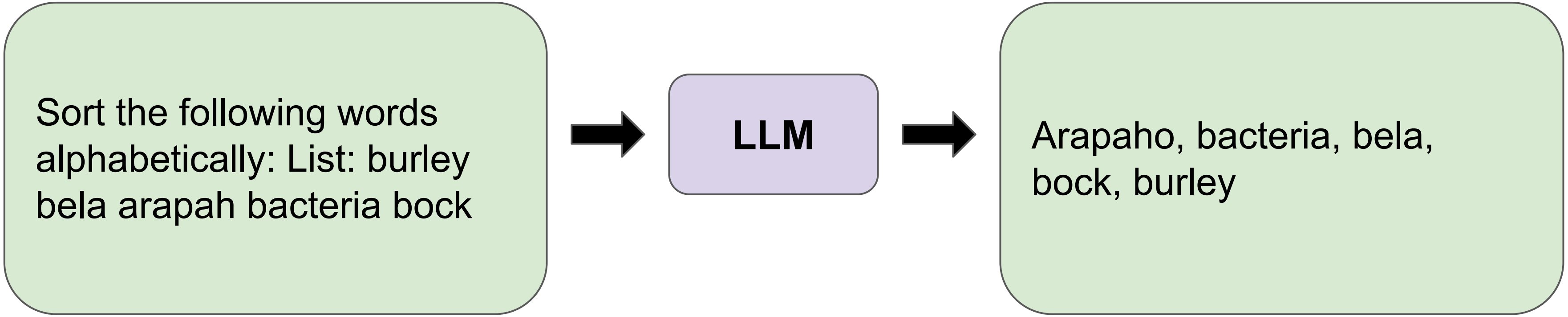
Guess the Instruction! Flipped Learning Makes Language Models Stronger Zero-Shot Learners

Seonghyeon Ye[1], Doyoung Kim[1],
Joel Jang[1], Joongbo Shin[2], Minjoon Seo[1]

[1] KAIST AI Kim Jaechul Graduate School

[2] LG AI Research

Zero-shot Task Generalization of Large Language Models

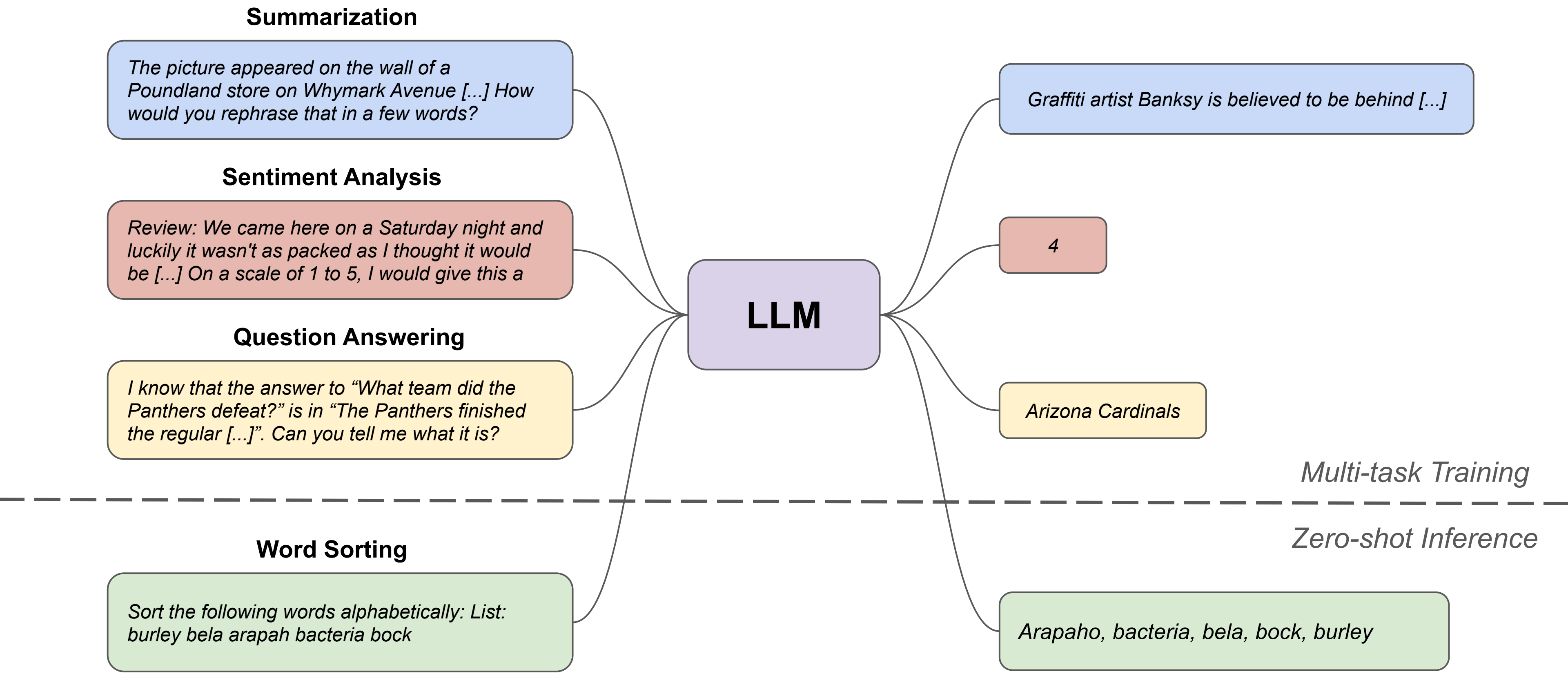


Sort the following words alphabetically: List: burley bela arapah bacteria bock

LLM

Arapaho, bacteria, bela, bock, burley

Instruction Tuning (Meta-training)



Summarization

The picture appeared on the wall of a Poundland store on Whymark Avenue [...] How would you rephrase that in a few words?

Sentiment Analysis

Review: We came here on a Saturday night and luckily it wasn't as packed as I thought it would be [...] On a scale of 1 to 5, I would give this a

Question Answering

I know that the answer to “What team did the Panthers defeat?” is in “The Panthers finished the regular [...]”. Can you tell me what it is?

Word Sorting

Sort the following words alphabetically: List: burley bela arapah bacteria bock

LLM

Graffiti artist Banksy is believed to be behind [...]

4

Arizona Cardinals

Arapaho, bacteria, bela, bock, burley

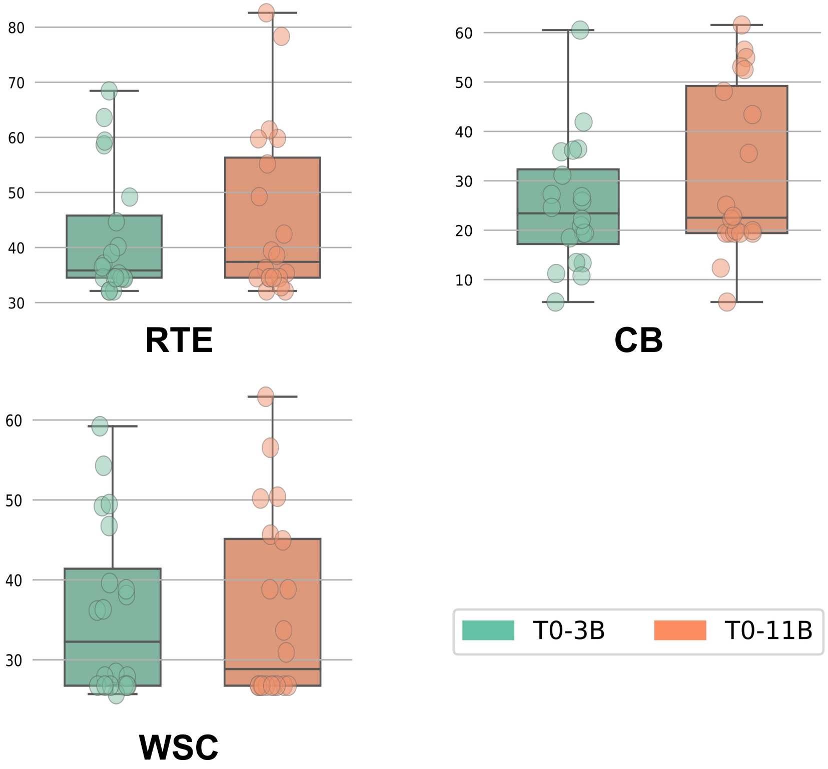
Multi-task Training

Zero-shot Inference

Limitation of Instruction Tuning

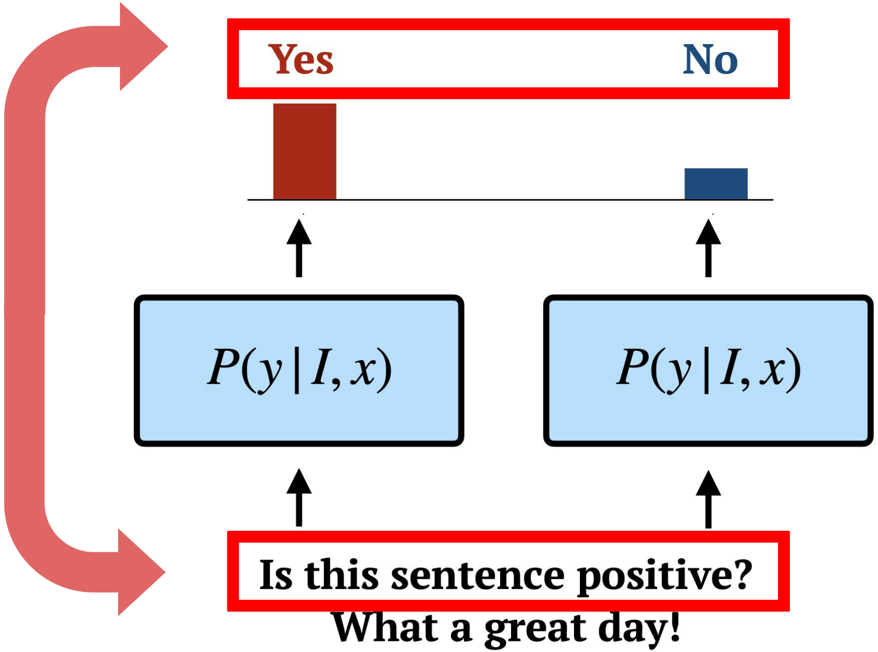
yes	no
true	false
positive	negative
right	wrong
correct	incorrect
agree	disagree
good	bad
guaranteed	impossible
always	never
affirmative	contradicting
exactly	not ever
undoubtedly	not at all
fine	disagreeable
good enough	cannot be
definitely	never
unquestionable	no way
yep	nope
yea	nah
without doubt	refused
willing	unwilling

Unseen Labels



Flipped Learning

$(I, x, y) =$ **(Is this sentence positive?, What a great day!, Yes)**



Yes **No**

$P(y \mid I, x)$

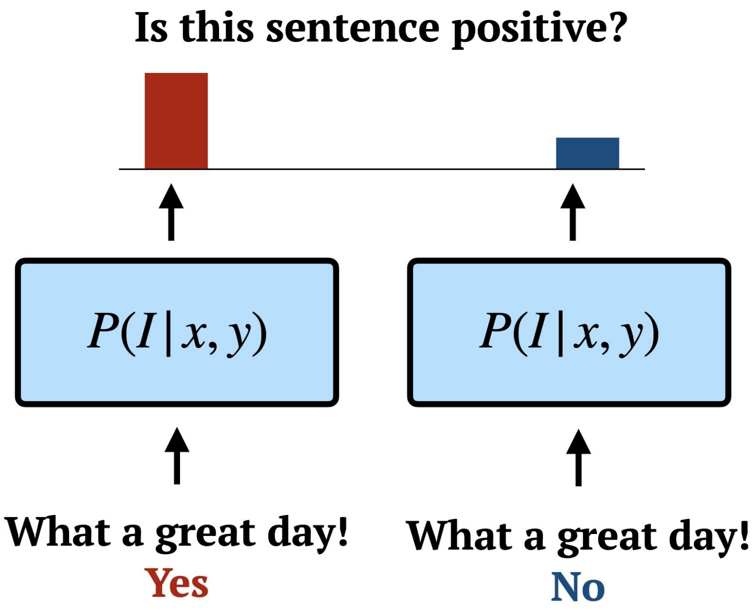
$P(y \mid I, x)$

Is this sentence positive?

What a great day!

Flipped Learning

$(I, x, y) =$ **(Is this sentence positive?, What a great day!, Yes)**



Is this sentence positive?

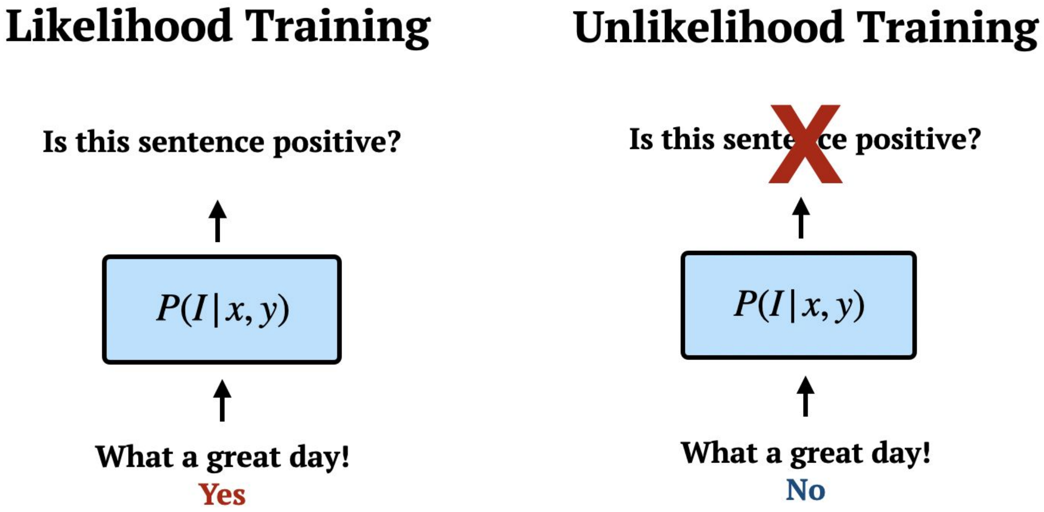
$P(I \mid x, y)$

$P(I \mid x, y)$

What a great day!
Yes

What a great day!
No

Flipped Learning

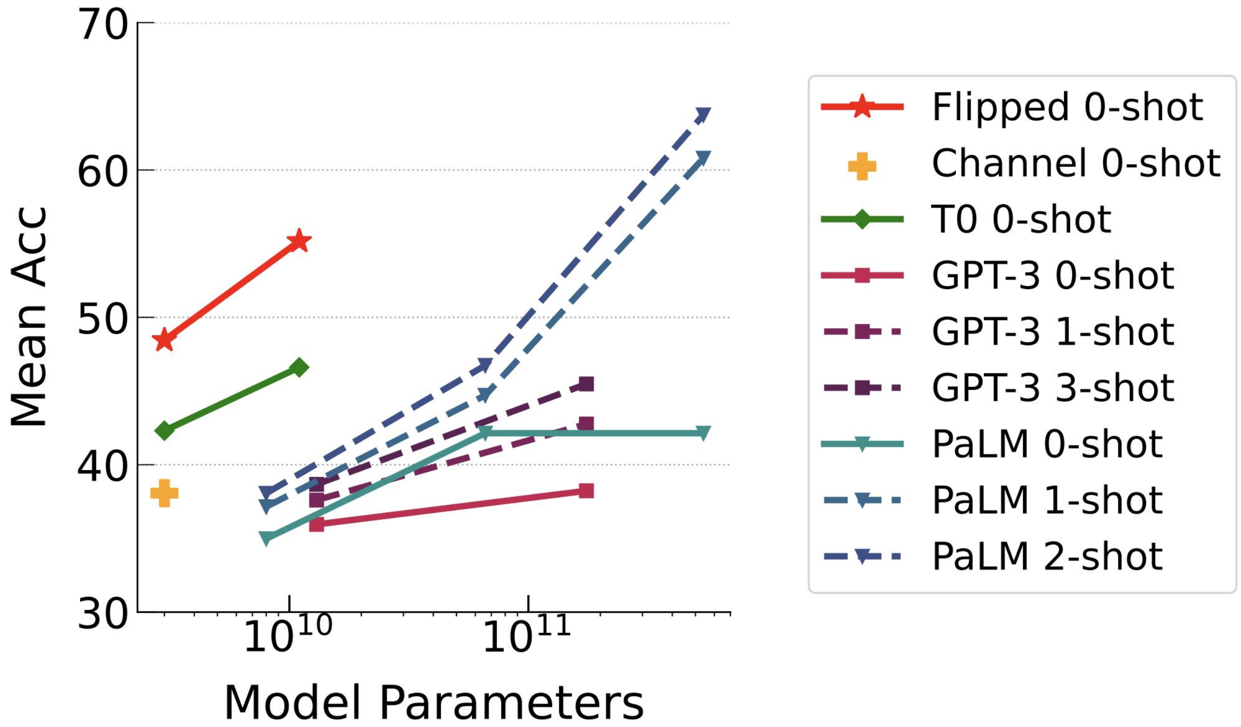


$$L_{LM} = -\sum_{t=1}^{T} \log P(I_t|x, l_c, I_{<t})$$

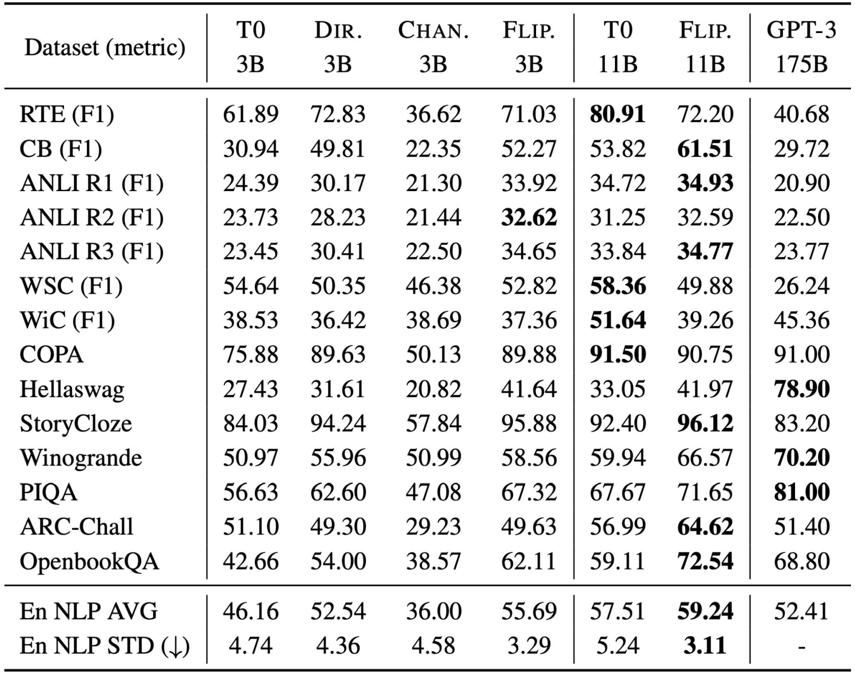
$$L_{UL} = -\sum_{t=1}^{T} \log(1 - P(I_t|x, l_{c'}, I_{<t}))$$

$$L = L_{LM} + \lambda L_{UL}$$

BIG-Bench Results (3B, 11B)



Result of Flipped



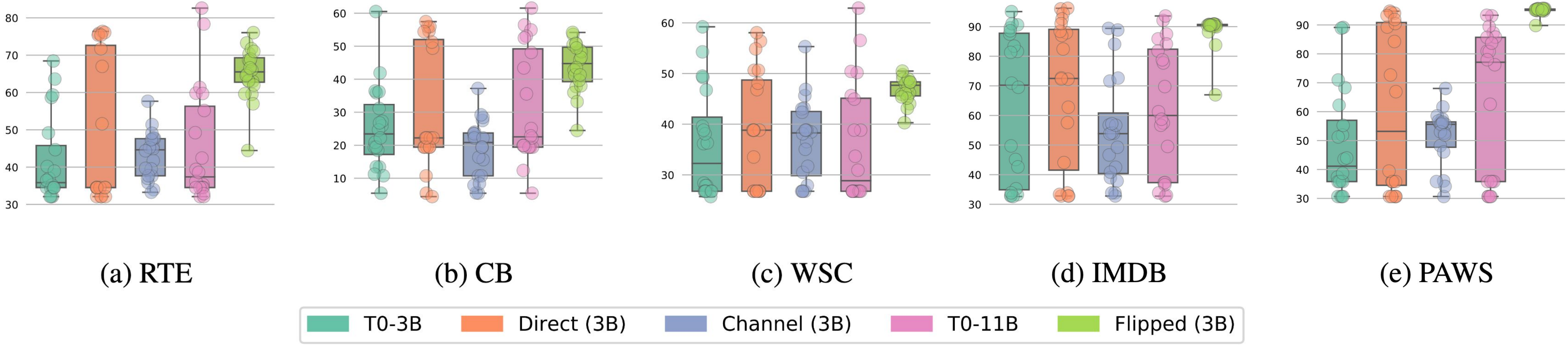
Dataset (metric)	T0 3B	DIR. 3B	CHAN. 3B	FLIP. 3B	T0 11B	FLIP. 11B	GPT-3 175B
RTE (F1)	61.89	72.83	36.62	71.03	**80.91**	72.20	40.68
CB (F1)	30.94	49.81	22.35	52.27	53.82	**61.51**	29.72
ANLI R1 (F1)	24.39	30.17	21.30	33.92	34.72	**34.93**	20.90
ANLI R2 (F1)	23.73	28.23	21.44	**32.62**	31.25	32.59	22.50
ANLI R3 (F1)	23.45	30.41	22.50	34.65	33.84	**34.77**	23.77
WSC (F1)	54.64	50.35	46.38	52.82	**58.36**	49.88	26.24
WiC (F1)	38.53	36.42	38.69	37.36	**51.64**	39.26	45.36
COPA	75.88	89.63	50.13	89.88	**91.50**	90.75	91.00
Hellaswag	27.43	31.61	20.82	41.64	33.05	41.97	**78.90**
StoryCloze	84.03	94.24	57.84	95.88	92.40	**96.12**	83.20
Winogrande	50.97	55.96	50.99	58.56	59.94	66.57	**70.20**
PIQA	56.63	62.60	47.08	67.32	67.67	71.65	**81.00**
ARC-Chall	51.10	49.30	29.23	49.63	56.99	**64.62**	51.40
OpenbookQA	42.66	54.00	38.57	62.11	59.11	**72.54**	68.80
En NLP AVG	46.16	52.54	36.00	55.69	57.51	**59.24**	52.41
En NLP STD (↓)	4.74	4.36	4.58	3.29	5.24	**3.11**	-

Flipped Models lead to **higher accuracy and lower variance** (robust to different instruction wordings)

Flipped Learning

Why does Flipped Learning works? ⇒ Label Generalization !

Previous work implies that during training of language models, the space of generation is easy to exploit than the models condition on. ⇒ output space overfitting



(a) RTE

(b) CB

(c) WSC

(d) IMDB

(e) PAWS

[1] Min et al (2022) Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?
[2] Webson and Pavlick (2022) Do Prompt-Based Models Really Understand the Meaning of Their Prompts?

Q & A