https://www.ruder.io/nlp-benchmarking/>
5. Elov, B. B., Khamroeva, Sh. M., Xusainova, Z. Y. (2023). The pipeline processing of NLP. E3S Web of Conferences 413, 03011. DOI: <https://doi.org/10.1051/e3sconf/202341303011>
6. Song, L., Zhang, J., Cheng, L., et al. (2023). NLPBench: Evaluating Large Language Models on Solving NLP Problems.
7. Storks, S., Gao, Q., Chai, J. Y. (2019). Recent Advances in Natural Language Inference: A Survey of Benchmarks. DOI: <https://doi.org/10.48550/arXiv.1904.01172>
8. Iazykova, T., Kapelyushnik, D., Bystrova, O., Kutuzov, A. (2021). Unreasonable Effectiveness of Rule-Based Heuristics in Solving Russian SuperGLUE Tasks. arXiv: 2105.01192v1 [cs.CL]
9. Shavrina, T., Shapovalova, O. (2017). To the methodology of corpus construction for machine learning: "Taiga" syntax tree corpus and parser. Proceedings of "CORPORA-2017" International Conference. 2017.
10. Dagan, I., Glickman, O., Magnini, B. (2005). The pascal recognising textual entailment challenge. Machine Learning Challenges Workshop. Springer, Berlin, Heidelberg.
11. Panchenko, Alexander, et al. "RUSSE'2018: a shared task on

word sense induction for the Russian language.” arXiv preprint arXiv:1803.05795 (2018).

11. Liang, P., et al. (2023). Holistic Evaluation of Language Models. arXiv:2211.09110

12. Stanford University – Human Centered Artificial Intelligence. (2022). Language Models Are Changing AI. We Need to Understand Them. URL: <https://hai.stanford.edu/news/language-models-are-changing-ai-we-need-understand-them>

13. Shavrina, T., et al. (2020). RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. arXiv:2010.15925

14. Srivastava, A., et al. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615

15. Colombo, P., et al. (2022). What are the best systems? New Perspectives on NLP Benchmarking. Advances in Neural Information Processing Systems 35 (NeurIPS 2022) Main Conference Track

16. Wołk, K. (2015). Neural-Based machine translation for the medical text domain. Based on European Medicines Agency leaflet texts. Procedia Computer Science 64:2-9.

17. Lavie, A., & Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 228–231).

18. Wang, Z. (2023). BIG – Bench: The Behemoth Benchmark for LLMs, Explained. URL: <https://deepgram.com/learn/big-bench-llm-benchmark-guide>

19. Wolf, T., et al. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771

20. Ferrer, J. (2023). What is Hugging Face? The AI Community's Open-Source Oasis. URL: <https://www.datacamp.com/tutorial/what-is-hugging-face>

21. Mikhailov, V., Taktasheva, E., Sigdel, E., Artemova, E. (2021). RuSentEval: Linguistic Source, Encoder Force! arXiv:2103.00573v2

22. Conneau, A., Kiela, D. (2018). SentEval: An Evaluation Toolkit for Universal Sentence Representations. arXiv:1803.05449
23. MERA official website. URL: <https://a-ai.ru/>
24. Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.
25. Zhu, L., Guan, M., Chen, J., Lu, Y., Zhang, Y., & Li, J. (2023). MERA: A Multilingual and Multimodal Evaluation Benchmark for NLP. arXiv preprint arXiv:2303.02552.
26. Baru, C., Bhandarkar, M., Curino, C., Danisch, M., Frank, M., Gowda, B. (2015). Discussion of BigBench: A Proposed Industry Standard Performance Benchmark for Big Data. In Lecture Notes in Computer Science (Vol. 8904). https://link.springer.com/chapter/10.1007/978-3-319-15350-6_4
27. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. (2019). "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems." In Advances in Neural Information Processing Systems 32 (NeurIPS 2019).
28. Ethayarajh, K., & Jurafsky, D. (2020). "Utility is in the Eye of the User: A Critique of NLP Leaderboards." arXiv preprint arXiv:2009.13888. DOI: 10.48550/arXiv.2009.13888.
29. Yin, W., Rajani, N. F., Radev, D., Socher, R., & Xiong, C. (2020). "Universal Natural Language Processing with Limited Annotations: Try Few-shot Textual Entailment as a Start." arXiv preprint arXiv:2010.02584. DOI: 10.48550/arXiv.2010.02584.