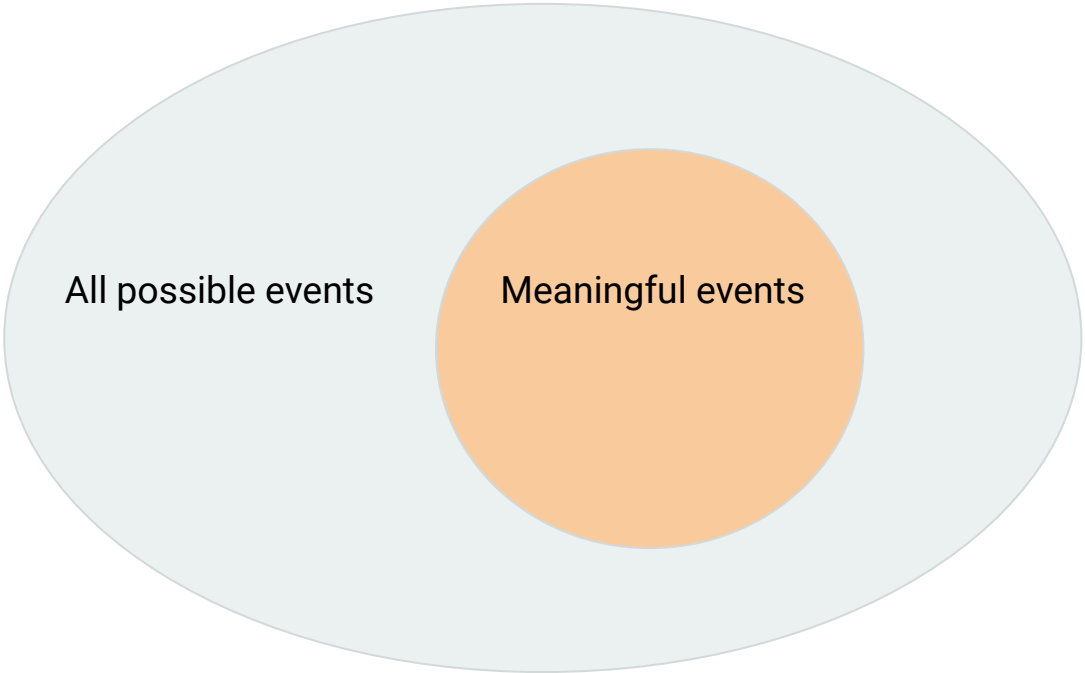
Russian SuperGLUE opportunities and new challenges for NLP benchmarks

Tatiana Shavrina

- AI Research Institute, Moscow, Russia
- SberDevices, Moscow, Russia

24.11.2021 Queen Mary University of London

Language Modelling Task in a Nutshell

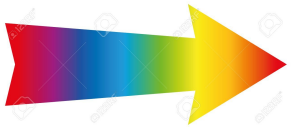


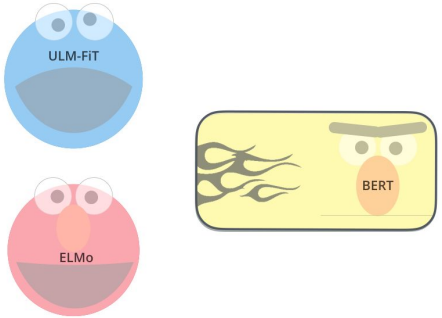
Language Probability Space

LM is Intelligence Distillation

Natural Intelligence

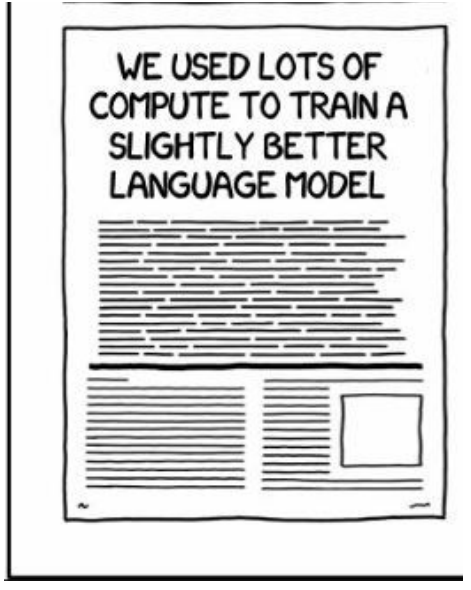
Knowledge Distillation in the Dark





LMs, Transformers

Types of NLP Papers



Efficient Transformers: A Survey
Tay et al. [Google Research]
arxiv.org/abs/2009.06732

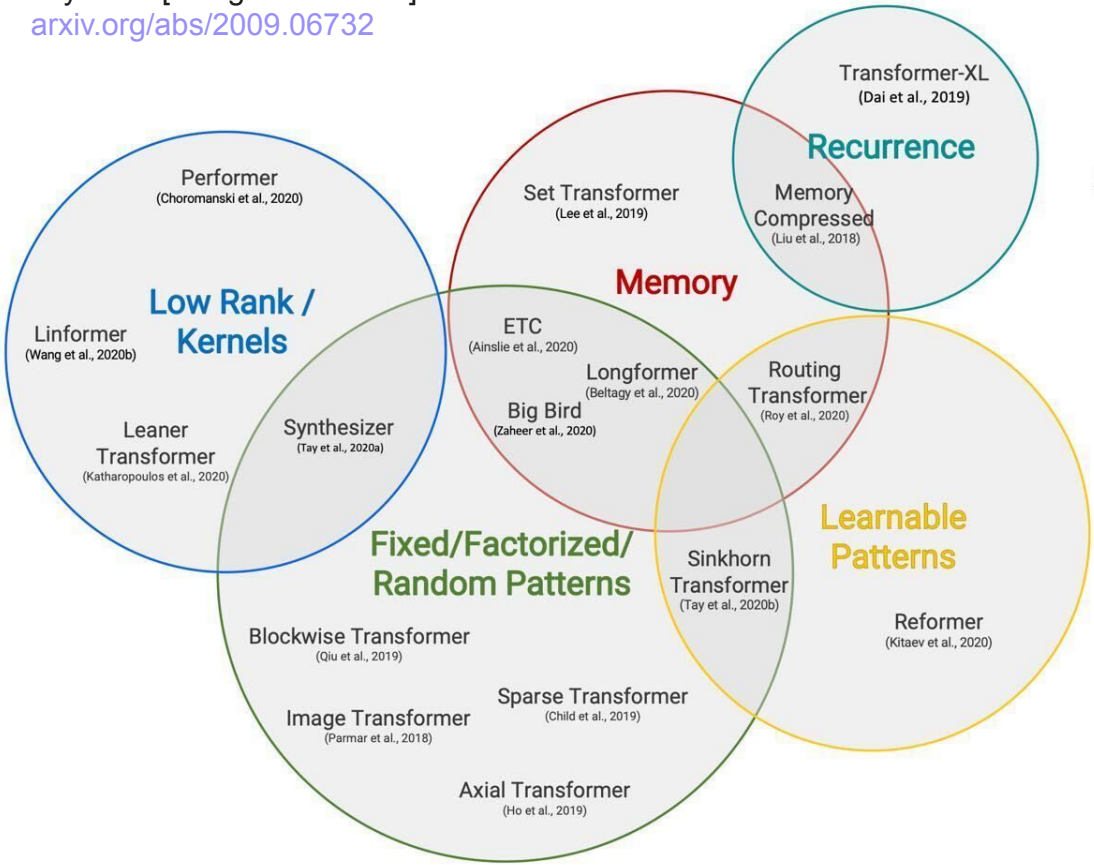
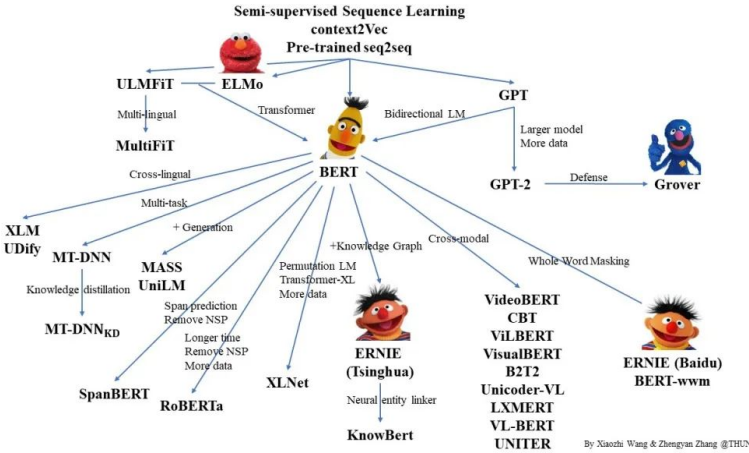


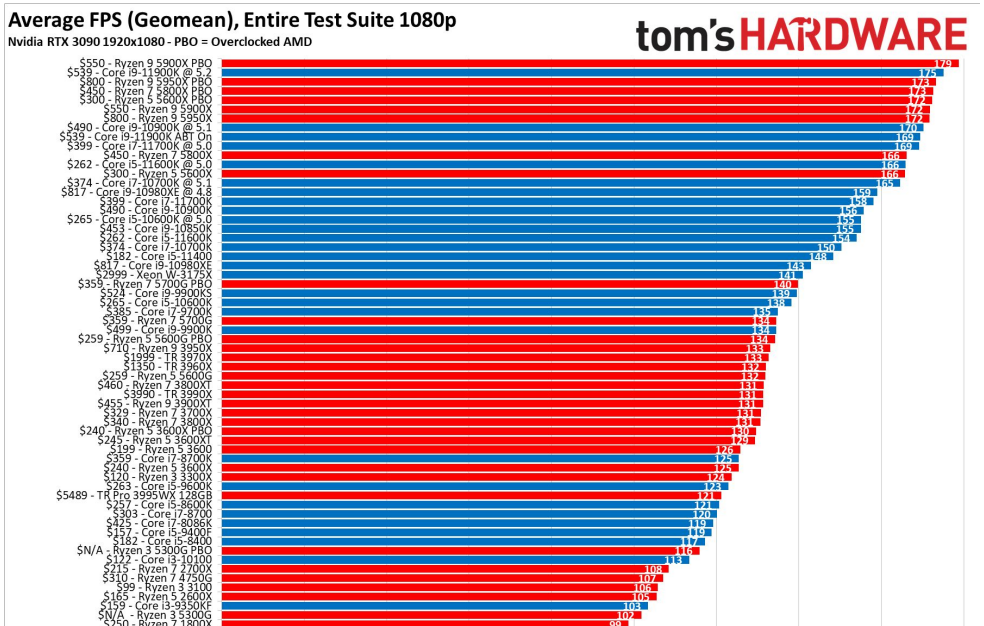
Figure 2: Taxonomy of Efficient Transformer Architectures.

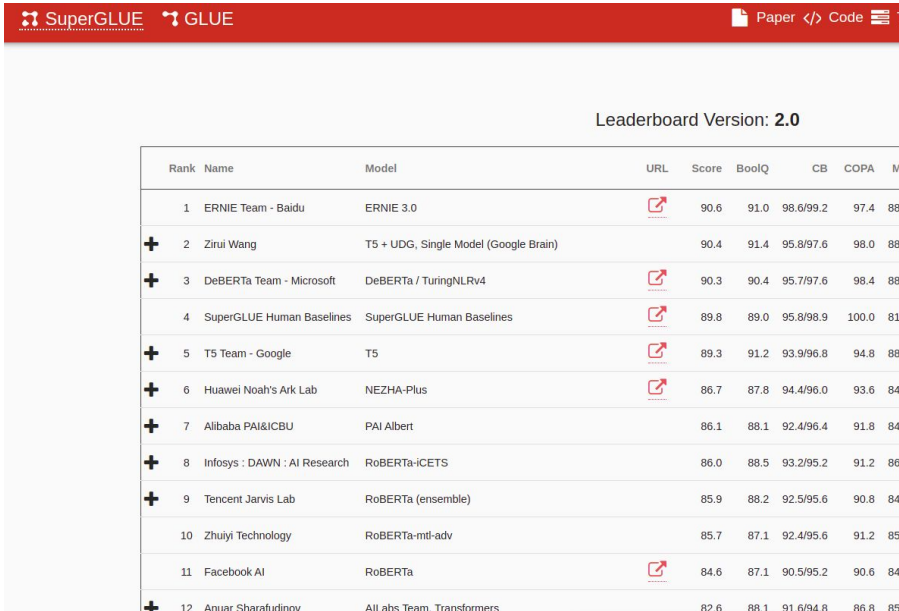


Benchmarking

From computational benchmarks to ML benchmarks

Cumulative score on multiple ML tasks





	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	M
	1	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88
+	2	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88
+	3	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88
	4	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81
+	5	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88
+	6	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84
+	7	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84
+	8	Infosys : DAWN : AI Research	RoBERTa-iCETS		86.0	88.5	93.2/95.2	91.2	86
+	9	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84
	10	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85
	11	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84
+	12	Anuar Sharafudinov	All.abs Team, Transformers		82.6	88.1	91.6/94.8	86.8	85

What is general language understanding evaluation?

The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems.

GLUE consists of

- A benchmark of 9 sentence- or sentence-pair language understanding tasks
- A diagnostic dataset designed to evaluate and analyze model performance with respect to a wide range of linguistic phenomena found in natural language,
- A public leaderboard for tracking performance on the benchmark

SuperGLUE

SuperGLUE GLUE | Paper Code Tasks Leaderboard FAQ Diagnostics Submit Login

Leaderboard Version: **2.0**

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	2	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
	3	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
	4	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
+	5	Huawei Noah's Ark Lab	NEZHA-Large		83.8	85.8	93.3/95.6	91.2	78.7/42.4	87.1/86.4	88.5	73.1	90.4	58.0	87.1/74.4
+	6	Infosys : DAWN : AI Research	RoBERTa-iCETS		77.4	84.7	88.2/91.6	85.8	78.4/37.5	82.9/82.4	83.8	69.1	65.1	35.2	93.8/68.8
	7	IBM Research AI	BERT-mtl		73.5	84.8	89.6/94.0	73.8	73.2/30.5	74.6/74.0	84.1	66.2	61.0	29.6	97.8/57.3
	8	Ben Mann	GPT-3 few-shot - OpenAI		71.8	76.4	52.0/75.6	92.0	75.4/30.5	91.1/90.2	69.0	49.4	80.1	21.1	90.4/55.3
	9	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	38.0	99.4/51.4
			BERT		69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	23.0	97.8/51.7
			Most Frequent Class		47.1	62.3	21.7/48.4	50.0	61.1/0.3	33.4/32.5	50.3	50.0	65.1	0.0	100.0/50.0
			CBoW		44.5	62.2	49.0/71.2	51.6	0.0/0.5	14.0/13.6	49.7	53.1	65.1	-0.4	100.0/50.0

Motivation

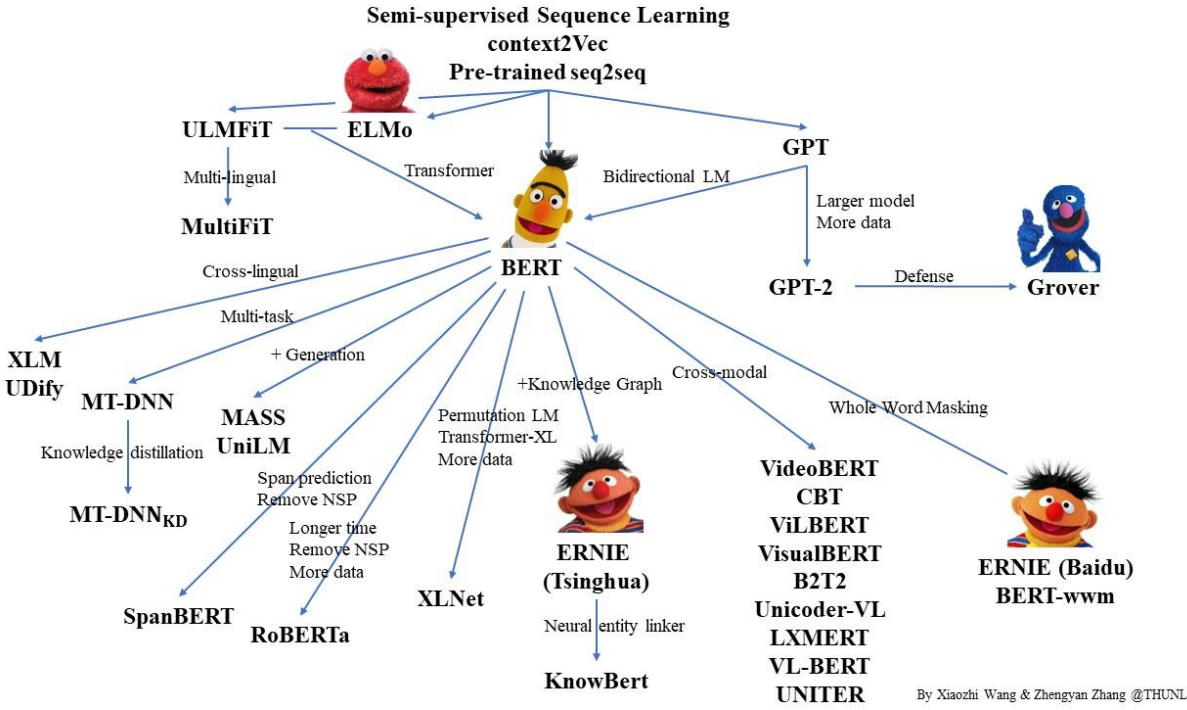
LMs went through the advanced stages of natural language modelling.

Universal transformers show ability to extract complicated relationships from texts.

Development of benchmark approach, testing general intellectual “abilities” in a text format.

Anglo-centric development of machine intelligence.

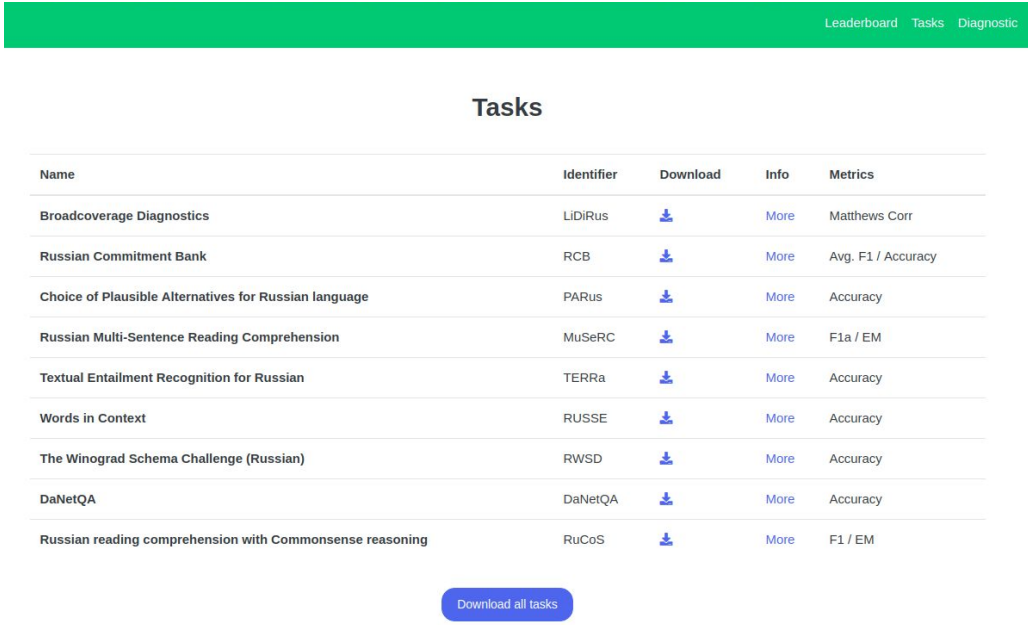
Russian suffers!



Tasks Overview

Six groups of tasks

1. Diagnostics: *LiDiRus*
2. Textual Entailment & NLI: *TERRa, RCB*
3. Common Sense: *RUSSe, PARus*
4. World Knowledge: *DaNetQA*
5. Machine Reading: *MuSeRC, RuCoS*
6. Logic: *RWSD*



Diagnostics: LiDiRus

Linguistic Diagnostic for Russian is a diagnostic dataset covering 33 linguistic phenomena

1. **Lexical Semantics:** lexical entailment, factivity, quantifiers, named entities, symmetry or collectivity, morphological negation, redundancy;
2. **Logic:** negation and double negation, intervals or numbers, upward/downward/non- mono-tone, temporal, conjunction and disjunction, conditionals, universal and existential;
3. **Predicate-Argument Structure:** core args, prepositional phrases, intersectivity, restrictivity, anaphora and coreference, coordination scope, active or passive voice, ellipsis or implicits, nominalization, relative clauses, datives, genitives and partitives;
4. **Knowledge:** common sense, world knowledge

Dataset size

1104 test examples
Translation of SuperGLUE Diagnostic Dataset;

Task

- Used as another “test” for TERRa;
- Linguistic features were preserved and sentences are in one-to-one correspondence.

AIRI

Testing Textual Entailment + linguistic noise

1) Bumblebees do not fly according to the same aerodynamic rules as airplanes.
2) Bumblebees fly more energy efficiently than airplanes.

Tag: knowledge: Common sense

Label: Entailment - No entailment

Testing Textual Entailment + linguistic noise

1) Bumblebees do not fly according to the same aerodynamic rules as airplanes.
2) Bumblebees fly more energy efficiently than airplanes.

Tag: knowledge: Common sense

Label: Entailment - No entailment

Testing Textual Entailment + linguistic noise

Text: The wonder sleigh was designed by the pupils of the House of Youth Technical Creativity. They are based on a tank that the guys made earlier, they only cut off the turret, built a new hull and painted it with Khokhloma - since we face the year of the rooster.

Label: Entailment - No entailment

Testing Textual Entailment + linguistic noise

Text: The wonder sleigh was designed by the pupils of the House of Youth Technical Creativity. They are based on a tank that the guys made earlier, they only cut off the turret, built a new hull and painted it with Khokhloma - since we face the year of the rooster.

Label: Entailment - No entailment

(Khokhloma national ornament casually contains peacocks,

flowers and other natural motifs)

Logic: RWSD

This dataset is a trancreation of Winograd Schema Challenge

Dataset size

606 train/204 val/154 test pairs of sentences

Data source

Editing and translation of the Winograd Schema Challenge

Task

Each sentence has two objects for coreference and segment markup

Example

Text: *The cup does not fit in the brown suitcase because it is too large.*

Coreference: True

Leaderboard

We have improved the datasets. Please, change the leaderboard for the version (1.0/1.1) you are looking for, clicked on the button below.
You can switch between the scores and inference speed leaderboard as well. Click on the button Performance.

Version 1.0 | Performance*

* More information about speed scores and RAM are available here.

Rank	Name	Team	Link	Score	LiDiRus	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQA	RuCoS
1	HUMAN BENCHMARK	AGI NLP		**0.811**	0.626	0.68 / 0.702	0.982	0.806 / 0.42	0.92	0.805	0.84	0.915	0.93 / 0.89
2	Golden Transformer v2.0	Avengers Ensemble		**0.755**	0.515	0.384 / 0.534	0.906	0.936 / 0.804	0.877	0.687	0.643	0.911	0.92 / 0.924
3	YaLM p-tune (3.3B freezed + 40k trainable params)	Yandex		**0.711**	0.364	0.357 / 0.479	0.834	0.892 / 0.707	0.841	0.71	0.669	0.85	0.92 / 0.916
4	ruT5-large finetune	SberDevices		**0.686**	0.32	0.45 / 0.532	0.764	0.855 / 0.608	0.775	0.773	0.669	0.79	0.86 / 0.859
5	ruRoberta-large finetune	SberDevices		**0.684**	0.343	0.357 / 0.518	0.722	0.861 / 0.63	0.801	0.748	0.669	0.82	0.87 / 0.867
6	Golden Transformer v1.0	Avengers Ensemble		**0.679**	0.0	0.406 / 0.546	0.908	0.941 / 0.819	0.871	0.587	0.545	0.917	0.92 / 0.924
7	ruT5-base finetune	Sberdevices		**0.635**	0.267	0.423 / 0.461	0.636	0.808 / 0.475	0.736	0.707	0.669	0.769	0.85 / 0.847
8	ruBert-large finetune	SberDevices		**0.62**	0.235	0.356 / 0.5	0.656	0.778 / 0.436	0.704	0.707	0.669	0.773	0.81 / 0.805
9	ruBert-base finetune	SberDevices		**0.578**	0.224	0.333 / 0.509	0.476	0.742 / 0.399	0.703	0.706	0.669	0.712	0.74 / 0.716
10	YaLM 1.0B few-shot	Yandex		**0.577**	0.124	0.408 / 0.447	0.766	0.673 / 0.364	0.605	0.587	0.669	0.637	0.86 / 0.859
11	RuGPT3XL few-shot	SberDevices		**0.535**	0.096	0.302 / 0.418	0.676	0.74 / 0.546	0.573	0.565	0.649	0.59	0.67 / 0.665
12	RuBERT plain	DeepPavlov		**0.521**	0.191	0.367 / 0.463	0.574	0.711 / 0.324	0.642	0.726	0.669	0.639	0.32 / 0.314
13	SBERT_Large_mt_ru_finetuning	SberDevices		**0.514**	0.218	0.351 / 0.486	0.498	0.642 / 0.319	0.637	0.657	0.675	0.697	0.35 / 0.347

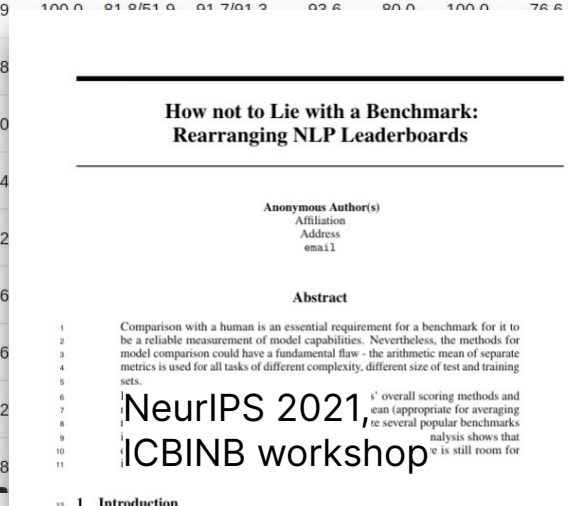
What does the "arms race" lead to?

SuperGLUE GLUE Paper Code Tasks Leaderboard FAQ Diagnostics

Leaderboard Version: **2.0**

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	
1	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7
2	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7
3	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3
4	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9								
5	T5 Team - Google	T5		89.3	91.2	93.9/96.8								
		NEZHA-Plus		86.7	87.8	94.4/96.0								
				86.1	88.1	92.4/96.4								
8	Infosys : DAWN : AI Research	RoBERTa-iCETS		86.0	88.5	93.2/95.2								
9	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6								
10	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6								
11	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2								
12	Anuar Sharafudinov	AILabs Team, Transformers		82.6	88.1	91.6/94.8								

1) instability of the result
2) huge computing power wasted on unreliable tests



How not to Lie with a Benchmark: Rearranging NLP Leaderboards

Anonymous Author(s)
Affiliation
Address
email

Abstract

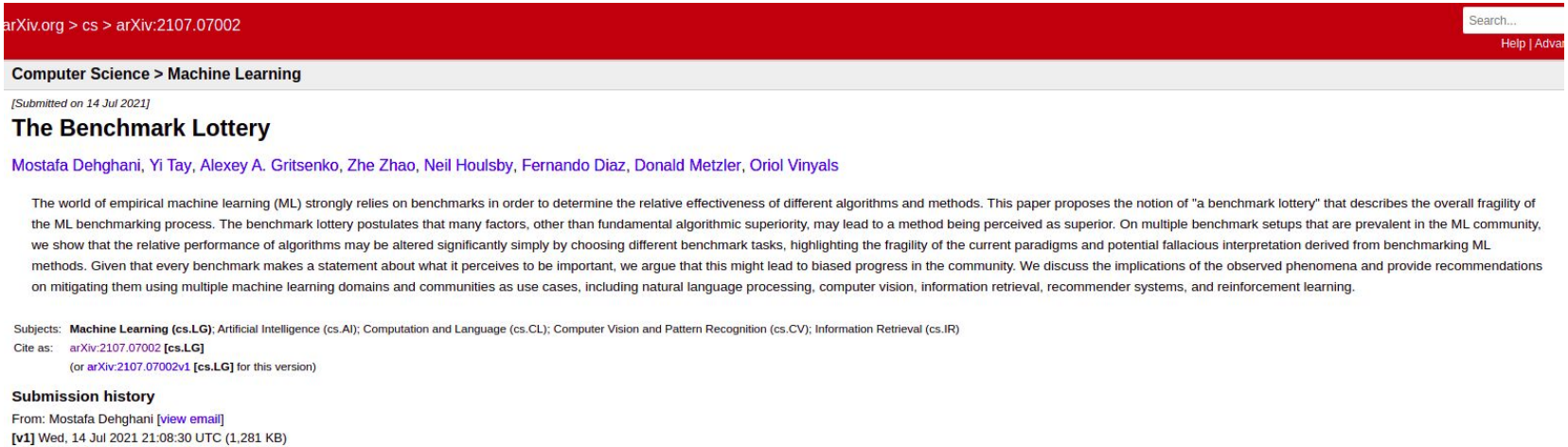
Comparison with a human is an essential requirement for a benchmark for it to be a reliable measurement of model capabilities. Nevertheless, the methods for model comparison could have a fundamental flaw - the arithmetic mean of separate metrics is used for all tasks of different complexity, different size of test and training sets.

NeurIPS 2021, ICBINB workshop

Some problems with benchmarks

We inherit methodological disadvantages:

- one main metric
- small test sets
- “benchmark lottery”



arXiv.org > cs > arXiv:2107.07002

Search...

Help | Advar

Computer Science > Machine Learning

[Submitted on 14 Jul 2021]

The Benchmark Lottery

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, Oriol Vinyals

The world of empirical machine learning (ML) strongly relies on benchmarks in order to determine the relative effectiveness of different algorithms and methods. This paper proposes the notion of "a benchmark lottery" that describes the overall fragility of the ML benchmarking process. The benchmark lottery postulates that many factors, other than fundamental algorithmic superiority, may lead to a method being perceived as superior. On multiple benchmark setups that are prevalent in the ML community, we show that the relative performance of algorithms may be altered significantly simply by choosing different benchmark tasks, highlighting the fragility of the current paradigms and potential fallacious interpretation derived from benchmarking ML methods. Given that every benchmark makes a statement about what it perceives to be important, we argue that this might lead to biased progress in the community. We discuss the implications of the observed phenomena and provide recommendations on mitigating them using multiple machine learning domains and communities as use cases, including natural language processing, computer vision, information retrieval, recommender systems, and reinforcement learning.

Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Computation and Language (cs.CL); Computer Vision and Pattern Recognition (cs.CV); Information Retrieval (cs.IR)

Cite as: arXiv:2107.07002 **[cs.LG]**

(or arXiv:2107.07002v1 **[cs.LG]** for this version)

Submission history

From: Mostafa Dehghani [view email]

[v1] Wed, 14 Jul 2021 21:08:30 UTC (1,281 KB)

Benchmark lottery

task selection bias

$$C_n^k = \frac{n!}{(n-k)! \cdot k!}.$$

the range is unstable if we choose different combinations of tasks
8 from 10 - 45 combinations
7 from 10 - 120
...
and if we use more tasks?

Benchmark lottery

task selection bias

$$C_n^k = \frac{n!}{(n-k)! \cdot k!}.$$

the range is unstable if we choose different combinations of tasks
8 from 10 - 45 combinations
7 from 10 - 120
...
and if we use more tasks?

Table 2: Relative order of different models when selecting different subsets of SuperGLUE. Selecting different subsets of tasks can produce *very different* outcomes for relative ranking of model architectures. Models that did not appear in Top-5 at all are Lightweight Conv, Dynamic Conv and Transparent Attention. For tasks, A=BoolQ, B=CB, C=CoPA, D=MultiRC, E=ReCoRD, F=RTE, G=WiC, H=WSC.

Tasks	Top-5 Performing Models (In Order)
H	Universal, Switch, Adaptive Softmax, Weighted, Vanilla
G	MoE, Switch, Vanilla, Funnel, Universal
A, B	Adaptive Softmax, Vanilla, MoE, Switch, Weighted
A, C	MoE, Switch, Adaptive Softmax, Vanilla, Universal
D, H	Switch, Universal, Adaptive Softmax, MoE, Weighted
B, E, H	Adaptive Softmax, Switch, MoE, Vanilla, Weighted
F, G, H	Switch, MoE, Adaptive Softmax, Universal, Vanilla
A, F, G	MoE, Switch, Vanilla, Adaptive Softmax, Vanilla
C, F, G, H	Switch, MoE, Adaptive Softmax, Vanilla, Universal
A, C, D, G	MoE, Switch, Adaptive Softmax, Vanilla, Universal
All	Switch, MoE, Adaptive Softmax, Vanilla, Universal

natural language understanding
GLUE, SuperGLUE

probing and interpretation
LINSPECTOR, SentEval

simple tasks
DecaNLP

cross-lingual knowledge transfer
XGLUE, XTREME

natural language generation
GEM

ethical biases
HateCheck

robustness
RobustnessGym, AdvGLUE

Complex NLP benchmarks by 2021

AIRI

DecaNLP

Question Answering, Machine Translation, Summarization, Natural Language Inference, Sentiment Analysis, Semantic Role Labeling, Relation Extraction, Goal-Oriented Dialogue, Semantic Parsing, Commonsense Reasoning

Leaderboard

Rank	Model	decaScore	Breakdown by Task			
1 June 20, 2018	MQAN *Salesforce Research*	590.5	SQuAD	74.4	QA-SRL	78.4
			IWSLT	18.6	QA-ZRE	37.6
			CNN/DM	24.3	WOZ	84.8
			MNLI	71.5	WikiSQL	64.8
			SST	87.4	MWSC	48.7

HateCheck

Targeting hateful texts and checking systematic biases in models

Protected Group	Slurs
Women	b*tch, sl*t, wh*re
Trans people	tr*nny, sh*male
Gay people	f*ggot, f*g, q*eer
Black people	n*gger, c*on
Disabled people	r*tard, cr*pple, m*ng
Muslims	m*zzie, J*hadi, camel f*cker
Immigrants	w*tbacks, r*pefugees

Table 5: Hateful slurs in HATECHECK

Target Group	n	B–D	B–F	P	SN
Women	421	*34.9*	52.3	**80.5**	*23.0*
Trans ppl.	421	69.1	69.4	**80.8**	*26.4*
Gay ppl.	421	73.9	74.3	**80.8**	*25.9*
Black ppl.	421	69.8	72.2	**80.5**	*26.6*
Disabled ppl.	421	71.0	*37.1*	**79.8**	*23.0*
Muslims	421	72.2	73.6	**79.6**	*27.6*
Immigrants	421	70.5	58.9	**80.5**	*25.9*

Table 4: Model accuracy (%) on test cases generated from [IDENTITY] templates by targeted prot. group.

GEM

Generative model evaluation

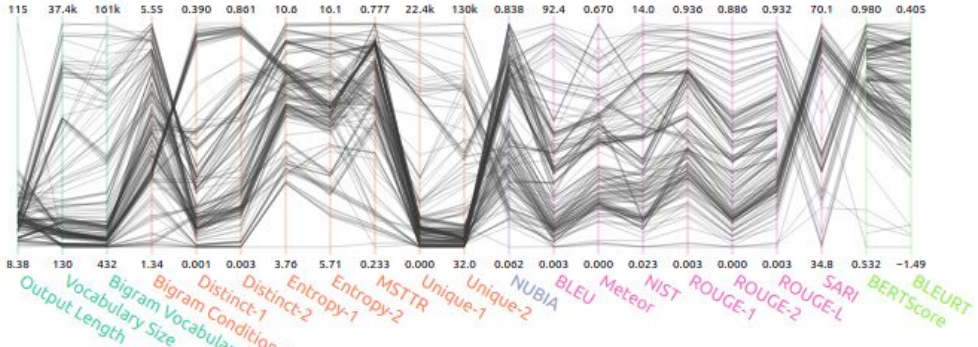
Submissions & Scores

data2text	common_gen_test	ByT5-base (Baseline) ByT5-large (Baseline)
		ByT5-small (Baseline) ByT5-xl (Baseline) T5-base (Baseline)
		T5-large (Baseline) T5-small (Baseline) T5-xl (Baseline)
		BART-base mT5_base mT5_large mT5_small mT5_xl
		NUIG-DSI POINTER t5-small
	cs_restaurants_test	ByT5-base (Baseline) ByT5-large (Baseline)
		ByT5-small (Baseline) ByT5-xl (Baseline) T5-base (Baseline)
		T5-large (Baseline) T5-small (Baseline) T5-xl (Baseline)

Measures

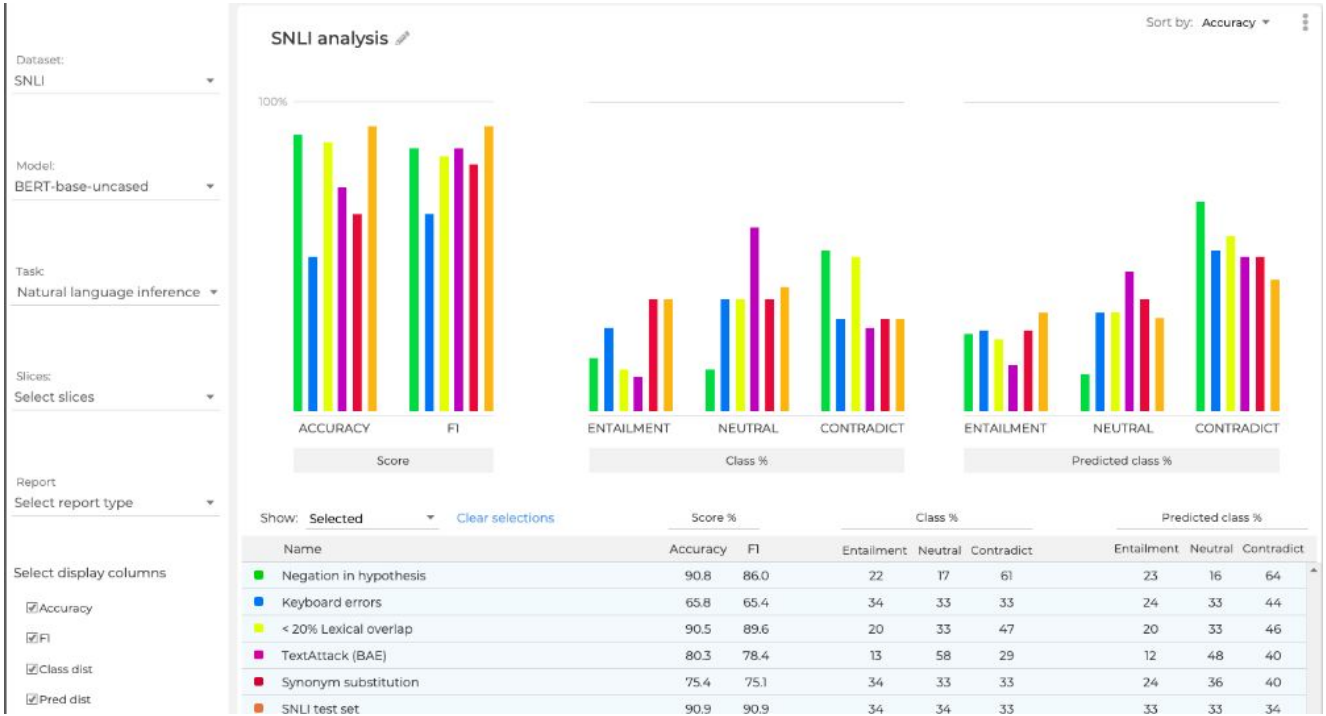
descriptive	Output Length Vocabulary Size Bigram Vocabulary Size
diversity	Bigram Conditional Entropy Distinct-1 Distinct-2 Entropy-1 Entropy-2 MSTTR
	Unique-1 Unique-2
faithful	NUBIA
lexical	BLEU Meteor NIST ROUGE-1 ROUGE-2 ROUGE-L SARI
semantic	BERTScore BLEURT

Visualization



Robustness Gym

Unifying the NLP Evaluation Landscape



4 standard evaluation paradigms:

1. subpopulations,
2. transformations,
3. evaluation sets,
4. adversarial attacks

Benchmark/ task	GLUE 2018	Super GLUE 2019	Deca NLP 2018	Hate check 2020	GEM 2021	*Sent Eval 2018*	*Lin spector 2018*	XGLUE 2020	Xtreme 2020	*Robust ness Gym 2020*	ADV Glue 2021
Intellectual tasks		+								+	+
Multilingual					+		+	+	+		
Leakage-resistant					+					+	+
Many tasks >5	+	+	+		+	+	+			+	+
Different complexity											
Many metrics				+	+	+	+			+	+

Benchmark/ task	GLUE 2018	Super GLUE 2019	Deca NLP 2018	Hate check 2020	GEM 2021	*Sent Eval 2018*	*Lin spector 2018*	XGLUE 2020	Xtreme 2020	*Robust ness Gym 2020*	ADV Glue 2021
Intellectual tasks		+									+
Multilingual					+		+	+	+		
Leakage-resistant					+					+	+
Many tasks >5	+	+	+		+	+	+			+	+
Different complexity											
Many metrics				+	+	+	+			+	+

Take-away points:

1) benchmarking is a key method of evaluating and validating new language models with real-life applications;
2) still there exist an “arms race” with hackathon methods;
3) the current paradigm encourages heavy computational costs of the LMs and allows them to “buy” the first lines of the ratings if more data or training is used;
4) benchmarks should also raise other issues, like computational efficiency, generalization abilities of the models, etc.

A small call for papers - ACL 2022

https://nlp-power.github.io/

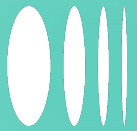
NLP Power! The First Workshop on Efficient Benchmarking in NLP.

- Computational efficiency and energy considerations;
- New practices in measuring linguistic competence in mono- and multilingual benchmarks;
- Critical analysis of existing benchmark evaluation and construction designs;
- Guidelines for reproducibility and reliability of the benchmark results;
- Construction of zero-shot and few-shot mono- and multilingual benchmarks;
- Novel approaches to benchmark evaluation considering task complexity, model architecture, number of parameters and result aggregation;
- Applications of utility theory, voting theory and microeconomics to benchmark evaluation;
- User and application-specific model evaluation;
- New metrics and tasks for computationally lean comparison between models and measuring interpretability;
- Analysis of human and model evaluation strategies in natural language understanding, text generation, knowledge transfer;
- Human evaluation protocols, specifically in the multilingual setting;
- Tracing biases and ethical issues in benchmark datasets and models.

References

- Russian SuperGLUE https://arxiv.org/pdf/2010.15925.pdf
- The Benchmark Lottery https://arxiv.org/abs/2107.07002
- How not to lie with a benchmark https://openreview.net/pdf?id=PPGfoNJnLKd
- ADVGLUE https://openreview.net/forum?id=GF9cSKI3A_q
- RobustnessGYM https://arxiv.org/abs/2101.04840
- SuperGLUE paper https://arxiv.org/abs/1905.00537
- What Will it Take to Fix Benchmarking in Natural Language Understanding? https://arxiv.org/abs/2104.02145
- Sebastian Ruder on benchmarking problems https://ruder.io/nlp-benchmarking/index.html

Thank you!



Artificial Intelligence
Research Institute