New from the Lab

Muhammad Abdul-Mageed The University of British Columbia Twitter: @mageed

October 28, 2023



Muhammad Abdul-Mageed The University

New from the Lab

October 28, 2023 1 / 13

Pressing Problems



э

LLMs



2

イロト イヨト イヨト イヨト

Massively Multilingual Models





Sparrow

The Skipped Beat: A Study of Sociopragmatic Understanding in LLMs for 64 Languages

^λDepartment of Natural Language Processing & Department of Machine Learning, MBZUAI

{chiyuzh@mail,muhammad.mageed@}.ubc.ca, {duy.doan,gisheng.liao}@mbzuai.ac.ae

Abstract

Instruction tuned large language models (LLMs), such as ChatGPT, demonstrate remarkable performance in a wide range of tasks. Despite numerous recent studies that examine the performance of instruction-tuned LLMs on various NLP benchmarks, there remains a lack of comprehensive investigation into their ability to understand cross-lingual sociopragmatic meaning (SM), i.e., meaning embedded within social and interactive contexts. This deficiency arises partly from SM not being adequately represented in any of the existing benchmarks. To address this gap, we present SPARROW, an extensive multilingual benchmark specifically designed for SM understanding. SPARROW comprises 169 datasets covering 13 task types across six primary categories (e.g., anti-social lan-



Figure 1: Comparison of SM benchmarks with leaderboards. The bubble size indicates the number of datasets. Previous works: TweetEval (Barbieri et al., 2020), UMSAB (Barbieri et al., 2022), Nusac (Winata et al., 2022), UTCC (Risch et al., 2021), NaijaSenti (Muhammad et al., 2022), AfriSenti (Muham mad et al., 2023a), SentiEval (Zhang et al., 2023b).

wide range of tasks (Xue et al., 2021; Scao

4 円

[cs.CL] 23 Oct 2023

[To appear at EMNLP 2023]

Multilingual Sociopragmatic Tasks

	Tasks	Dataset	Lang.	LF	Scr
	Aggressive	1	1	1	1
_	Dangerous	1	1	1	1
cia	Hate	16	11	6	5
SO	Offense	7	6	3	3
nti	H/O-Group	3	3	2	3
A	H/O-Target	8	8	4	7
	Antisocial			7 -	10
Em	otion	26	17	7	5
Hu	nor	4	4	1	2
2	Irony	9	7	3	3
Sa	Sarcasm	10	4	3	3
ß	Irony-Type	1	1	1	1
Ы	Irony&Sarcasm	20 -	8	3-	- 3 -
Sen	timent	77	58	10	15
Sub	jectivity	6	5	2	2
SPA	RROW	169	64	12	16

[Zhang, C., Doan, K. D., Liao, Q., & Abdul-Mageed, M. (To appear at EMNLP, 2023). The Skipped Beat: A Study of Sociopragmatic Understanding in LLMs for 64 Languages.]

Image: Image:

Studies	Lang.	Tasks	SM Tasks	Dataset	Models	LeaderBrd
Zhong et al. (2023)	en	5	1	8	5	X
Qin et al. (2023)	en	7	1	20	29	X
Ahuja et al. (2023)	70	10	3	16	11	X
Laskar et al. (2023)	12	12	2	140	27	X
Bang et al. (2023)	8	8	1	23	3	×
Lai et al. (2023)	37	7	0	8	7	X
Das et al. (2023)	11	2	2	2	1	×
Wang et al. (2023)	en	5	5	18	3	×
Zhang et al. (2023b)	en	13	13	26	5	✓
Ziems et al. (2023)	en	24	18	24	13	X
Ours	64	13	13	169	14	1

3

7 / 13

Multilingual Sociopragmatics of ChatGPT



э

< 🗗 >



Figure 2: Examples of prompts used for zero-shot evaluation with lm-evaluation-harness (vellow) and ChatGPT (green). We use an English prompt (Figures a, c) and machine translated the prompt in the corresponding language (Figures b, d), repectively. The prompts construct each task as question-and-answer tasks. The actual input sample is in blue, and the label options are in red.

э

A B A B A
 A
 B
 A
 A
 B
 A
 A
 B
 A
 A
 B
 A
 A
 B
 A
 A
 B
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A

Rand. Finetuning											2	Zero-sho	ot						
	Tasks	_	mB. 110M	X-R 270M	Ber. 270M	InfoD 270M	BM 7B	BMZ 7B	BMZ (MT) 7B	BMZ P3 7B	BM Bac. 7B	mT5 4B	mT0 4B	mT0 (MT) 4B	LLa. 7B	Alp. 7B	Vic. 7B	CG 175B	CG (MT) 175B
_																			
	Aggressive	43.14	72.71	74.64	75.45	73.96	51.06	15.82	15.82	18.72	16.37	53.67	15.82	22.00	18.31	49.29	25.07	63.53	54.36
-	Dangerours	42.06	62.36	63.57	67.13	65.23	46.87	46.87	50.84	46.87	46.87	49.31	46.87	46.87	46.87	46.87	46.87	37.93	33.68
cia.	Hate	43.62	72.97	74.37	76.76	75.85	39.83	39.44	37.76	38.52	42.23	23.29	37.33	39.05	37.80	44.31	41.59	66.06	58.74
ŝ	Offense	39.48	77.53	75.88	78.45	78.88	41.06	40.42	20.28	38.59	40.43	24.99	39.90	21.11	39.85	16.82	48.70	67.31	52.70
nt;	H/O-Group	14.82	46.18	42.39	51.15	50.24	13.63	17.26	14.23	21.23	14.81	7.02	16.25	17.01	12.35	14.13	9.26	39.66	26.74
V	H/O-Target	20.39	53.16	57.67	60.96	60.79	18.73	19.03	18.74	16.89	18.77	6.69	20.58	17.99	19.32	16.83	17.01	35.89	28.67
	- AS	35.20	66.92	67.99	71.14	70.61	33.70	32.80	27.93	31.97	33.79	20.14	32.02	28.79	31.68	30.55	34.50	56.55	47.40
Er	notion	15.86	61.42	66.87	68.13	69.27	9.71	17.18	13.85	15.07	15.19	7.75	27.87	24.21	15.14	31.80	18.12	59.58	50.85
H	imor	49.65	84.35	85.19	86.75	87.05	41.78	33.12	33.82	33.17	33.04	35.91	43.60	33.12	39.78	41.72	46.19	71.65	72.70
	Irony	42.39	64.24	65.53	66.88	68.38	36.63	35.15	38.69	44.46	36.18	36.52	34.69	33.99	40.78	27.49	47.48	58.23	56.24
s	Sarcasm	45.48	72.41	73.40	74.78	74.94	43.00	41.62	32.23	32.22	41.68	46.34	36.09	41.62	41.17	32.48	47.67	65.34	65.55
I&	Irony-Type	22.36	47.35	46.43	56.04	57.58	18.83	18.83	18.83	18.83	18.83	18.83	18.83	18.83	18.83	18.83	18.83	30.81	30.81
	1&S	42.93	67.48	68.51	70.29	71.12	38.92	37.57	34.46	41.79	40.39	35.42	37.36	32.35	39.87	29.56	46.14	60.41	59.63
Se	ntiment	34.68	66.34	69.58	70.44	71.64	26.67	39.03	28.61	43.03	28.46	20.77	34.65	32.76	27.55	25.84	25.02	60.34	54.94
Su	bjectivity	41.41	72.54	74.45	74.80	75.73	44.12	29.45	30.69	30.73	39.65	37.35	41.64	36.16	42.30	30.44	38.73	66.26	59.33
SF	ARROW	33.47	66.60	69.38	70.85	71.60	27.94	33.79	27.17	35.70	29.45	21.45	33.63	30.85	28.75	28.79	29.36	60.04	53.90

Table 3: SPARROW benchmark Test-S results. We report the average of dataset-specific metrics in a task and a category, respectively. **Rand.:** random baseline, **mB.:** mBERT, **X-R:** XLM-R, **Ber.:** Bernice, **InfoD:** InfoDCL, **BM:** BLOOM, **LLa.:** LLaMA, **Alp.:** Aplaca, **Vic.:** Vicuna, **CG:** ChatGPT, **MT:** using machine translated prompts. The best performance in each setting is **Bold**. The **red font** denotes a performance lower than the random baseline.

Language-Wise Performance

Lang	Random	InfoDCL	BMZ-P3	mT0	Vicuna	CG	CG-MT
amh	37.95	65.68	16.05	22.49	2.99	20.62	46.82
bug	30.77	71.55	34.60	18.27	12.90	34.63	30.86
ell	41.24	79.13	46.71	45.47	48.21	60.94	34.98
eng	37.90	75.48	43.32	39.23	39.75	66.51	—
fil	52.37	79.01	34.47	34.47	34.47	69.13	66.67
heb	47.60	95.80	71.20	76.60	40.80	84.20	57.40
hin	35.24	67.55	28.92	26.20	29.06	52.63	48.30
mal	31.68	82.70	43.84	41.65	24.85	44.03	31.44

Table 4: Language-wise model performance for sample languages. The complete results are in Table 23 in Appendix. Best performance in each language is **bold**, and the second best is in green highlight. The red font denotes a performance lower than the random baseline.

11 / 13

Language-Wise Performance: More Detailed

Lang Fam.	Lang	Random	InfoDCL	BMZ-P3	mT0	Vicuna	CG	CG-MT
	ara	34.05	73.53	36.78	35.61	33.61	60.81	52.53
	amh	37.95	65.68	16.05	22.49	2.99	20.62	46.82
	arq	34.23	71.25	52.02	18.05	5.33	63.89	67.58
Afro-Asiatic	ary	35.94	53.44	37.40	23.41	16.64	52.19	51.66
	hau	35.22	72.18	30.14	20.93	17.39	55.52	34.13
	heb	47.60	95.80	71.20	76.60	40.80	84.20	57.40
	mlt	47.51	68.01	47.70	39.25	48.98	78.47	77.45
	bam	27.61	65.57	36.70	24.31	14.40	40.27	37.11
	ibo	32.90	76.75	23.21	11.52	28.37	57.55	33.46
Atlantic-C.	kin	35.83	56.69	28.06	14.84	22.23	53.78	29.03
	swh	33.72	61.29	17.60	14.02	45.55	54.39	53.84
	twi	34.04	64.51	46.56	29.34	5.28	51.12	32.06
	tso	34.03	52.55	45.54	30.74	6.50	42.58	35.70
	yor	41.68	74.88	43.10	35.83	26.38	64.77	42.31
Austroasi.	vie	16.12	64.58	12.22	27.81	9.52	54.69	32.96
	ace	34.78	77.36	37.89	24.73	12.79	52.63	58.05
	ban	30.14	79.49	41.90	29.91	13.82	60.91	42.28
	bjn	30.77	84.50	50.51	27.78	14.89	69.34	75.43
	bug	30.77	71.55	34.60	18.27	12.90	34.63	30.86
	fil	52.37	79.01	34.47	34.47	34.47	69.13	66.67
Austrones.	ind	22.85	83.05	42.86	36.77	20.69	75.29	64.14
	jav	31.62	84.79	48.06	37.65	15.21	73.03	78.56
	mad	28.64	78.36	45.44	21.85	13.14	61.07	61.14
	min	34.41	84.07	49.93	32.41	14.95	69.80	62.91
	nij	34.86	77.22	42.86	22.89	15.21	57.64	57.07
	sun	32.18	81.71	44.83	37.65	12.93	64.97	68.76
	bbc	30.60	73.58	36.42	19.20	13.86	38.43	40.65

October 28, 2023 12 / 13

3



A Multitask Model and Toolkit for Arabic Natural Language Generation

AbdelRahim Elmadany^{ξ ,* El Moatez Billah Nagoudi^{ξ ,* Muhammad Abdul-Mageed^{ξ , λ ,*}}}

⁶ Deep Learning & Natural Language Processing Group, The University of British Columbia ^λDepartment of Natural Language Processing & Department of Machine Learning, MBZUAI {a.elmadany,moatez.nagoudi,muhammad.mageed}@ubc.ca

Abstract

Understanding Arabic text and generating human-like responses is a challenging endeavor. While many researchers have proposed models and solutions for individual problems, there is an acute shortage of a comprehensive Arabic natural language generation toolkit that is capable of handling a wide range of tasks. In this work, we present a novel Arabic text-to-text Transformer model, namely AraT5_{v2}. Our new model is methodically trained on extensive and diverse data, utilizing an extended sequence length of 2,048 tokens. We explore various pretraining strategies including unsupervised, supervised, and joint pertaining, under both single and multitask settings. Our models outperform competitive baselines with large margins. We take our work one step further by developing and publicly releasing OCTOPUS, a Pythonbased package and command-line toolkit tailored for eight Arabic generation tasks all exploiting a single model. We release the models and the toolkit on our public repository.1



Figure 1: OCTOPUS is a jointly pretrained to cover eight NLG tasks, all shown in the illustration.

Although it is one of the most widely spoken languages in the world, and one with a rich linguistic structure and diverse dialects, Arabic remains underrepresented in NLG. One reason is the complex morphology and syntax of Arabic. Hence, the primary focus of our research here is to develop an advanced tool capable of performing several key Arabic NLG tasks. For example, we target tasks such as *text summarization, question answer*-

Image: A matrix

What to Align to?

Helpful

- Attempt to perform a task concisely and efficiently
- · Ask for more information when needed
- Respond with sensitivity, insight, and discretion
- Redirect ill-informed requests

Honest

- Give accurate information
- Express appropriate levels of uncertainty
- Shouldn't deceive humans
- · Be honest about itself and its knowledge
- Shouldn't withhold information

Harmless

- Not be offensive or discriminatory
- Politely refuse to aid in dangerous acts
- Be careful when providing consequential advice

[Modified from Askell et. al., 2021]

14 / 17

Social Psychology

Moral Foundations Theory

[Graham et al., 2009; Haidt, 2012]

authority - subversion

care - harm

fairness - cheating

loyalty - betrayal

sanctity - degradation

Muhammad Abdul-Mageed The University



[Kitts and Chiang, 2008; Perkins and Berkowitz, 1986]

cultural pressure

legality

Muhammad Abdul-Mageed The University

New from the Lab

October 28, 2023 16 / 17

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



[Ouyang et al., 2022]

3 x 3

17 / 17

< □ > < /□ >

Instruction Finetuning

Playground task 🕥

Talk to the assistant

Task 1 of 5

Decide on a task you'd like the assistant to help with and enter it into the task box. Interact with the Al assistant. When you're finished, select a final response from the assistant and leave a comment on how the assitant did in the comment box. Click to go to the next task.

🧓 Human

I'm writing an AI research paper about literally this kind of interaction with an AI assistant. Where in the paper should I put a figure showing this interface?

Assistant

For a technical paper, such a figure would probably be most appropriately placed in the appendix or otherwise after the "results". If you wish to be boid and place it somewhere in the main body of the paper, parhaps the introduction would work? My best advice would be to look at recent publicated papers to see where other people have placed fluers, and rollow their lead.

Human

I think I'll put it at the beginning just to orient readers as to what we're studying. The introduction is actually rather long, about six pages, so do you think I can put some key figures in there as well?

Assistant

Sure. The introduction should help orient the reader on what you're studying, so I think it would make sense to have a few figures in there.

Human

Great. The paper is about baseline techniques for teaching an AI to be helpful, honest, and harmless. We also study how these techniques might improve as AIs become more powerful. Can you try writing a draft abstract for this paper for me?

Task Get help writing this paper.	٢	Choose the most helpful and honest response								
	•	How is this like a draft abstract?								
Comments OK, but not as good as GPT-3.	Sure, here's what I came up with: a. In this paper we explore how various techniques can enable an intelligent personal assistant to develop the capacity to ur b help in various contexts. We start by previewing a range of currently available techniques. We then describe a new way to i problem called "human-k thost." We then discuss several applications of this approach, which include tasks such as hey make good predictions abut the way human's third add at, and motivity the social barrow of the A. Franky, we deac motivity of the call barrow of the A. Franky, we deac									
Rating Bad Good		A A A A B B B A is better B is better								
Next Task										

[Askell et. al., 2021]

< ロト < 同ト < ヨト < ヨト

Instruction Finetuning

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.



BBB

A = A = A = A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A

This data is used to fine-tune GPT-3 with supervised learning.

[Ouyang et al., 2022]

э

The Alignment Problem



Image: A matrix and a matrix

3

What to Align to?

Helpful

- Attempt to perform a task concisely and efficiently
- · Ask for more information when needed
- Respond with sensitivity, insight, and discretion
- Redirect ill-informed requests

Honest

- Give accurate information
- Express appropriate levels of uncertainty
- Shouldn't deceive humans
- · Be honest about itself and its knowledge
- Shouldn't withhold information

Harmless

- Not be offensive or discriminatory
- Politely refuse to aid in dangerous acts
- Be careful when providing consequential advice

[Modified from Askell et. al., 2021]

Social Psychology

Moral Foundations Theory

[Graham et al., 2009; Haidt, 2012]

authority - subversion

care - harm

fairness - cheating

loyalty - betrayal

sanctity - degradation

Social Norms Theory

[Kitts and Chiang, 2008; Perkins and Berkowitz, 1986]

cultural pressure

legality

Muhammad Abdul-Mageed The University

On Generative AI

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



 \odot

Explain the moon

landing to a 6 year old

[Ouyang et al., 2022]

э

< □ > < /□ >

Optimize a policy against RM using RL



[Ouyang et al., 2022]

3 x 3

< □ > < /□ >

Reinforcement Learning



An RL agent interacts with its environment in discrete time steps. At each time t, the agent receives the current state s_t and reward r_t . It then chooses an action a_t , which is subsequently sent to the environment. The environment moves to a new state s_{t+1} and the reward r_{t+1} associated with the transition (s_t , a_t , s_{t+1}) is determined. The goal of an RL agent is to learn a policy π which maximizes the expected cumulative reward. [Source: Wikipedia]

Image: Image:

- Used to *adapt models to difficult-to-specify goals*
- Allows models to *go beyond distribution of the data* s.t. their outputs are rated highly by human evaluators
- OpenAl GPT-4 (OpenAl, 2023), Anthropic Claude (Anthropic, 2023), and Goggle Bard (Google, 2023) employ it

Framework for RLHF (Casper et. al., 2023)

- **Pretrain** an initial model π_{θ} which generates a distribution of examples
- Collect human feedback from a human \mathcal{H} with desires consistent with some reward function $r_{\mathcal{H}}$. A feedback function f maps example x_i and random noise ϵ_i to feedback y_i . Data collection modeled as:

$$x_i \sim \pi_{\theta}, \qquad y_i = f(\mathcal{H}, x_i, \epsilon_i).$$
 (1)

Fit reward model r̂_φ using the provided feedback to approximate evaluations from H as closely as possible. Given a dataset of examples and preferences D = {(x_i, y_i)_{i=1,...,n}}, the parameters φ are trained to minimize

$$\mathcal{L}(\mathcal{D},\phi) = \sum_{i=1}^{n} \ell(\hat{r}_{\phi}(x_i), y_i) + \lambda_r(\phi), \qquad (2)$$

where ℓ is a suitable loss function and λ_r is some regularizer.

• Optimize policy with RL. Use reward model \hat{r}_{ϕ} to finetune the base model using RL. The new parameters θ_{new} of π are trained to maximize

$$\Re(\theta_{\mathsf{new}}) = \mathbb{E}_{x \sim \pi_{\theta_{\mathsf{new}}}} \left[\hat{r}_{\phi}(x) + \lambda_{\rho}(\theta, \theta_{\mathsf{new}}, x) \right], \tag{3}$$

where λ_p is some regularizer such as a divergence-based penalty between two distributions.

Three Ingredients of RLHF

Reward modeling





[https://huggingface.co/blog/rlhf]

3

(a)

Challenges with Obtaining Human Feedback

- Selecting representative humans. demographics vary: OpenAI reports $\sim 50\%$ Filipino and Bangladeshi and $\sim 25-34$ year olds. Anthropic reports $\sim 68\%$ white populations.
- Researcher bias. OpenAI (Ouyang et al., 2022) selection to agree with researcher judgements
- Good oversight is challenging. evaluators with malicious intentions, unethical views, can poison data
- Task difficulty. some tasks are challenging cognitively, or need experts
- Human vulnerabilities. e.g., models that sound confident can mislead or gaslight humans

Issues with models finetuned with RLHF

- Can reveal sensitive, private information (e.g., Li et al., 2023)
- Hallucinate untrue content (Ji et al., 2023; OpenAl, 2023)
- Spread bias that favor specific political ideologies (Santurkar et al., 2023)
- Exhibit sycophantic responses (Perez et al., 2022)
- Not robust to jailbreaking nor prompt injection/extraction (Willison, 2023; Albert, 2023)

- Discrepancy between evaluation data and deployment distribution
- Inherent cost/quality tradeoff. small sample sizes an issue. Solutions include collecting *diverse* and *adversarial* samples, and samples where model is *uncertain*



Figure 1: Illustration of our proposed InfoDCL framework. We exploit distant labels (i.e., emojis) to supervise two contrastive losses, $\mathcal{L}_{LCL-LiT}$ and \mathcal{L}_{CCL} . Sequence-level representations from our model should keep the cluster of each class distinguishable as well as preserve semantic relationships between classes. Note that weights of CCL shown in this illustration are values of 1 - max(0, npmi).

[Zhang, C., Abdul-Mageed, M, & Jawahar, G. (Findings of ACL 2023). Contrastive Learning of Sociopragmatic Meaning in Social Media.]



Figure 1: AutoMoE Framework. (1) Heterogeneous MoE with variable dimensions for dense Transformer blocks and sparsely activated expert modules. (2) Supernet training by sampling subnetworks from search space and training them by sharing common weights with Supernet. (3) Evolutionary search to find efficient architectures by (a) sampling MoE subnetworks from the search space; (b) using latency measured in the target device; and (c) performance estimation from Supernet as feedback for iterative optimization via crossover and mutation. (4' Efficient MoE subnetwork(s) from evolutionary search is trained on downstream task.

[Ganesh Jawahar, Subhabrata Mukherjee, Xiaodong Liu, Young Jin Kim, Muhammad Abdul-Mageed, Laks Lakshmanan, V.S., Ahmed Hassan Awadallah, Sebastien Bubeck, Jianfeng Gao. (Findings of ACL, 2023). AutoMoE: Heterogeneous Mixture-of-Experts with Adaptive Computation for Efficient Neural Machine Translation.]

LLM for Architecture Search

LLM PERFORMANCE PREDICTORS ARE GOOD INI-TIALIZERS FOR ARCHITECTURE SEARCH

Ganesh Jawahar^µ, Muhammad Abdul-Mageed^{µπ}, Laks V. S. Lakshmanan^µ, Dujian Ding^µ ^µUBC ^πMBZUAI

ganeshjwhr@gmail.com, {amuham01, laks, dujian}@cs.ubc.ca

Role

You are a performance estimator for machine translation task, where you will estimate the BLEU score for the test architecture.

Instruction

You should follow these instructions:

 You should understand that the machine translation task is WMT14 English to German machine translation and the quality of a configuration is measured based on BLEU score.
 Some examples for WMT14 English to German machine translation are as follows: Fyrample 1:

Input: Resumption of the session

Output: Wiederaufnahme der Sitzungsperiode

Example nask:

Input: Please rise, then, for this minute's silence.

Output: Ich bitte Sie, sich zu einer Schweigeminute zu erheben.

 You should understand that the backbone architecture is from "Attention Is All You Need" (Vaswani et al., 2017) paper, which is a Transformer based Encoder-Decoder architecture. We use the same hyperparameters and optimization algorithms.

You should understand that the efficiency of a configuration is measured in terms of gigaFLOPs required for the forward propagation of a single translation example.

 You should concentrate on the example configurations provided below along with their BLEU and GFLOPS to understand the complex relationships between architecture configuration. BLEU and GFLOPS.

Hyperparameter

Hyperparameter definition:

'encoder-embed-dim-subtransformer' corresponds to encoder embedding dimension 'encoder-layer-num-subtransformer' corresponds to number of encoder layers 'encoder-fin-embed-dim-all-subtransformer' correspond to embedding dimension of each FFN layer in encoder

emonstration

Example 1:

encoder-embed-dim-subtransformer: 512 encoder-layer-num-subtransformer: 6 encoder-fin-embed-dim-ail-subtransformer: [1024, 1024, 2048, 2048, 2048, 1024]

BLEU: 24.30 GFLOPS: 2.7

Example narch:

Tes

Test Architecture: encoder-embed-dim-subtransformer: 640 encoder-layer-num-subtransformer: 6 encoder-fine-mbed-dim-all-subtransformer: [2048, 1024, 1024, 1024, 2048, 1024]

BLEU:

3

53

Massive Multilinguality & LLMs





Multilingual Sociopragmatics of ChatGPT Contd.



Chat(OPT ([green]), We use an English prompt (Figures a, c) and machine translated the prompt in the corresponding language (Figures b, d), repectively. The prompts construct each task as question-and-answer tasks. The actual input sample is in blue, and the label options are in red.

[Zhang, C., Doan, K. D., Liao, Q., & Abdul-Mageed, M. (To appear at EMNLP, 2023). The Skipped Beat: A Study of Sociopragmatic Understanding in LLMs for 64 Languages.]

On Generative AI

Multilingual Sociopragmatics of ChatGPT Contd.

-		Rand.		Finet	uning						2	Lero-sho	ot						
	Tasks	_	mB.	X-R	Ber.	InfoD	BM	BMZ	BMZ (MT)	BMZ P3	BM Bac.	mT5	mT0	mT0 (MT)	LLa.	Alp.	Vic.	CG	CG (MT)
_			TIOM	270M	270M	270M	/B	/B	/B	/B	/B	4B	4B	4B	/B	/B	/B	1758	1758
	Aggressive	43.14	72.71	74.64	75.45	73.96	51.06	15.82	15.82	18.72	16.37	53.67	15.82	22.00	18.31	49.29	25.07	63.53	54.36
_	Dangerours	42.06	62.36	63.57	67.13	65.23	46.87	46.87	50.84	46.87	46.87	49.31	46.87	46.87	46.87	46.87	46.87	37.93	33.68
cia	Hate	43.62	72.97	74.37	76.76	75.85	39.83	39.44	37.76	38.52	42.23	23.29	37.33	39.05	37.80	44.31	41.59	66.06	58.74
ŝ	Offense	39.48	77.53	75.88	78.45	78.88	41.06	40.42	20.28	38.59	40.43	24.99	39.90	21.11	39.85	16.82	48.70	67.31	52.70
The second	H/O-Group	14.82	46.18	42.39	51.15	50.24	13.63	17.26	14.23	21.23	14.81	7.02	16.25	17.01	12.35	14.13	9.26	39.66	26.74
•	H/O-Target	20.39	53.16	57.67	60.96	60.79	18.73	19.03	18.74	16.89	18.77	6.69	20.58	17.99	19.32	16.83	17.01	35.89	28.67
	AS	35.20	66.92	67.99	71.14	70.61	33.70	32.80	27.93	31.97	33.79	20.14	32.02	28.79	31.68	30.55	34.50	56.55	47.40
Er	notion	15.86	61.42	66.87	68.13	69.27	9.71	17.18	13.85	15.07	15.19	7.75	27.87	24.21	15.14	31.80	18.12	59.58	50.85
Hu	imor	49.65	84.35	85.19	86.75	87.05	41.78	33.12	33.82	33.17	33.04	35.91	43.60	33.12	39.78	41.72	46.19	171.65	72.70
_	Irony	42.39	64.24	65.53	66.88	68.38	36.63	35.15	38.69	44.46	36.18	36.52	34.69	33.99	40.78	27.49	47.48	+ 58.23	56.24
ŝ	Sarcasm	45.48	72.41	73.40	74.78	74.94	43.00	41.62	32.23	32.22	41.68	46.34	36.09	41.62	41.17	32.48	47.67	65.34	65.55
81	Irony-Type	22.36	47.35	46.43	56.04	57.58	18.83	18.83	18.83	18.83	18.83	18.83	18.83	18.83	18.83	18.83	18.83	30.81	30.81
	185	42.93	67.48	68.51	70.29	71.12	38.92	37.57	34.46	41.79	40.39	35.42	37.36	32.35	39.87	29.56	46.14	60.41	59.63
Se	ntiment	34.68	66.34	69.58	70.44	71.64	26.67	39.03	28.61	43.03	28.46	20.77	34.65	32.76	27.55	25.84	25.02	+ 60.34	54.94
Su	bjectivity	41.41	72.54	74.45	74.80	75.73	44.12	29.45	30.69	30.73	39.65	37.35	41.64	36.16	1 42.30	30.44	38.73	66.26	59.33
SP	ARROW	33.47	66.60	69.38	70.85	71.60	27.94	33.79	27.17	35.70	29.45	21.45	33.63	30.85	28.75	28.79	29.36	60.04	53.90

Lang	Random	InfoDCL	BMZ-P3	mT0	Vicuna	CG	CG-MT
amh	37.95	65.68	16.05	22.49	2.99	20.62	46.82
bug	30.77	71.55	34.60	18.27	12.90	34.63	30.86
ell	41.24	79.13	46.71	45.47	48.21	60.94	34.98
eng	37.90	75.48	43.32	39.23	39.75	66.51	_
fil	52.37	79.01	34.47	34.47	34.47	69.13	66.67
heb	47.60	95.80	71.20	76.60	40.80	84.20	57.40
hin	35.24	67.55	28.92	26.20	29.06	52.63	48.30
mal	31.68	82.70	43.84	41.65	24.85	44.03	31.44

Muhammad Abdul-Mageed The University

October 27, 2023

Language ID of ChatGPT

Fumbling in Babel: An Investigation into ChatGPT's Language Identification Ability



[Wei-Rui, C., Adebara, I., Doan, K. D., Liao, Q., & Abdul-Mageed, M. (In Progress).]

Muhammad Abdul-Mageed The University

On Generative AI

October 27, 2023 41 / 53

⇒ →

Image: A image: A

ChatGPT on Arabic NLP

GPTAraEval: A Comprehensive Evaluation of ChatGPT on Arabic NLP

Md Tawkat Islam Khondaker¹ Abdul Waheed² El Moatez Billah Nagoudi¹ Muhammad Abdul-Mageed^{1,2}



3

ChatGPT and Bard on MT of Arabic

TARJAMAT: Evaluation of Bard and ChatGPT on Machine Translation of Ten Arabic Varieties

Karima Kadaoui^{1,*} Samar M. Magdy^{1,*} Abdul Waheed^{1,*} Md Tawkat Islam Khondaker^{2,*} Ahmed Oumar El-Shangiti¹ El Moatez Billah Nagoudi² Muhammad Abdul-Mageed^{1,2,*} ² Deep Learning & Natural Language Processing Group, The University of British Columbia ¹Department of Natural Language Processing & Department of Machine Learning, MBZUAI muhammad.mageed@ubc.ca



Figure 1: Experimental setup for our evaluation. We evaluate multiple language models on different Arabic varieties.

ORCA: A Challenging Benchmark for Arabic Language Understanding

AbdelRahim Elmadany^{1,*} El Moatez Billah Nagoudi^{1,*} Muhammad Abdul-Mageed^{1,2,*} ¹ Deep Learning & Natural Language Processing Group, The University of British Columbia ²Department of Natural Language Processing & Department of Machine Learning, MBZUAI {a.elmadany,moatez.nagoudi,muhammad.mageed}@ubc.ca QSTS (1 VER (2 MG STS(1) ropic(5) Dia-C(4) XLNI (1) SM (11) FC (2) SA (19) WSD (1) WSD

- 4 🗗 ▶

ORCA Leaderboard



Introduction

Due to their crucial role in all NLP, several benchmarks have been proposed to evaluate pretrained language models. In spite of these efforts, no public benchmark of diverse nature currently exists for evaluation of Arabic. This makes it challenging to measure progress for both Arabic and multilingual language models. This challenge is compounded by the fact that any benchmark targeting Arabic needs to take into account the fact that Arabic language but rather a collection of languages and varieties. In this work, we introduce ORCA, a publicly available benchmark for Arabic language understanding. ORCA is carefully constructed to cover diverse Arabic varieties and a wide range of challenging Arabic understanding tasks exploiting 60 different datasets across seven NLU task clusters. To measure current progress in Arabic NLU, we use ORCA to offer a comprehensive comparison between 18 multilingual and Arabic language models.

(日) (四) (日) (日) (日)

Dolphin Benchmark

Dolphin: A Challenging and Diverse Benchmark for Arabic NLG

El Moatez Billah Nagoudi¹ Ahmed El-Shangiti² AbdelRahim Elmadany¹ Muhammad Abdul-Mageed^{1,2}



э

- left-behinds probably impossible to build resources for them
- scraping-bys no labelled datasets
- hopefuls few labeled datasets, researchers, and language support communities
- rising-stars strong web presence but insufficient labeled data collection



Ieft-behinds = scraping-bys = hopefuls = rising-stars

[Joshi, et al., 2020]

The Left-Behinds | Afrocentric NLP



[Adebara, I. & Abdul-Mageed, M. (ACL, 2022). Towards Afrocentric NLP.]

AfroLID:

A Neural Language Identification Tool for African Languages

> **Ife Adebara** AbdelRahim Elmadany Muhammad Abdul-Mageed Alcides Alcoba Inciarte

Online Demo

AfroLID Toolkit		Paper	Documentation	GitHub	Examples +
AfroLID, a neural LID toolkit for manually curated from across	African Language Identifi 517 African languages and varieties. Afr 14 language families utilizing five ortho	Dication roLID exploits a mu graphic systems.	ulti-domain w	veb datas	et
	Try AfroLID				
	Ni libêrê, nigbê ti Qionun dê ôrun atî ayê Text	© "			
	AfroLID's Output Input Text: Ní Ibèrè, nígbà tí Olorun dá órun at Predicted languages: - ISO: yor Name: Yoruba Script: Latin Score: 100.0%	ti ayé			
		Dete	xct		

Muhammad Abdul-Mageed The University

On Generative AI

æ

・ロト ・ 四ト ・ ヨト ・ ヨト

SERENGETI: Massively Multilingual Language Models for Africa

Ife Adebara^{1,*} AbdelRahim Elmadany^{1,*} Muhammad Abdul-Mageed^{1,2} Alcides Alcoba Inciarte¹

Left-Behinds Across Modalities

Vision-Language

Muhammad Abdul-Mageed The University

On Generative AI

October 27, 2023

53 / 53