Benchmarking Transformer Language Models on Natural Language Understanding Tasks

Vladislav Mikhailov

November 2, 2023

SALUTE DEVICES

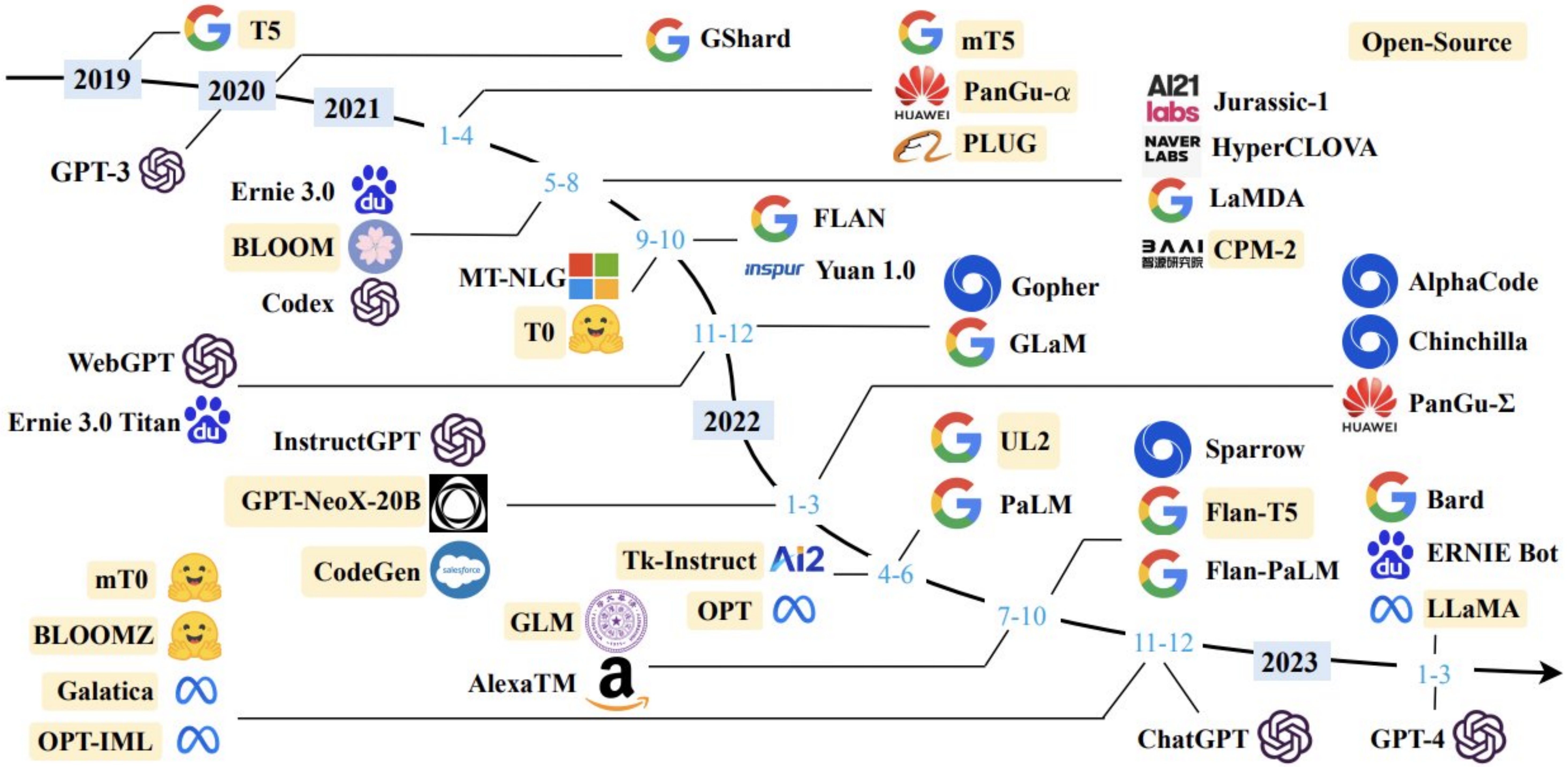
HSE University

Outline

- Motivation and main contributions
- Key results and retrospective
- Other research contributions

Background

The rapid development and proliferation of large language models (LLMs) [5]



Source: a blog post by Brian Wang

Background

- Benchmarking as a standard approach to evaluating LLMs
- Benchmark is a collection of datasets, task-specific metrics, and an aggregation procedure

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
3	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
4	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	92.7/94.7
5	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4

The SuperGLUE public leaderboard [1]

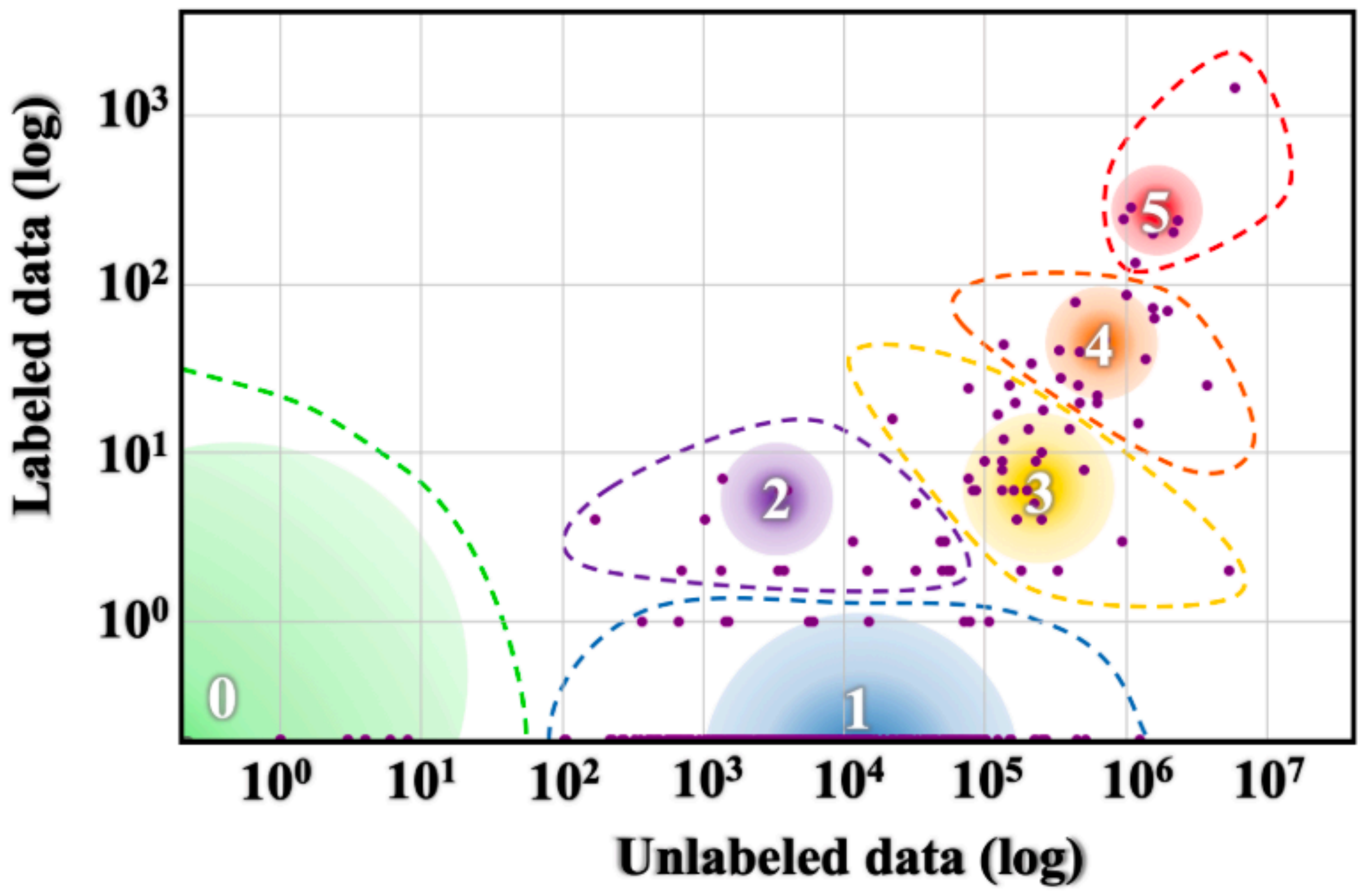
Background

- Benchmarking is becoming more complex:
  - **TURINGBENCH [2]**: the Turing test in natural language generation
  - **BigBench [3]**: more than 100 tasks
  - **HELM [4]**: user-oriented evaluation scenarios
  - **MMLU [5]**: massive multi-task language understanding

Alan Turing sitting on a bench

Low linguistic diversity in NLP

- NLP is generally focused on English



The distribution of resources in the world's languages [6]. The size of the gradient circle represents the number of languages in the class. The color spectrum represents the total speaker population size from low to high.

Low linguistic diversity in NLP

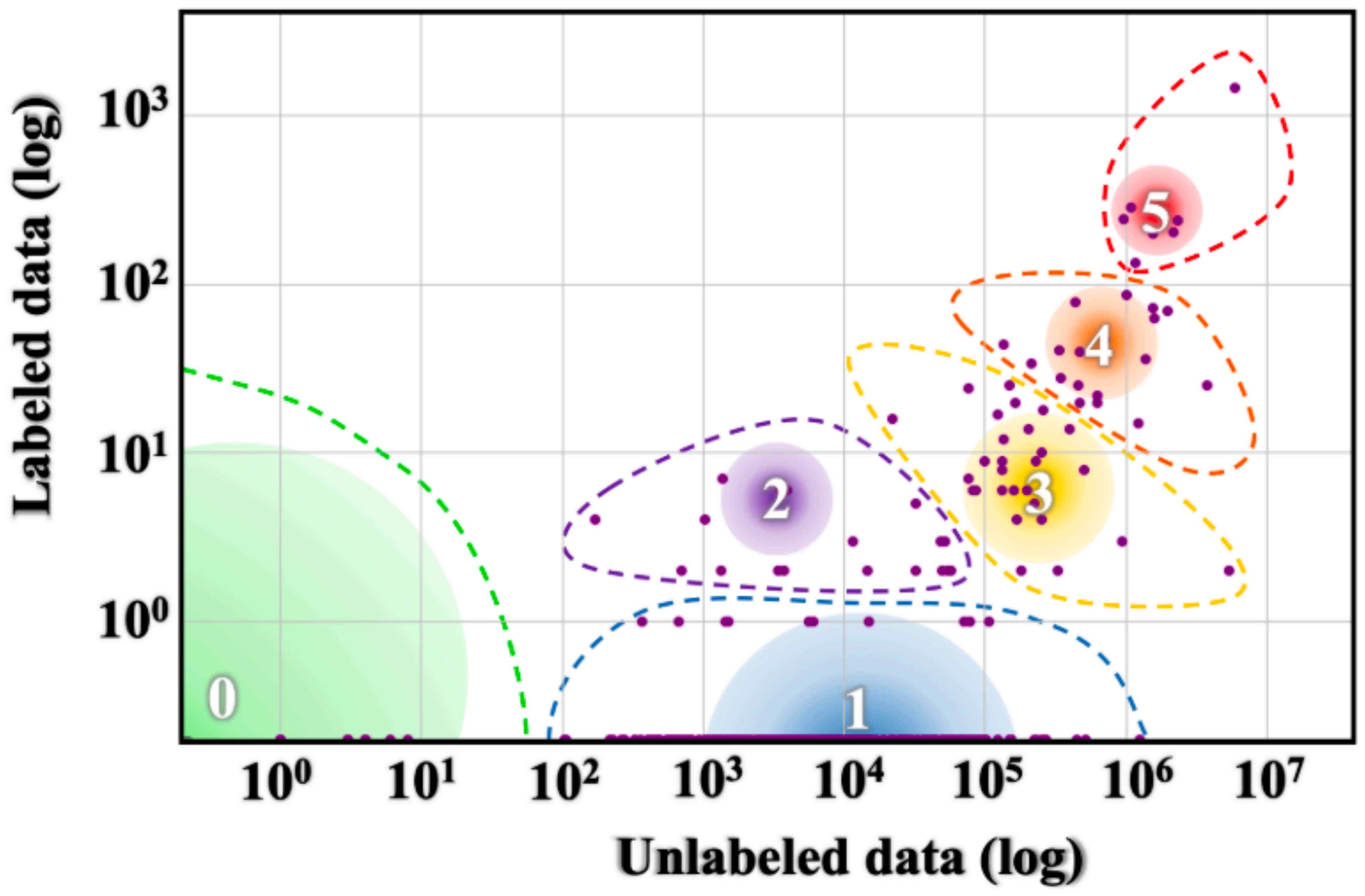
- NLP is generally focused on English

🔍 Cross-lingual benchmarks

XTREME [7]
XGLUE [8]

🔍 Monolingual benchmarks

FLUE [9]
KLEJ [10]



The distribution of resources in the world's languages [6]. The size of the gradient circle represents the number of languages in the class. The color spectrum represents the total speaker population size from low to high.

Low linguistic diversity in NLP

- NLP is generally focused on English

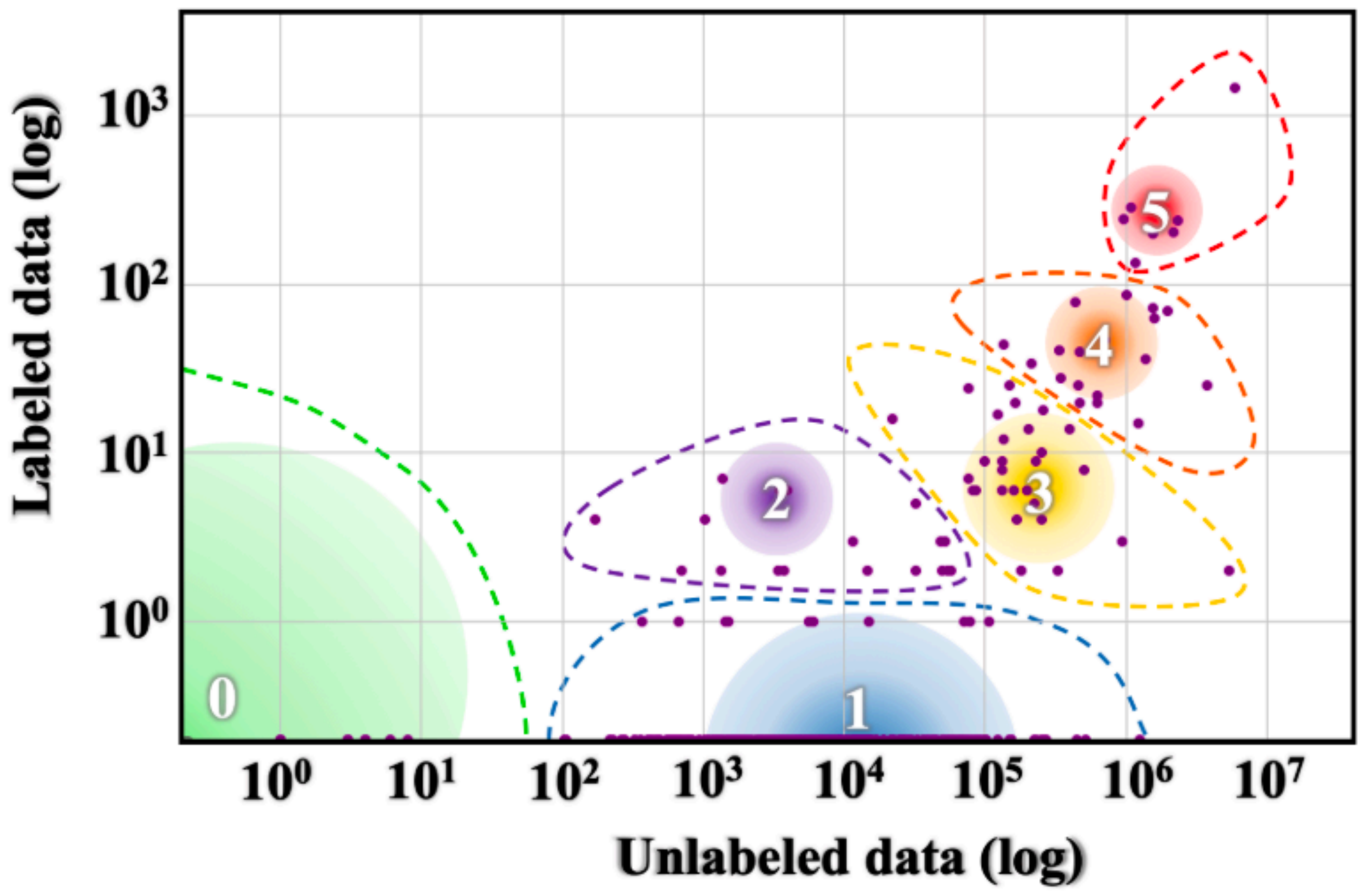
Cross-lingual benchmarks

XTREME [7]
XGLUE [8]

Monolingual benchmarks

FLUE [9]
KLEJ [10]

⁉️ Russian is not well-addressed



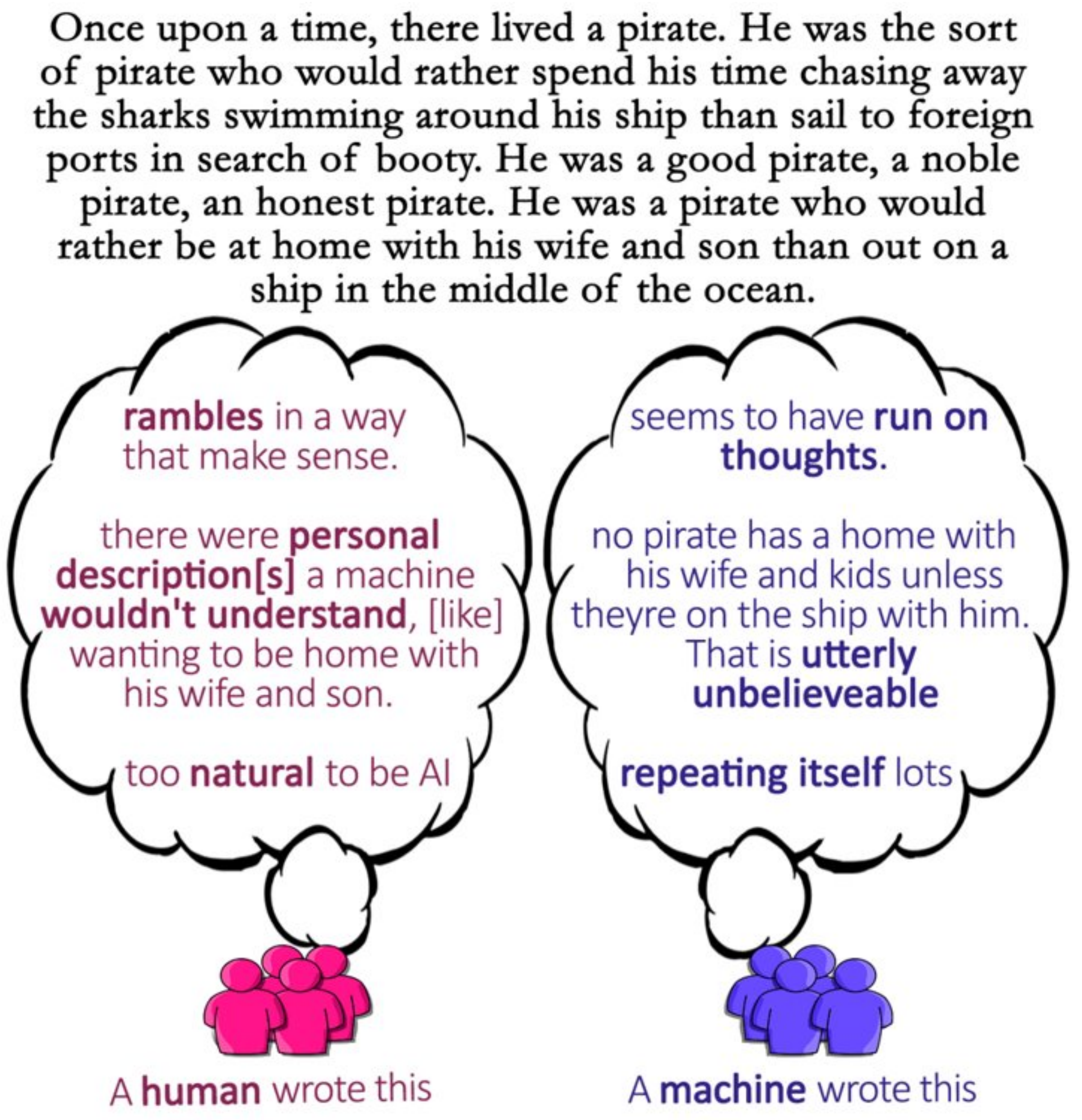
The distribution of resources in the world's languages [6]. The size of the gradient circle represents the number of languages in the class. The color spectrum represents the total speaker population size from low to high.

Low linguistic diversity in NLP

- Our contribution:
  - Russian SuperGLUE (Russian General Language Understanding Evaluation)
    A multi-task benchmark designed analogically to SuperGLUE for English
  - RuCoLA (Russian Corpus of Linguistic Acceptability)
    A single-task benchmark designed similarly to CoLA for English [11]
  - RuATD (Russian Artificial Text Detection)
    A two-task benchmark modelled after the Turing test [12]

Turing test in natural language generation

- Humans struggle to identify neural texts



Humans' explanations of why GPT3 texts are human-like (left) or model-like (right) [13]

Turing test in natural language generation

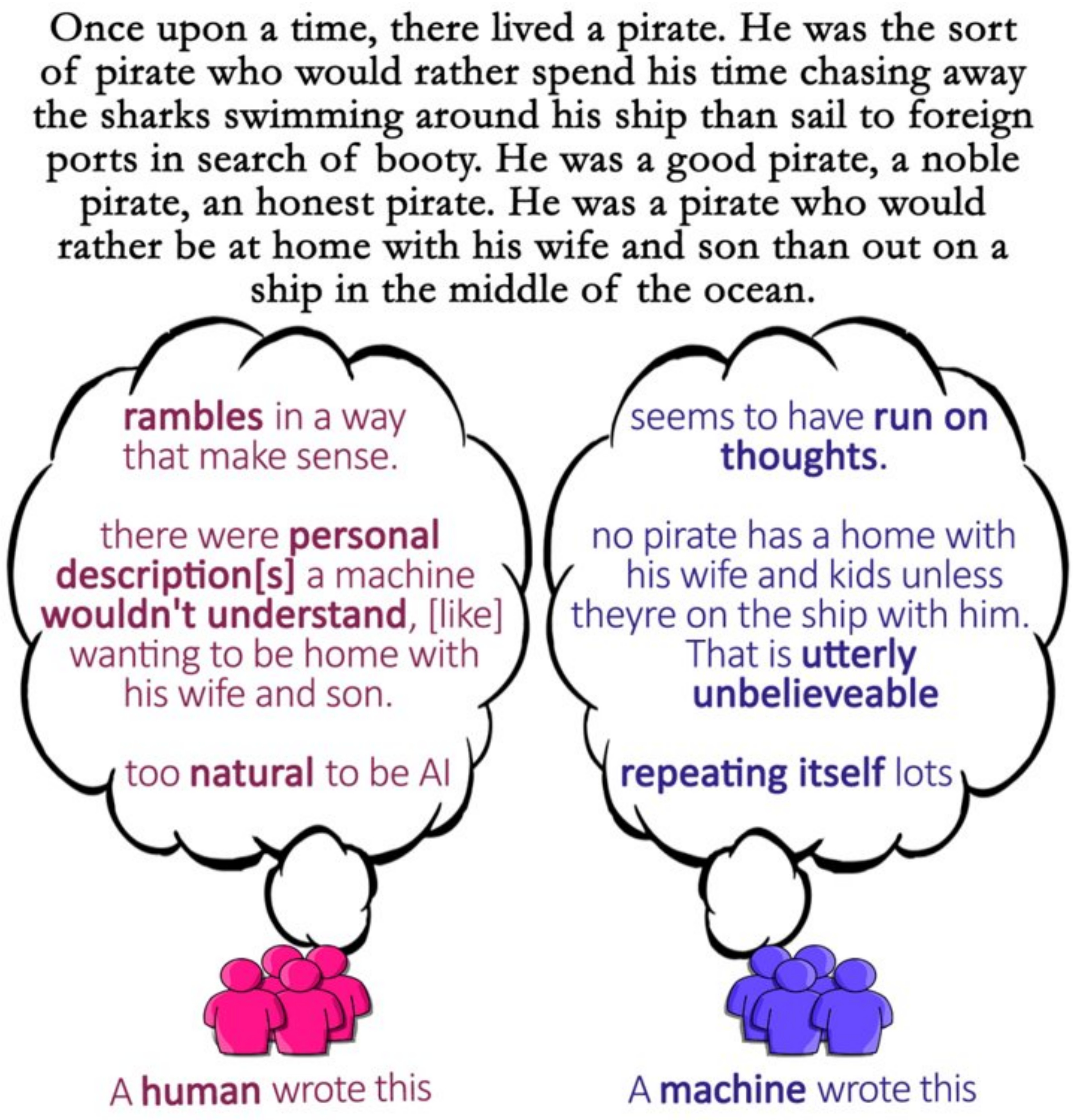
- Humans struggle to identify neural texts

🔍 Benchmarks

TURINGBENCH [2]
M4 [14]

🔍 Detectors

TF-IDF [15]
RoBERTa [16]



Humans' explanations of why GPT3 texts are human-like (left) or model-like (right) [13]

Turing test in natural language generation

- Humans struggle to identify neural texts

🔍 Benchmarks

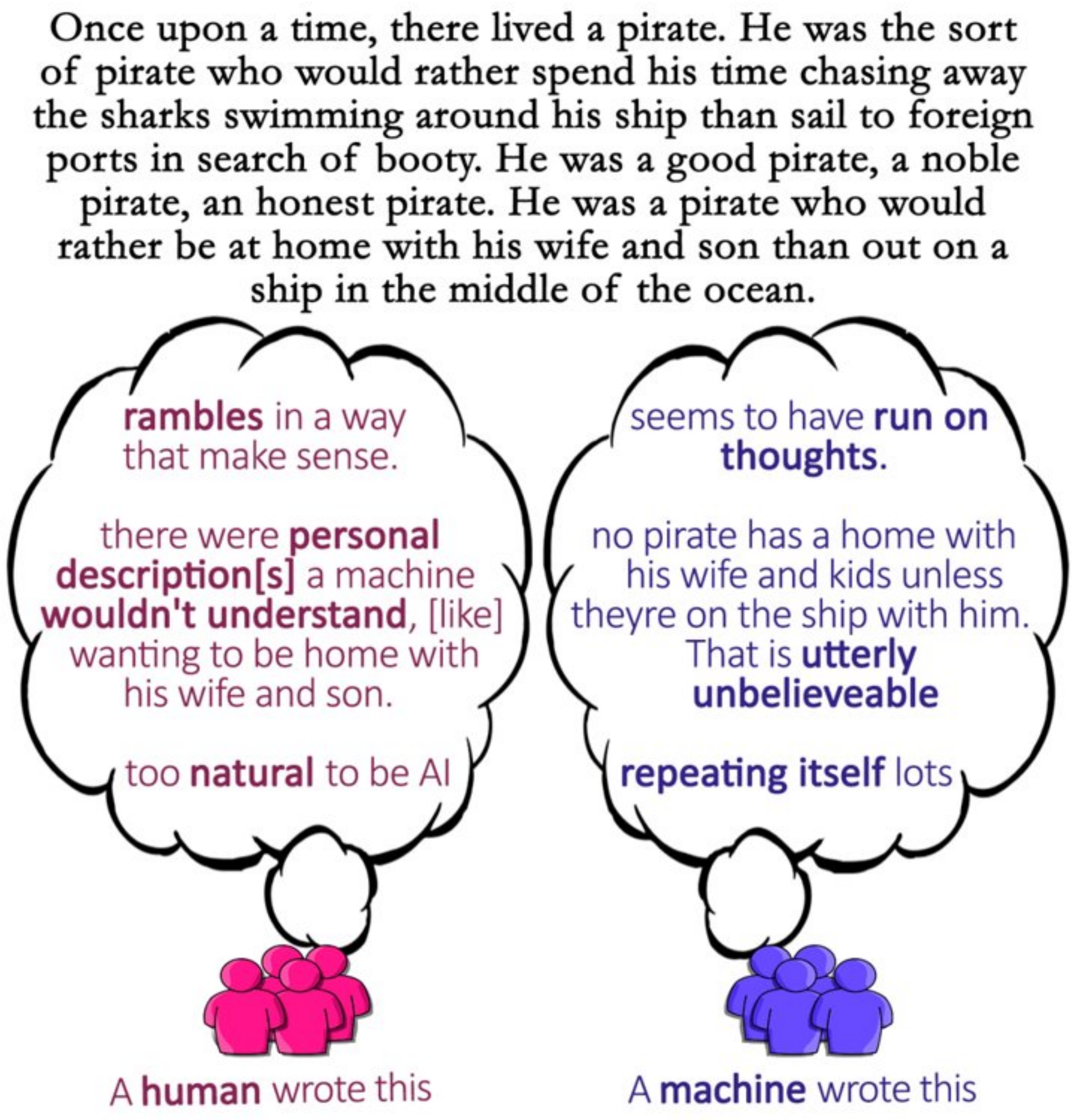
TURINGBENCH [2]
M4 [14]

🔍 Detectors

TF-IDF [15]
RoBERTa [16]

⁉️ Most existing detectors are not:

- interpretable
- robust to unseen generative LLMs



Humans' explanations of why GPT3 texts are human-like (left) or model-like (right) [13]

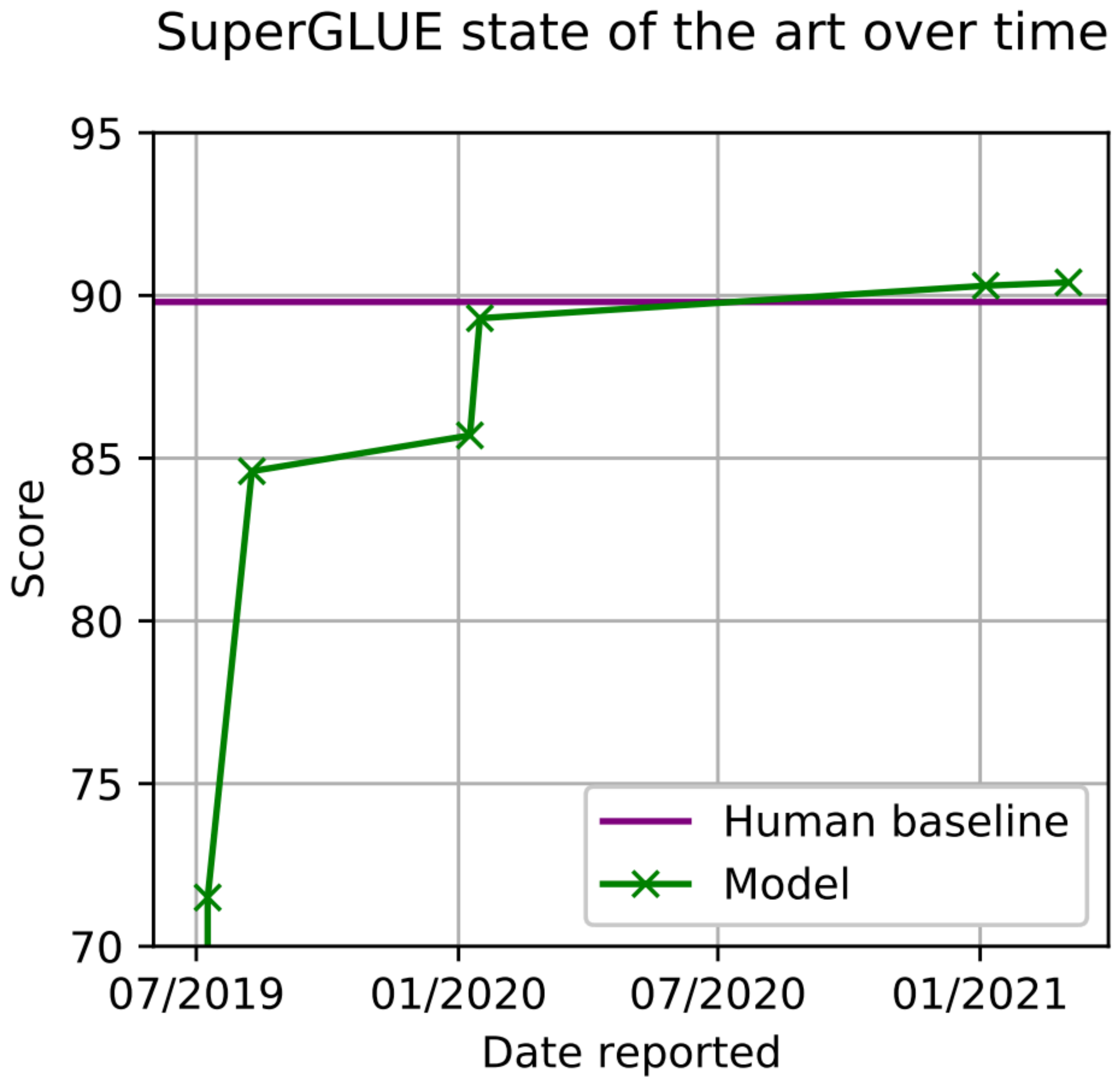
Turing test in natural language generation

- Our contribution:
  - A novel artificial text detector based on Topological Data Analysis (TDA)

    Outperforms/performs on par with existing detectors in 3 domains
    Interpretable and more robust to unseen GPT2 models

Aggregation procedures in NLP benchmarking

- The arithmetic mean is commonly used to rank LLMs on multi-task benchmarks, **but**:
  - Implies that all metrics are homogeneous
  - Declares the models best even if they outperform the others only on the outlier tasks



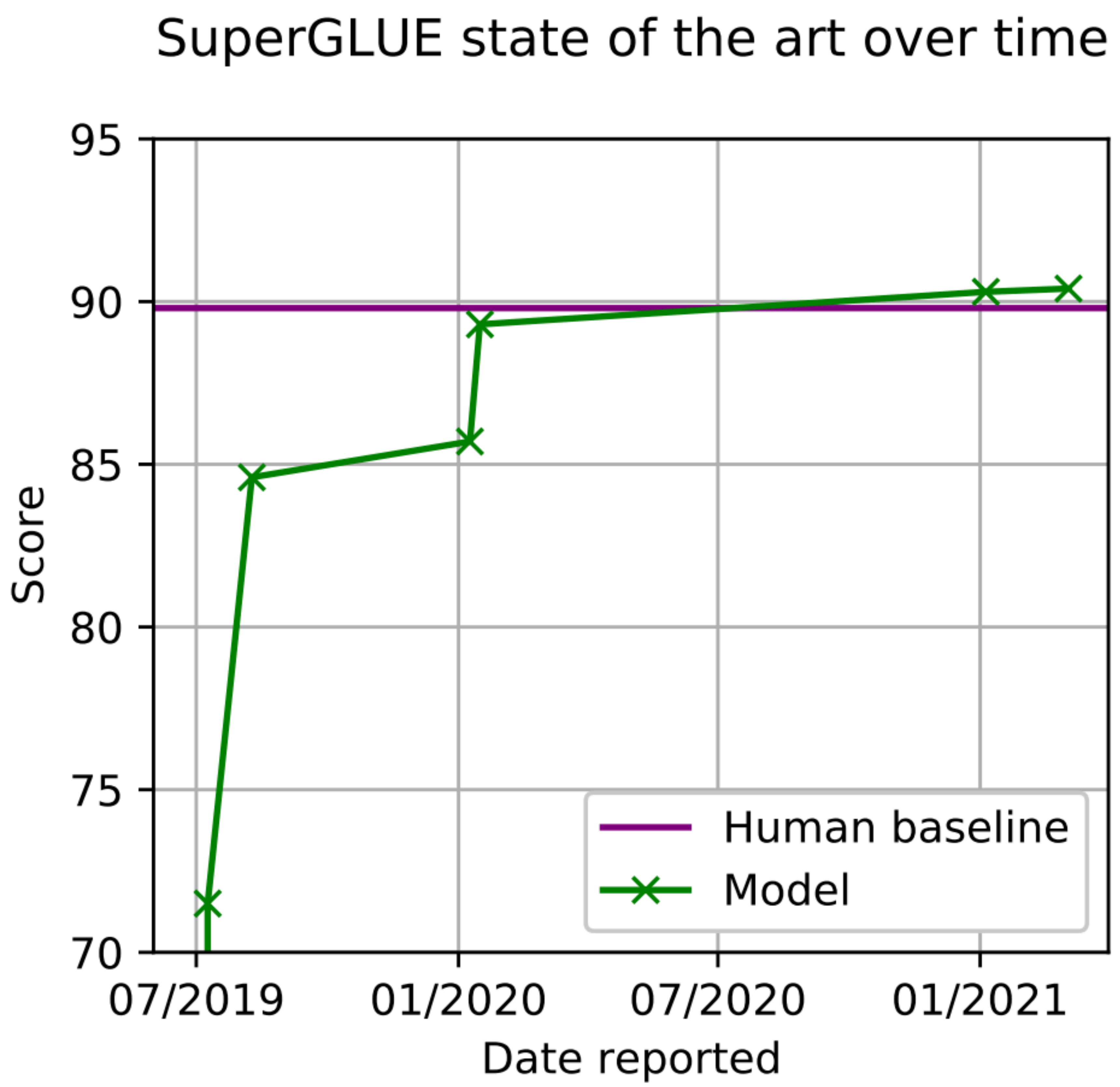
Saturation of the SuperGLUE benchmark over time based on the arithmetic mean aggregation [3]

Aggregation procedures in NLP benchmarking

- The arithmetic mean is commonly used to rank LLMs on multi-task benchmarks, **but**:
    - Implies that all metrics are homogeneous
    - Declares the models best even if they outperform the others only on the outlier tasks

New evaluation principles

Pareto efficiency [17]
DynaScore [18]



Saturation of the SuperGLUE benchmark over time based on the arithmetic mean aggregation [3]

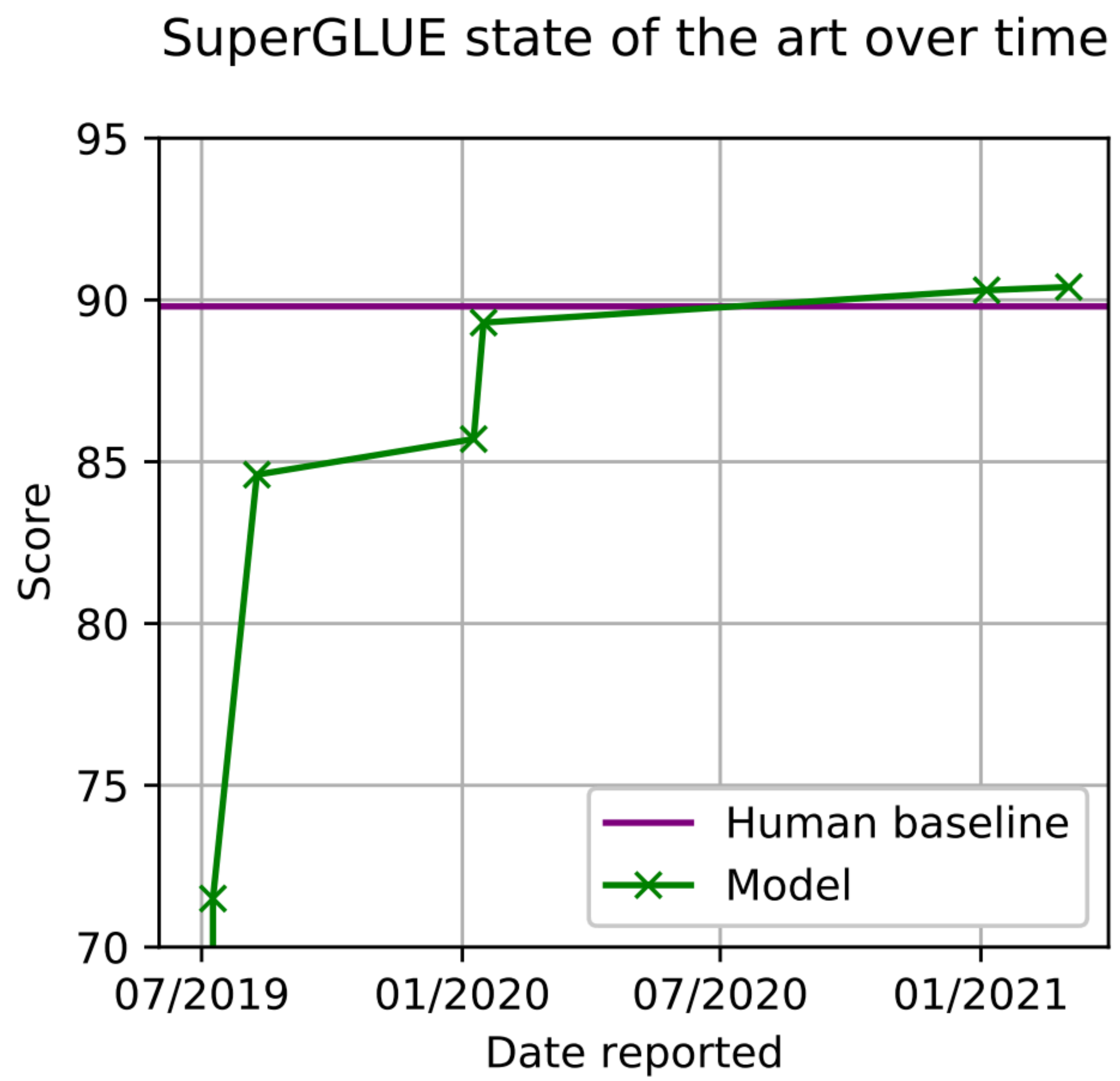
Aggregation procedures in NLP benchmarking

- The arithmetic mean is commonly used to rank LLMs on multi-task benchmarks, **but**:
  - Implies that all metrics are homogeneous
  - Declares the models best even if they outperform the others only on the outlier tasks

🔍 New evaluation principles

Pareto efficiency [17]
DynaScore [18]

⁉️ How to aggregate performances?



Saturation of the SuperGLUE benchmark over time based on the arithmetic mean aggregation [3]

Aggregation procedures in NLP benchmarking

- Our contribution:
  - Vote’n’Rank, a framework for ranking and selecting the best-performing LLMs

    8 novel performance aggregation procedures based on the social choice theory
  - Re-interpreting standard NLP and multi-modal benchmarks

    4 case studies conducted on the GLUE [19], SuperGLUE, and VALUE [20] benchmarks

Research goal

- Develop standardised evaluation resources and tools that:
  - provide an exhaustive comparison of existing and upcoming LLMs for Russian

Research goal

- Develop standardised evaluation resources and tools that:
    - provide an exhaustive comparison of existing and upcoming LLMs for Russian
    - enable the inclusion of Russian into the cross-lingual research directions

Research goal

- Develop standardised evaluation resources and tools that:
    - provide an exhaustive comparison of existing and upcoming LLMs for Russian
    - enable the inclusion of Russian into the cross-lingual research directions
    - address practical aspects of benchmarking and artificial text detection

Russian SuperGLUE: A Russian Language Understanding Benchmark

EMNLP 2020

Tasks

Dataset	|Train|	|Dev|	|Test|	Task	Metrics	Domain
DaNetQA	392	295	295	MRC	Acc.	Wikipedia
MuSeRC	500	100	322	MRC	$F1_a$/EM	news, fairy tales, academic texts, fiction, summaries of TV series and books
RuCoS	72k	4.3k	4.1k	MRC	F1/EM	news (Lenta, Deutsche Welle)
RUSSE	19.8k	8.5k	12.1k	WSD	Acc.	Wikipedia, RNC, dictionaries
PARus	400	100	500	NLI	Acc.	blogs, photography encyclopedia
RWSD	606	204	154	Coref.	Acc.	fiction
RCB	438	220	348	NLI	F1/Acc.	news, fiction
TERRa	2616	307	3198	NLI	Acc.	news, fiction
LiDiRus	✗	✗	1104	NLI	MCC	news, Wikipedia, Reddit, academic texts

MRC=machine reading comprehension. WSD=word sense disambiguation. NLI=natural language inference. Coref.=coreference resolution.

Empirical evaluation

- The BERT-based LLMs underperform humans on most NLU tasks
- The LLMs the exceed the human level on RUSSE (word sense disambiguation)

Model	Overall	LiDiRus MCC	RCB F1/Acc.	PARus Acc.	MuSeRC $F1_a$/EM	TERRa Acc.	RUSSE Acc.	RWSD Acc.	DaNetQA Acc.	RuCoS F1/EM
TF-IDF	43.4	5.9	30.1/44.1	48.6	58.7/24.2	47.1	66.0	66.2	62.1	25.6/25.1
ruBERT	54.6	18.6	43.2/46.8	61.0	65.6/25.6	63.9	**89.4**	67.5	74.9	25.5/25.1
mBERT	54.2	15.7	38.3/42.9	58.8	62.6/25.3	62.0	84.0	67.5	79.0	37.1/36.7
Human	**80.2**	**62.6**	**68.0/70.2**	**98.2**	**80.6/42.0**	**92.0**	74.7	**84.0**	**87.9**	**93.0/92.4**

Retrospective

Rank	Name	Team	Link	Score	LiDiRus	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQA	RuCoS
1	HUMAN BENCHMARK	AGI NLP	i	**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	Mistral 7B LoRA	Saiga team	i	**0.763**	0.46	0.529 / 0.573	0.824	0.927 / 0.787	0.888	0.758	0.786	0.919	0.83 / 0.816
3	FRED-T5 1.7B finetune	SberDevices	i	**0.762**	0.497	0.497 / 0.541	0.842	0.916 / 0.773	0.871	0.823	0.669	0.889	0.9 / 0.902
4	Golden Transformer v2.0	Avengers Ensemble	i	**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
5	LLaMA-2 13B LoRA	Saiga team	i	**0.718**	0.398	0.489 / 0.543	0.784	0.919 / 0.761	0.793	0.74	0.714	0.907	0.78 / 0.76
6	Saiga 13B LoRA	Saiga team	i	**0.712**	0.436	0.439 / 0.5	0.694	0.898 / 0.704	0.865	0.728	0.714	0.862	0.85 / 0.83
7	YaLM p-tune (3.3B frozen + 40k trainable params)	Yandex	i	**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
8	FRED-T5 large finetune	SberDevices	i	**0.706**	0.389	0.456 / 0.546	0.776	0.887 / 0.678	0.801	0.775	0.669	0.799	0.87 / 0.863
9	RuLeanALBERT	Yandex Research	i	**0.698**	0.403	0.361 / 0.413	0.796	0.874 / 0.654	0.812	0.789	0.669	0.76	0.9 / 0.902
10	FRED-T5 1.7B (only encoder 760M) finetune	SberDevices	i	**0.694**	0.421	0.311 / 0.441	0.806	0.882 / 0.666	0.831	0.723	0.669	0.735	0.91 / 0.911
11	ruT5-large finetune	SberDevices	i	**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
12	ruRoberta-large finetune	SberDevices	i	**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
13	gpt-3.5-turbo zero-shot	Saiga team	i	**0.682**	0.422	0.484 / 0.505	0.888	0.817 / 0.532	0.795	0.596	0.714	0.878	0.68 / 0.667

The performance gap between humans and LLMs: 25.8 —> 4.9 🦾

More than 2,000 private submissions

Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian

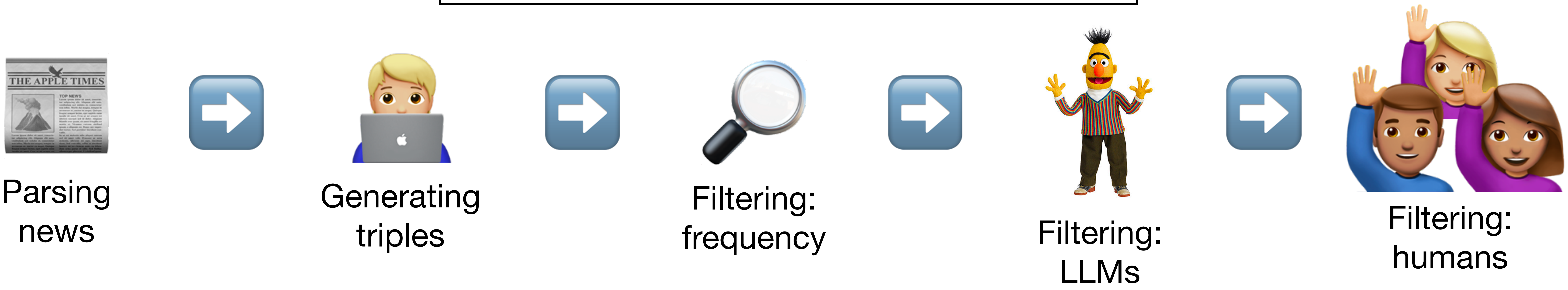
COLING 2020

Dataset creation

Passage: The mother of two boys who were abandoned by their father at *Moscow's Sheremetyevo airport* has taken them. This was reported to *TASS* by the press service of the *Ministry of education and science of the Khabarovsk territory*. Now the younger child attends kindergarten, and the older one goes to school. In educational institutions, full-time psychologists work with them as necessary. Also, the *Ministry of social protection of the population* is considering the issue of free health improvement for children in the summer. A few days after *Viktor Gavrilov* abandoned his children at the airport, he turned himself in to investigators in the city of *Bataysk*, *Rostov region*.

Query: On January 26, <placeholder> abandoned his sons, aged five and seven, in Sheremetyevo.

Correct Entities: Viktor Gavrilov



Empirical evaluation

- ruBERT-base performs the best among the baselines
- Performance gap between humans and LLMs:
  - MuSeRC: 9 (Macro-average F1) & 8.4 (exact match)
  - RuCoS: 58.6 (F1-score) & 58.5 (exact match)

Model	**MuSeRC** $\mathbf{F1_a/EM}$	**RuCoS** **F1/EM**
TF-IDF	58.9/24.4	25.6/25.1
mBERT	66.8/33.6	30.6/29.6
ruBERT-conv	71.7/32.9	26.4/25.9
ruBERT-base	71.7/33.6	34.4/33.9
Human	**80.6/42.0**	**93.0/92.4**

F1=F1-score; EM=exact match.

Retrospective

- Nowadays, the LLMs match or outperform humans on these tasks 🦾
- The best-performing LLMs:
  - FRED-T5 (SaluteDevices)
  - RuLeanALBERT (Yandex Research)
  - YaLM (Yandex)
- Russian takes the third place regarding the number of machine reading comprehension resources [21]

RuCoLA: Russian Corpus of Linguistic Acceptability

EMNLP 2022

Task

- Formulation: binary classification
- Metrics: Matthews Correlation Coefficient (MCC) & accuracy (Acc.)
- Categories: morphology, syntax, semantics, and hallucinations

Label	Set	Category	Sentence
✓	**In-domain**	✗	*Ya obnaruzhil ego lezhaschego odnogo na krovati.* I found him lying in the bed alone.
*	**In-domain**	SYNTAX	*Ivan prileg, chtoby on otdokhnul.* Ivan laid down in order that he has a rest.
✓	**Out-of-domain**	✗	*Ja ne chital ni odnogo iz ego romanov.* I have not read any of his novels.
*	**Out-of-domain**	HALLUCINATION	*Ljuk ostanavlivaet udachu ot etogo.* Luke stops luck from doing this.

Corpus creation

In-domain set

Out-of-domain set

Generating sentences

Annotating acceptabiliy

Annotating violation categories

Empirical evaluation

- ruRoBERTa achieves the best results among the LLMs
- The LLMs generalise well to the out-of-domain set
- The human performance is higher on the out-of-domain set, which can be attributed to the "unnaturalness" of the machine-specific features
- LLMs are least sensitive to morphological and semantic violations

Baseline	**Overall**		**In-domain**		**Out-of-domain**	
	Acc.	**MCC**	**Acc.**	**MCC**	**Acc.**	**MCC**
Non-neural models						
Majority	68.05 $\pm$ 0.0	0.0 $\pm$ 0.0	74.42 $\pm$ 0.0	0.0 $\pm$ 0.0	64.58 $\pm$ 0.0	0.0 $\pm$ 0.0
Linear	67.34 $\pm$ 0.0	0.04 $\pm$ 0.0	75.53 $\pm$ 0.0	0.17 $\pm$ 0.0	62.86 $\pm$ 0.0	-0.02 $\pm$ 0.0
Acceptability measures from LMs						
ruGPT-3	55.79 $\pm$ 0.0	0.27 $\pm$ 0.0	59.39 $\pm$ 0.0	0.19 $\pm$ 0.0	53.82 $\pm$ 0.0	0.30 $\pm$ 0.0
Russian language models						
ruBERT	75.9 $\pm$ 0.42	0.42 $\pm$ 0.01	78.82 $\pm$ 0.57	0.4 $\pm$ 0.01	74.3 $\pm$ 0.71	0.42 $\pm$ 0.01
ruRoBERTa	<u>80.8</u> $\pm$ 0.47	<u>0.54</u> $\pm$ 0.01	<u>83.48</u> $\pm$ 0.45	<u>0.53</u> $\pm$ 0.01	<u>79.34</u> $\pm$ 0.57	<u>0.53</u> $\pm$ 0.01
ruT5	71.26 $\pm$ 1.31	0.27 $\pm$ 0.03	76.49 $\pm$ 1.54	0.33 $\pm$ 0.03	68.41 $\pm$ 1.55	0.25 $\pm$ 0.04
Cross-lingual models						
XLM-R	65.73 $\pm$ 2.33	0.17 $\pm$ 0.04	74.17 $\pm$ 1.75	0.22 $\pm$ 0.03	61.13 $\pm$ 2.9	0.13 $\pm$ 0.05
RemBERT	76.21 $\pm$ 0.33	0.44 $\pm$ 0.01	78.32 $\pm$ 0.75	0.4 $\pm$ 0.02	75.06 $\pm$ 0.55	0.44 $\pm$ 0.01
Human	**84.08**	**0.63**	**83.55**	**0.57**	**84.59**	**0.67**

Retrospective

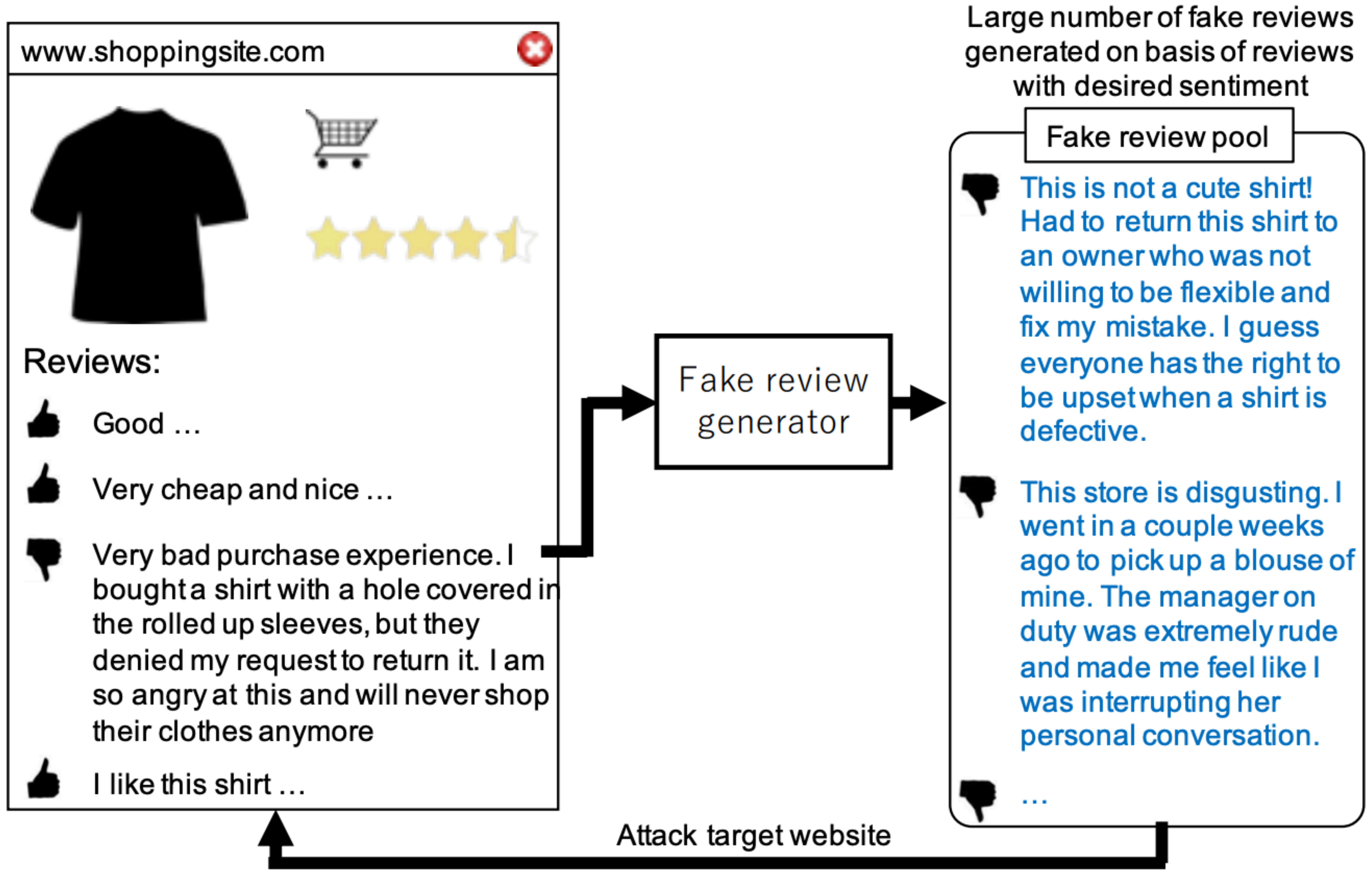
rucola | dataset & baseline ↗ | paper ↗ | leaderboard | FAQ | eng ▾ | sign in

overall | by source

Rank	Team	Model	Date	Acc	MCC
1	RuCoLA Team	Human Benchmark	3/11/2022	0.84	0.63
2	smth_binomial	ruRoBERTa-v2	23/3/2023	0.81	0.57
3	RuCoLA Team	ruRoBERTa-large	30/8/2022	0.82	0.56
4	TopaCoLA	TDA-RoBERTa	26/2/2023	0.81	0.56
5	yp	BertGram	3/7/2022	0.81	0.56
6	Random Submit	pred_1	11/6/2022	0.81	0.55
7	Mindful Squirrel	RuRobertaLargeAug_v2	30/5/2022	0.81	0.55

Still challenging for LLMs and humans

RuCoLA-based models were used for filtering the Russian DALL-E's pretraining corpus



Example of generating fake product reviews [22]

Findings of the RuATD Shared Task 2022 on Artificial Text Detection

Dialogue 2022

Method

Task	Model	Size	N	%	Domain	Task	Model	Size	N	%	Domain
Back-translation	Human M-BART50 M2M-100 OPUS-MT	35,588	12.9	88.0	RNC, Wikipedia, news, diaries, WikiMatrix, Tatoeba, SD	**Machine translation**	Human M-BART50 M2M-100 OPUS-MT	35,860	11.5	89.0	WikiMatrix, Tatoeba
Open-ended generation	Human ruGPT3-small ruGPT3-medium ruGPT3-large	37,499	141.5	85.0	RNC, Wikipedia, news, diaries, SD, social media	**Text summarisation**	Human M-BART M-BART50 ruT5-base	17,164	33.5	86.0	RNC, Wikipedia, news, diaries, SD
Paraphrase generation	Human mT5-small mT5-large ruGPT2-large ruGPT3-large ruT5-base	44,298	13.0	85.0	RNC, SD, social media, Wikipedia, news, diaries	**Text simplification**	Human mT5-large ruGPT3-small ruGPT3-medium ruGPT3-large ruT5-large	44,700	18.3	86.0	RNC, SD, social media, Wikipedia, news, diaries

N=average number of tokens; %=percentage of high-frequency tokens; SD=strategic documents; RNC=Russian National Corpus.

Empirical evaluation

- In total, 38 submissions:
    - 30 submissions (the first task)
    - 8 submissions (the second task)
- The performance depends on the length (the higher the length, the better)
- Authorship attribution for language generation is not trivial
- Humans achieve only 0.66 accuracy

Rank	Detection of neural texts		Authorship attribution	
	Team	**Acc.**	**Team**	**Acc.**
1	MSU	0.829	Posokhov Pavel	0.650
2	Igor	0.827	Yixuan Weng	0.647
3	Orzhan	0.826	Orzhan	0.646
4	mariananieva	0.824	MSU	0.628
5	Ivan Zakharov	0.822	ruBERT baseline	0.598
6	Yixuan Weng	0.818	Nikita Selin	0.590
7	ilya koziev	0.817	Victor Krasilnikov	0.550
8	miso soup	0.811	Petr Grigoriev	0.458
9	Eduard Belov	0.810	TF-IDF baseline	0.443

Retrospective

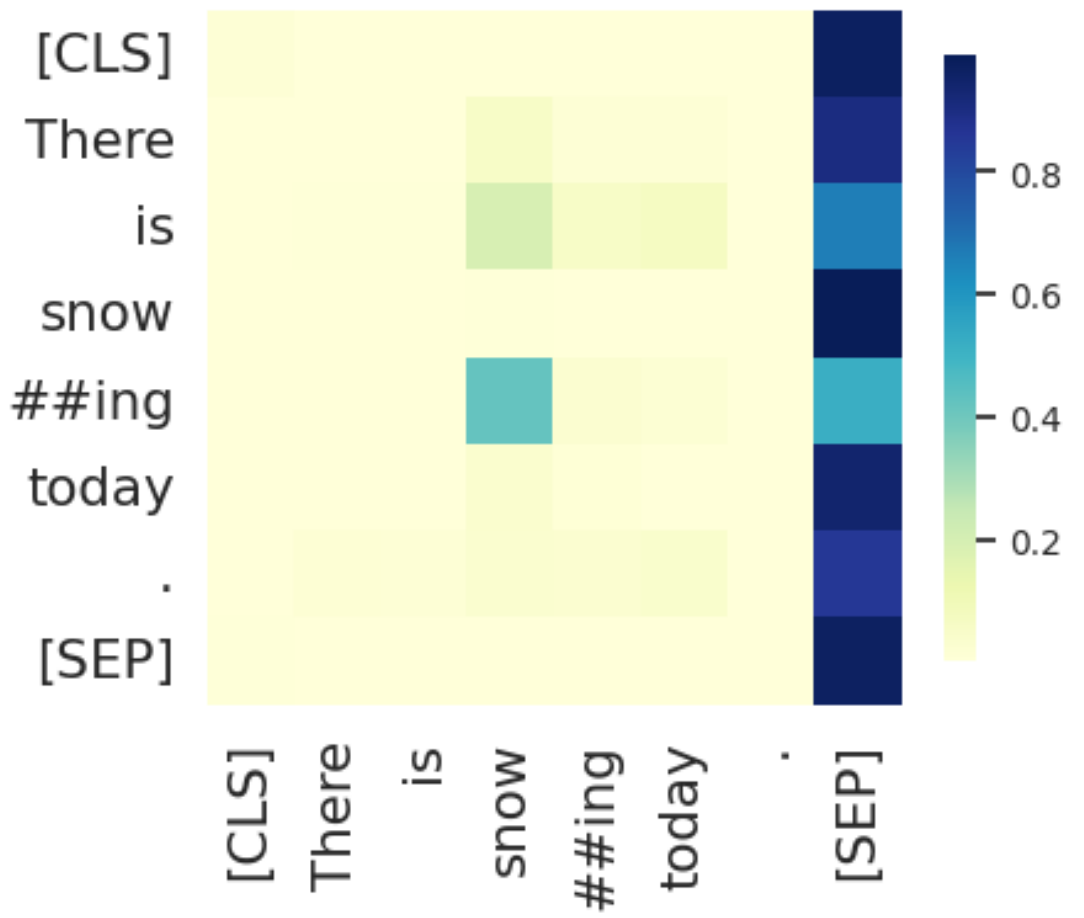
- Transformer-based detectors outperform humans by up to 16.3% of the accuracy score
- The RuATD benchmark has been included in the SemEval-2024 task on multi-generator, multi-domain, and cross-lingual artificial text detection [14]

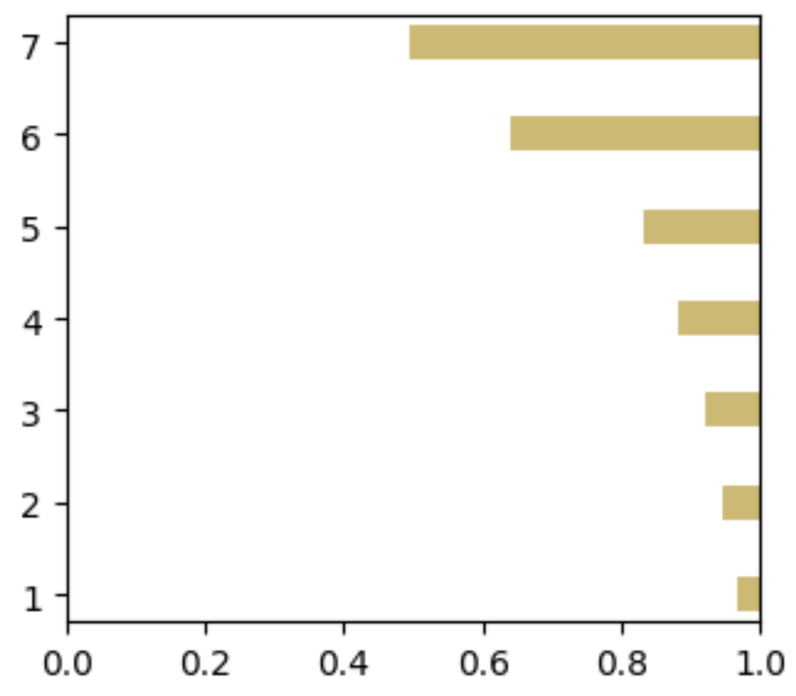
ChatGPT: "The LLMs have become a powerful tool for generating text that closely resembles human language, but their misuse can have serious consequences. Misuse can lead to the amplification of biases present in the training data, the generation of misinformation, and privacy violations. Therefore, it is important to use these models responsibly, with careful consideration of the potential risks involved."

Artificial Text Detection via Examining the Topology of Attention Maps

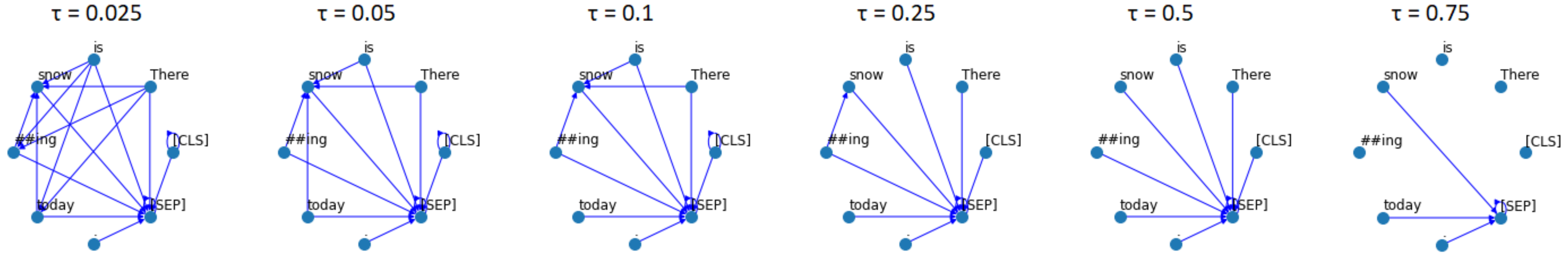
EMNLP 2021

Method



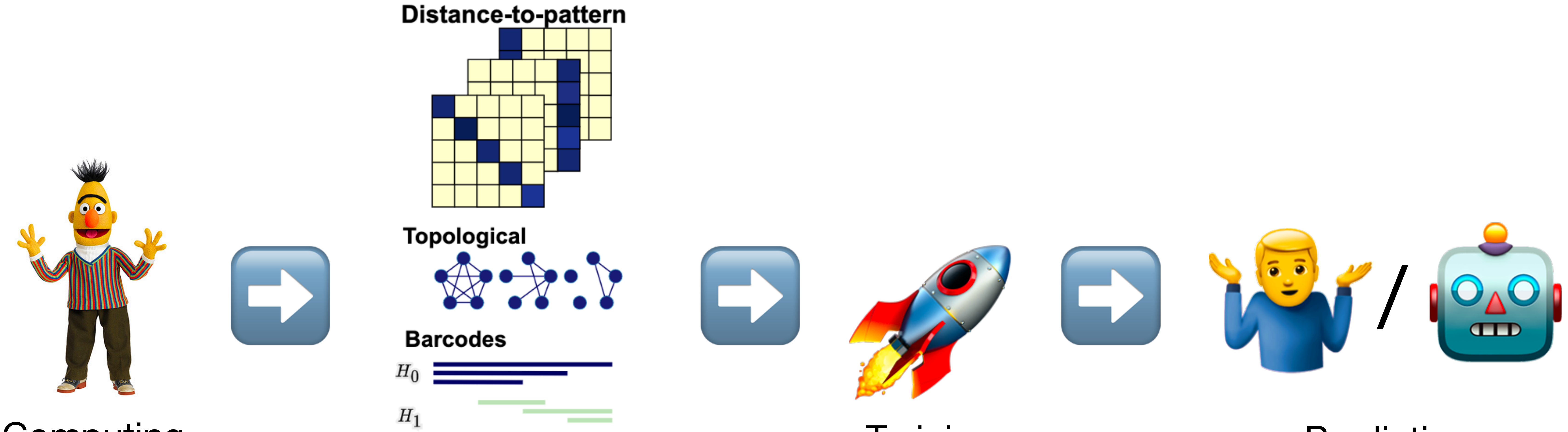


Example of the attention map (left) and barcodes (right)



Example of filtration

Method



Computing Attentions

Computing features

Training detector

Predicting: human or LLM?

Empirical evaluation

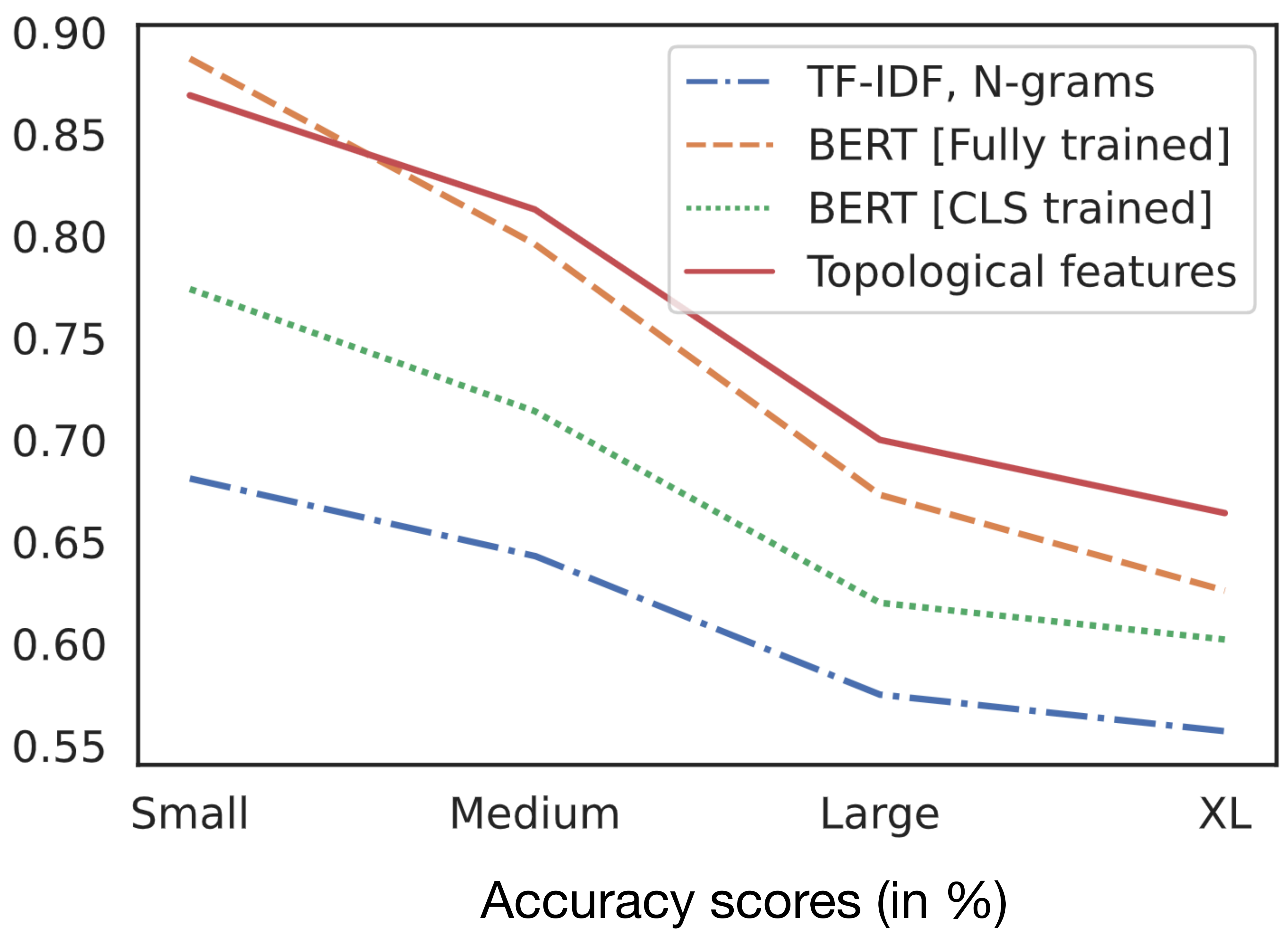
- TDA-based detectors:
    - outperform the count-based and neural baselines
    - perform on par with the finetuned BERT

Model	Reddit & GPT-2 Small	Amazon Reviews & GPT-2 XL	RealNews & GROVER
TF-IDF, N-grams	68.1	54.2	56.9
BERT [CLS trained]	77.4	54.4	53.8
BERT [Fully finetuned]	**88.7**	**60.1**	**62.9**
BERT [SLOR]	78.8	59.3	53.0
Topological features	86.9	59.6	63.0
Features derived from barcodes	84.2	60.3	61.5
Features based on the distance to patterns	85.4	61.0	62.3
All features	**87.7**	**61.1**	**63.6**

Accuracy scores (in %). SLOR is an acceptability measure that accounts for length and unigram probability [26]

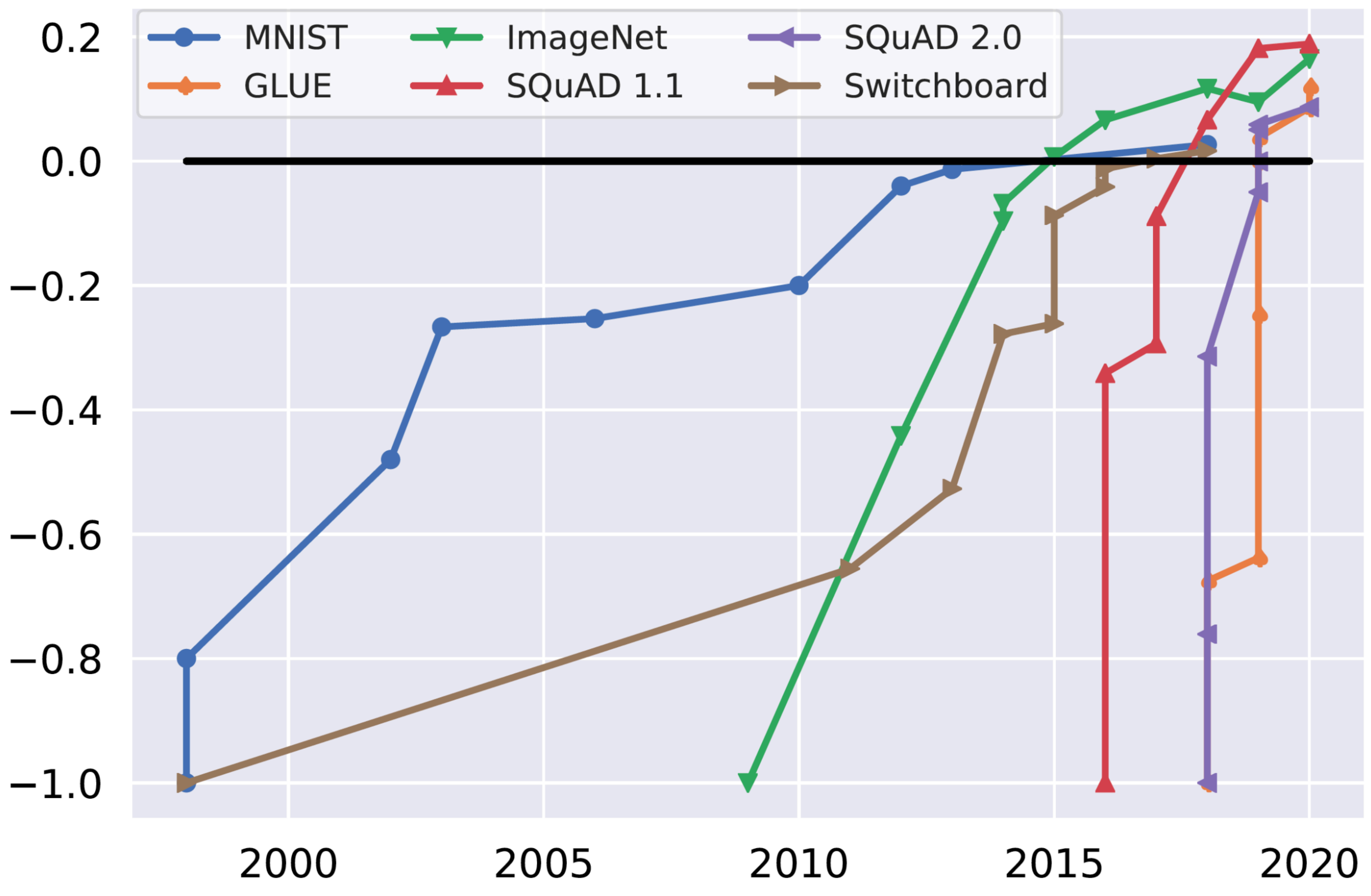
Empirical evaluation

- TDA-based detectors:
  - outperform the count-based and neural baselines
  - perform on par with the finetuned BERT
  - more generalisable to unseen GPT-2 LLMs, but perform slightly worse on the GPT-2-small test set



Retrospective

- TDA is becoming more popular in NLP
- ATD is becoming more and more difficult
- Our methodology has been adapted to:
  - promote new state-of-the-art results in speech processing tasks [23]
  - reach the human-level performance in acceptability judgments tasks [24]



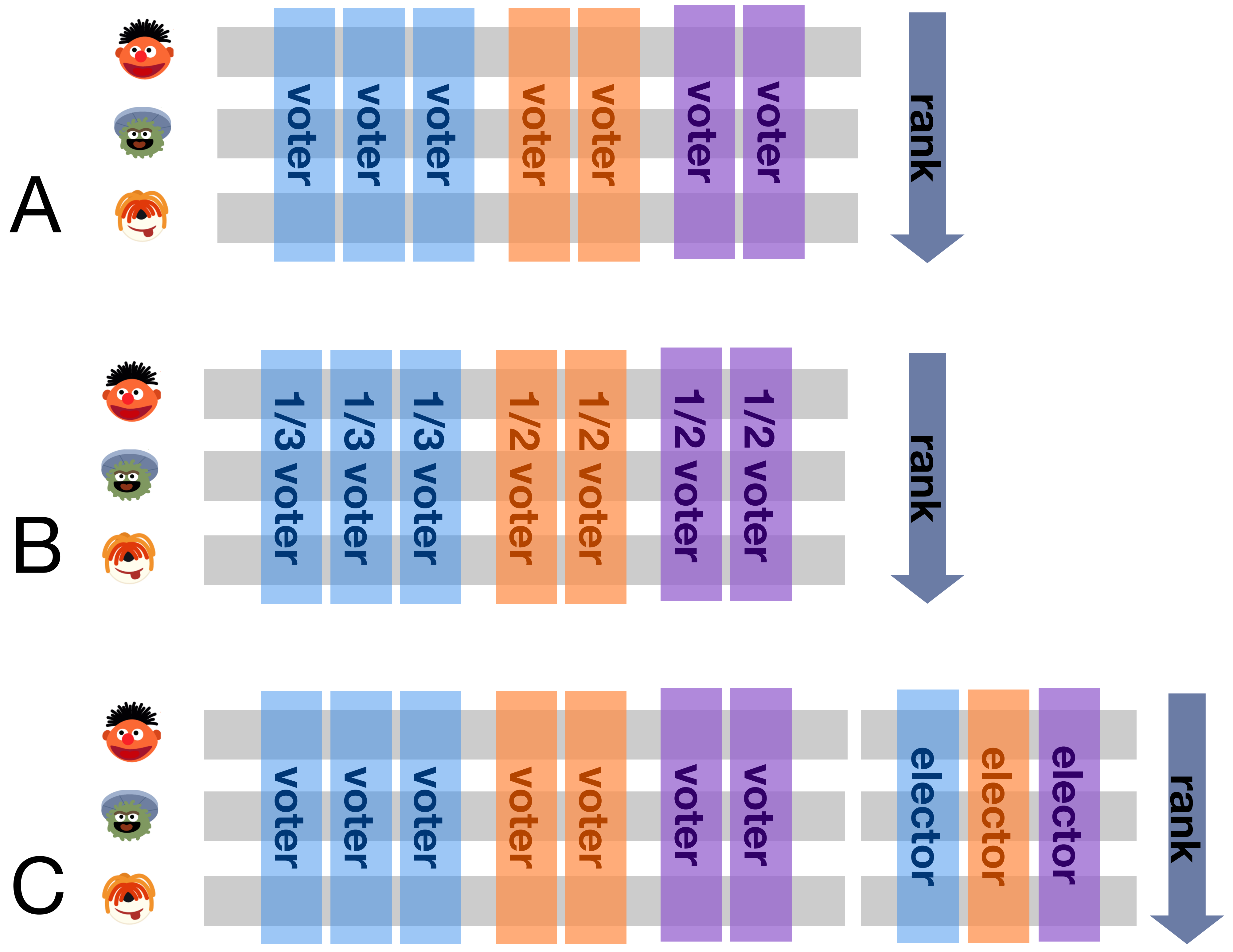
Example of saturated benchmarks [25]

Vote’n’Rank: Revision of Benchmarking with Social Choice Theory

EACL 2023

Method

- Aggregation procedures:
  - Scoring rules
  - Iterative scoring rules
  - Majority-relation based rules
- Scenarios:
  A. Basic aggregation
  B. Weighted aggregation
  C. Two-step aggregation



Voter=task; elector=interim ranking. 1/N=task group weight.

Method

- **Scoring rules:** the total score for the system is the sum of scores in each task based on the scoring vector *c.*

- System scores, where |M| is the number of systems:

  - Plurality rule: c = (1, 0, …, 0)

    A = 2; B = C = D = 1

  - Borda rule: c = (|M| - 1, |M| - 2, …, 1, 0)

    B = 9, C = 8, D = 7, A = 6

  - Dowdall rule: c = (1, 1/2, …, 1/|M|)

    A = B = 2.75, C = 2.5, D = 2.41

Rank	Task 1	Task 2	Task 3	Task 4	Task 5
1	m_A	m_A	m_B	m_C	m_D
2	m_B	m_C	m_D	m_B	m_B
3	m_C	m_D	m_C	m_D	m_C
4	m_D	m_B	m_A	m_A	m_A

Method

- **Iterative scoring rules**: having *c,* let us iteratively calculate the total score of the system. Stop the procedure when it is impossible to break ties or there is only one alternative left.
  - Threshold rule: c = (1, 1, …, 1,1, 0)

    C = 5, B = D = 4, A = 2

    The worst ranking matters the most

Rank	Task 1	Task 2	Task 3	Task 4	Task 5
1	m_A	m_A	m_B	m_C	m_D
2	m_B	m_C	m_D	m_B	m_B
3	m_C	m_D	m_C	m_D	m_C
4	m_D	m_B	m_A	m_A	m_A

Method

- **Iterative scoring rules**: having *c,* let us iteratively calculate the total score of the system. Stop the procedure when it is impossible to break ties or there is only one alternative left.
  - Baldwin rule: c = (|M| - 1, |M| - 2, …, 1, 0), (|M| - 2, |M| - 3, …, 1, 0, 0), …, (1, 0, …, 0) and discards systems with the minimum sum of scores at each iteration.

Relies on Borda: B = 9, C = 8, D = 7, A = 6

Eliminates A and uses c = (2, 1, 0): B = 6, C = 5, D = 4

Eliminates D and uses c = (1, 0): B = 3, C = 2

Rank	Task 1	Task 2	Task 3	Task 4	Task 5
1	m_A	m_A	m_B	m_C	m_D
2	m_B	m_C	m_D	m_B	m_B
3	m_C	m_D	m_C	m_D	m_C
4	m_D	m_B	m_A	m_A	m_A

Rank	Task 1	Task 2	Task 3	Task 4	Task 5
1	m_B	m_C	m_B	m_C	m_D
2	m_C	m_D	m_D	m_B	m_B
3	m_D	m_B	m_C	m_D	m_C

Example for the Baldwin rule.

Method

- **Majority-relation based rules**

Let us define a majority relation μ over the set of alternatives as the following binary relation: $m_A \mu m_B$ iff m_A is ranked higher than m_B by more criteria.

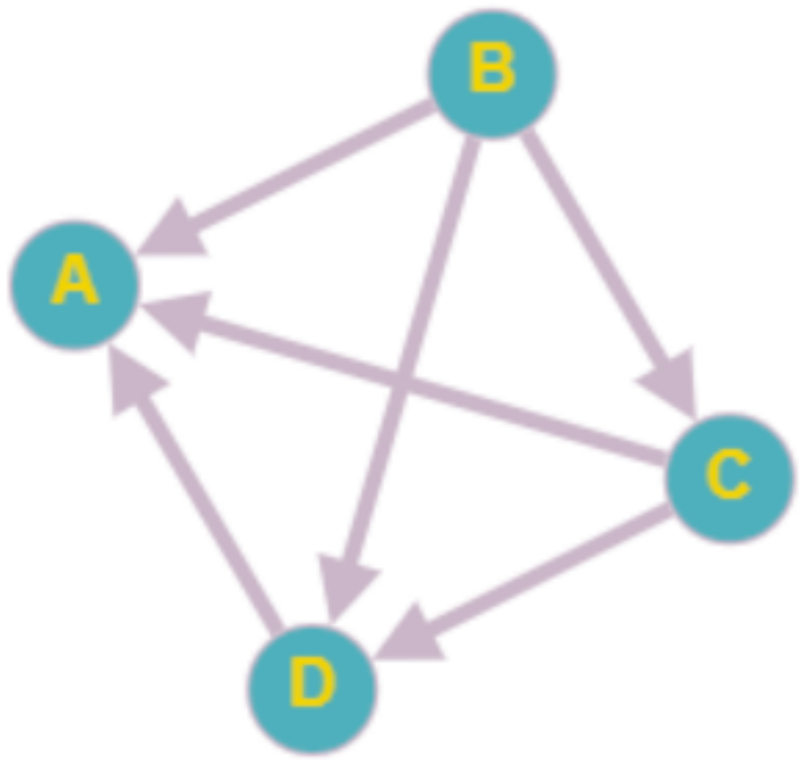
- *Condorcet rule*. m_C is the Condorcet winner (CW) iff $m_C \mu m$ for any $m \in M$.

B is the CW

Selects the system that dominates all systems

in pair-wise comparison

Rank	Task 1	Task 2	Task 3	Task 4	Task 5
1	m_A	m_A	m_B	m_C	m_D
2	m_B	m_C	m_D	m_B	m_B
3	m_C	m_D	m_C	m_D	m_C
4	m_D	m_B	m_A	m_A	m_A



Method

- **Majority-relation based rules**

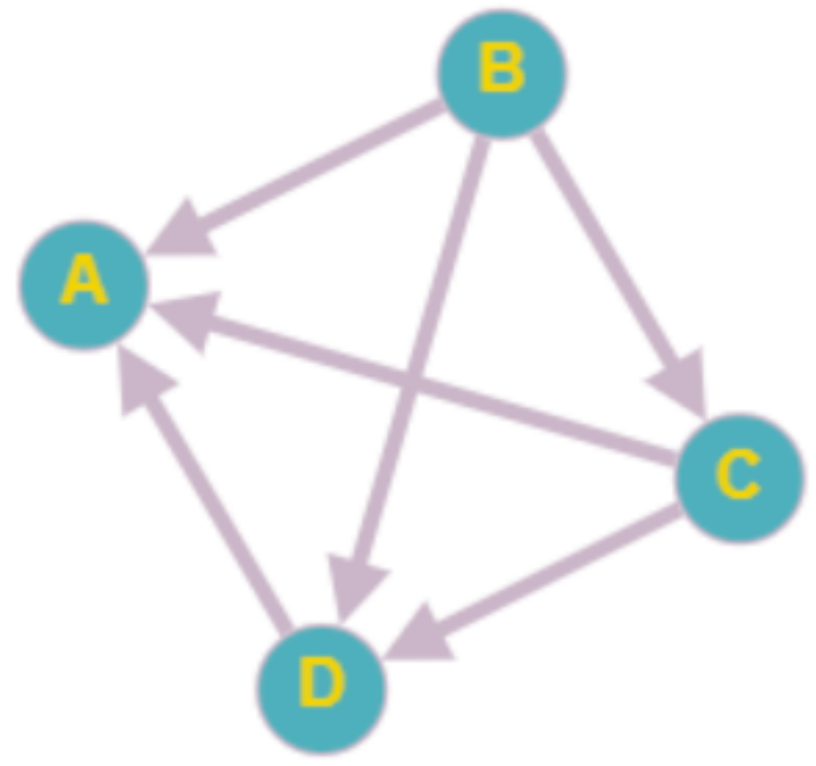
Let us define a majority relation μ over the set of alternatives as the following binary relation: $m_A \mu m_B$ iff m_A is ranked higher than m_B by more criteria.

Copeland rule. Define the lower counter set of systems m_A as a set of systems dominated by m_A via μ: $L(m_A) = \{m \in M, m_A \mu m\}$. In a similar way, define the upper counter set of systems m_A as a set of systems that dominate m_A via μ: $U(m_A) = \{m \in M, m \mu m_A\}$. Define $u(m) = |L(m)| - |U(m)|$. The final decision is provided by the alternatives with the highest $u(m)$.

B = 3, C = 1, D = -1, A = -3

Selects a system that wins more than loses

Rank	Task 1	Task 2	Task 3	Task 4	Task 5
1	m_A	m_A	m_B	m_C	m_D
2	m_B	m_C	m_D	m_B	m_B
3	m_C	m_D	m_C	m_D	m_C
4	m_D	m_B	m_A	m_A	m_A



Method

- **Majority-relation based rules**

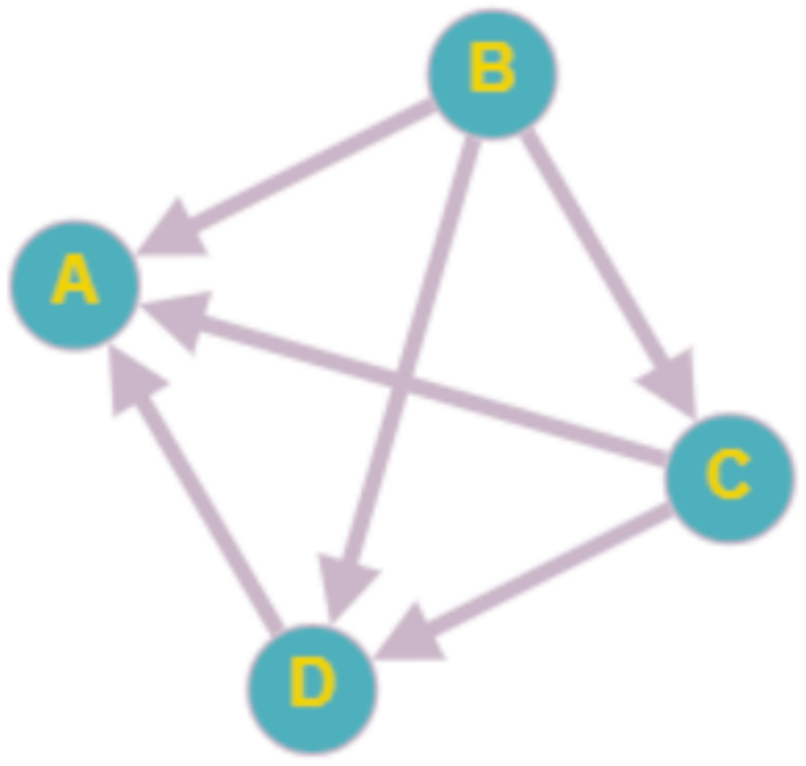
Let us define a majority relation μ over the set of alternatives as the following binary relation: $m_A \mu m_B$ iff m_A is ranked higher than m_B by more criteria.

Minimax rule. Let $s(m_A, m_B)$ be the number of criteria for which system m_A is ranked higher than system m_B if $m_A \mu m_B$ or $s(m_A, m_B) = 0$ otherwise. The systems are ranked according to the formula $\text{rank}(m_A) = -\max_B s(m_B, m_A)$.

B = 0, C = D = A = -3

Selects a system with a minimum number of defeats

Rank	Task 1	Task 2	Task 3	Task 4	Task 5
1	m_A	m_A	m_B	m_C	m_D
2	m_B	m_C	m_D	m_B	m_B
3	m_C	m_D	m_C	m_D	m_C
4	m_D	m_B	m_A	m_A	m_A



Empirical evaluation

- Baselines:
  - σ^{am} : the arithmetic mean aggregation
  - σ^{gm} : the geometric mean aggregation
  - σ^{og} : optimality gap [26], an aggregation metric that identifies the amount by which the system fails to get a minimum score of 0.95
- Case studies:
  1. Re-interpreting benchmarks: GLUE, SuperGLUE, VALUE
  2. Robustness to omitting scores
  3. Ranking based on user preferences

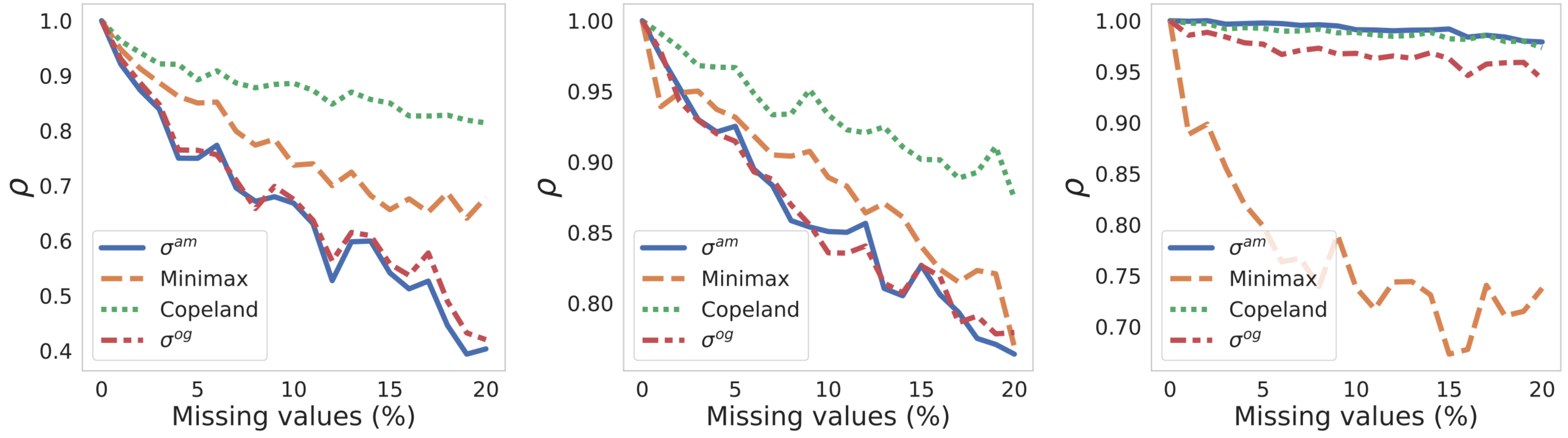
Re-interpreting benchmarks

Rank	σ^{am}	σ^{gm}	σ^{og}	**Copeland**	**Minimax**	**Plurality**	**Dowdall**	**Borda**
1	91.18	90.89 $\updownarrow 0$	0.074 $\updownarrow 0$	29.00 $\updownarrow 0$	0 $\updownarrow 0$	2.00 $\updownarrow 0$	4.95 $\updownarrow 0$	260.50 $\updownarrow 0$
2	91.07	90.78 $\updownarrow 0$	0.075 $\uparrow 4$	25.00 $\uparrow 1$	-5.50 $\uparrow 1$	2.00 $\uparrow 13$	4.08 $\uparrow 13$	256.00 $\updownarrow 0$
3	90.88	90.56 $\updownarrow 0$	0.076 $\downarrow 1$	24.00 $\downarrow 1$	-6.00 $\uparrow 1$	1.50 $\updownarrow 0$	3.82 $\updownarrow 0$	247.50 $\updownarrow 0$
4	90.86	90.48 $\updownarrow 0$	0.076 $\updownarrow 0$	22.00 $\uparrow 3$	-6.50 $\downarrow 2$	1.00 $\uparrow 1$	3.41 $\updownarrow 0$	241.50 $\updownarrow 0$
5	90.74	90.44 $\updownarrow 0$	0.077 $\updownarrow 0$	22.00 $\uparrow 10$	-7.00 $\uparrow 2$	1.00 $\downarrow 3$	3.27 $\downarrow 3$	233.50 $\uparrow 1$
6	90.66	90.34 $\updownarrow 0$	0.078 $\uparrow 1$	22.00 $\downarrow 2$	-7.00 $\uparrow 9$	0.50 $\updownarrow 0$	2.57 $\downarrow 1$	229.50 $\uparrow 1$
7	90.48	90.11 $\updownarrow 0$	0.082 $\uparrow 3$	16.00 $\downarrow 1$	-7.00 $\downarrow 1$	0.00 $\downarrow 3$	2.55 $\updownarrow 0$	220.50 $\downarrow 2$

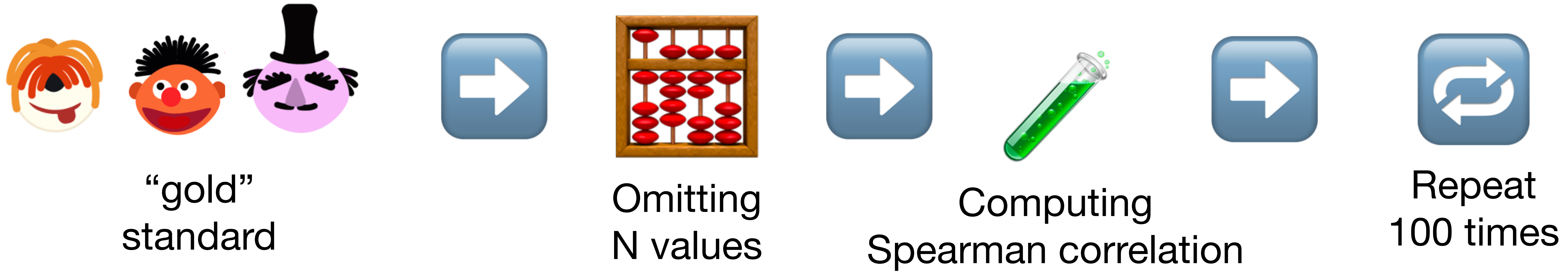
Results of re-ranking the GLUE benchmark. Changes in the system ranks are depicted with arrows, whilst the superscripts denote scores assigned by the aggregation procedure. Notations: =HUMAN; =ERNIE; =STRUCTBERT+CLEVER; =DEBERTA+CLEVER; =DEBERTA/TURINGNLRV4; =MACALBERT+DKM; =T5; =ALBERT+DAAF+NAS; =FUNNEL. The superscript values stand for the voting rules' scores, whilst the subscript values indicate changes in the ranking positions. $\uparrow x$ means up x positions, $\downarrow x$ means down x positions, $\updownarrow$ means no changes.

Humans still can take the leading positions!

Robustness to omitting scores



(a) GLUE (b) SGLUE (c) VALUE

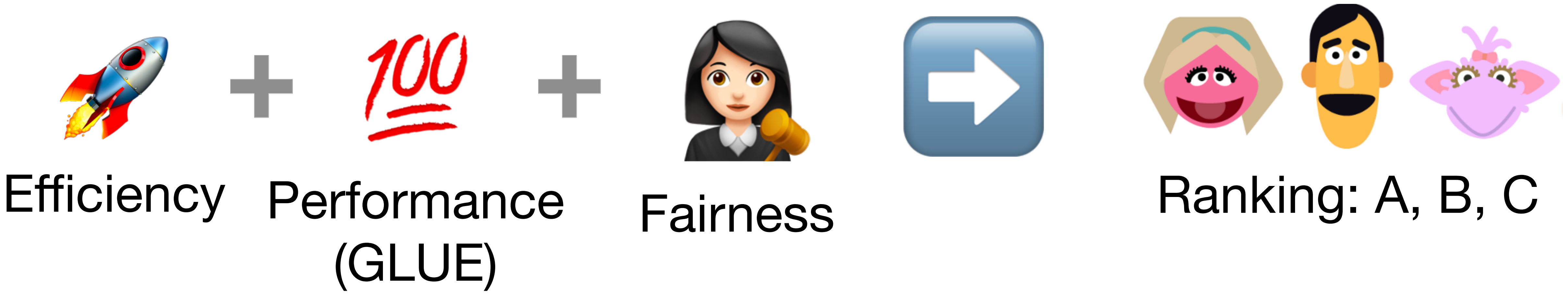


More robust, but Minimax is indecisive on VALUE

Ranking based on user preferences

Rank	$\sigma^{am}_{\text{Performance}}$	Borda	Weighted Borda	Weighted 2-step Borda	Borda Performance	Borda Efficiency	Borda Fairness
1	DeBERTa 82.73	DistilRoBERTa 267.0 ↑4	DistilBERT 10.75 ↑5	ALBERT 4.30 ↑2	DeBERTa 56.5 ↕0	DistilRoBERTa 223.0 ↑4	ALBERT 19.0 ↑2
2	RoBERTa 82.52	DistilBERT 245.0 ↑4	ALBERT 9.83 ↑1	BERT 3.60 ↑2	RoBERTa 49.0 ↕0	DistilBERT 216.0 ↑4	DistilBERT 18.0 ↑4
3	ALBERT 80.94	BERT 166.0 ↑1	BERT 8.96 ↑1	DistilBERT 3.40 ↑3	ALBERT 32.5 ↕0	BERT 120.0 ↑1	BERT 14.0 ↑1
4	BERT 79.20	ALBERT 154.0 ↓1	DistilRoBERTa 8.63 ↑1	DeBERTa 3.00 ↓3	BERT 32.0 ↕0	ALBERT 103.0 ↓1	DeBERTa 11.00 ↓3
5	DistilRoBERTa 78.56	RoBERTa 144.0 ↓3	DeBERTa 7.17 ↓4	DistilRoBERTa 2.90 ↕0	DistilRoBERTa 17.0 ↕0	RoBERTa 91.0 ↓3	GPT2 11.0 ↑1
6	DistilBERT 77.89	GPT2 10.0 ↑1	RoBERTa 7.04 ↓4	RoBERTa 2.60 ↓3	DistilBERT 11.0 ↕0	GPT2 84.0 ↑1	DistilRoBERTa 7.00 ↑1
7	GPT2 75.95	DeBERTa 70.50 ↓6	GPT2 5.47 ↕0	GPT2 0.90 ↕0	GPT2 8.0 ↕0	DeBERTa 3.00 ↓6	RoBERTa 4.00 ↓5

Results of re-ranking the GLUE benchmark using the *Borda* rule in the simulated user-oriented scenario.
Notations: = ALBERT; =BERT; =DistilBERT; =RoBERTa; =DistilRoBERTa; =DeBERTa; =GPT2.



Best-performing systems get penalised for low efficiency and satisfactory fairness

Other contributions

- **Developing & evaluating LLMs:**
  - mGPT: Few-shot Learners Go Multilingual (TACL 2023, to be presented at EMNLP 2023)
  - BLOOM: A 176B-Parameter Open-Access Multilingual Language Model (under review at JMLR)
  - A Family of Pretrained Transformer Language Models for Russian (under review)
- **Creating probing suites (*ACL Workshops):**
  - RuSentEval, Morph Call, Shaking Syntactic Trees (Perturbations)
- **Organised conference events:**
  - NLP Power! The First Workshop on Efficient Benchmarking (ACL 2022)
  - Tutorial on Artificial Text Detection (INLG 2022)

Publications

* denotes equal contribution

- RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. Tatiana Shavrina, Alena Fenogenova, Anton Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, **Vladislav Mikhailov**, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. EMNLP 2020. CORE A.

- Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian. Alena Fenogenova, **Vladislav Mikhailov**, and Denis Shevelev. COLING 2021. CORE A.

- RuCoLA: Russian Corpus of Linguistic Acceptability. **Vladislav Mikhailov***, Tatiana Shamardina*, Max Ryabinin*, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. EMNLP 2022. CORE A.

Publications

* denotes equal contribution

- Findings of the RuATD Shared Task 2022 on Artificial Text Detection in Russian. Tatiana Shamardina*, **Vladislav Mikhailov***, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. Dialogue 2022. Scopus.

- Artificial Text Detection via Examining the Topology of Attention Maps. Laida Kushnareva*, Daniil Cherniavskii*, **Vladislav Mikhailov***, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. EMNLP 2021. CORE A.

- Vote'n'Rank: Revision of Benchmarking with Social Choice Theory. Mark Rofin*, **Vladislav Mikhailov***, Mikhail Florinskiy*, Andrey Kravchenko, Elena Tutubalina, Tatiana Shavrina, Daniel Karabekyan, and Ekaterina Artemova. EACL 2023. CORE A.

Thank you for your attention

References

[1] SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. Wang et al., 2019.

[2] TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. Uchendu et al., 2022.

[3] Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. Srivastava et al., 2022.

[4] Holistic Evaluation of Language Models. Liang et al., 2022.

[5] Measuring Massive Multitask Language Understanding. Hendrycks et al., 2021.

[6] The State and Fate of Linguistic Diversity and Inclusion in the NLP World. Joshi et al., 2020.

References

[7] XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. Hu et al., 2020.

[8] XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. Liang et al., 2020.

[9] FlauBERT: Unsupervised Language Model Pre-training for French. Le et al., 2020.

[10] KLEJ: Comprehensive Benchmark for Polish Language Understanding. Rybak et al., 2020.

[11] Neural Network Acceptability Judgments. Warstadt et al., 2019.

[12] The Turing test: Verbal Behavior as the Hallmark of Intelligence. Turing and Haugeland, 1950.

References

[13] All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text. Clark et al., 2021.

[14] M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection. Wang et al., 2023.

[15] Release Strategies and The Social Impacts of Language Models. Solaiman et al., 2019.

[16] TweepFake: About detecting deepfake tweets. Fagni et al., 2021.

[17] How Linguistically Fair Are Multilingual Pre-Trained Language Models? Choudhury and Deshpande, 2021.

[18] Dynaboard: An Evaluation-As-A-Service Platform for Holistic Next-Generation Benchmarking. Ma et al., 2021.

References

[19] GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Wang et al., 2018.

[20] VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation. Li et al., 2021.

[21] QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. Rogers et al., 2023.

[22] Generating Sentiment-Preserving Fake Online Reviews Using Neural Language Models and Their Human- and Machine-based Detection. Adelani et al., 2020.

[23] Topological Data Analysis for Speech Processing. Tulchinskii et al., 2022.

[24] Acceptability Judgements via Examining the Topology of Attention Maps. Cherniavskii et al., 2022.

References

[25] Dynabench: Rethinking Benchmarking in NLP. Kiela et al., 2021.

[26] Deep Reinforcement Learning at the Edge of the Statistical Precipice. Agarwal et al., 2021.