Editing Models with Task Arithmetic

Gabriel Ilharco[*1] **Marco Tulio Ribeiro**[2] **Mitchell Wortsman**[1] **Suchin Gururangan**[1]
Ludwig Schmidt[1,3] **Hannaneh Hajishirzi**[1,3] **Ali Farhadi**[1]
[1]University of Washington [2]Microsoft Research [3]Allen Institute for AI

International Conference on Learning (ICLR) 2023

Outline

1) Motivation
2) Literature Review
3) Main Idea
4) Applications
5) Intuition
6) Limitations
7) Future Directions

Motivation: Modifying the Behaviour of Pre-trained Models

Mitigating biases from pre-trained models

INSIDER

Newsletters Log in Subscribe

HOME > NEWS

ChatGPT could be used for good, but like many other AI models, it's rife with racist and discriminatory bias

Article | Published: 23 March 2022

Large pre-trained language models contain human-like biases of what is right and wrong to do

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf & Kristian Kersting

Nature Machine Intelligence **4**, 258–268 (2022) | Cite this article

2681 Accesses | **7** Citations | **155** Altmetric | Metrics

AI

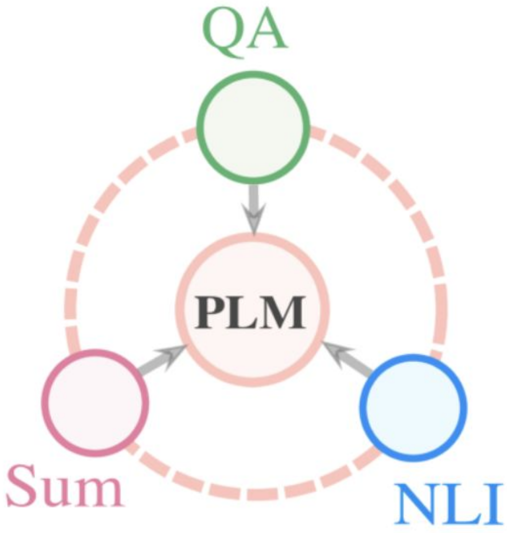
Here are a few ways GPT-3 can go wrong

Liz O'Sullivan, John Dickerson / 9:45 AM CDT • August 7, 2020

Comment

Motivation: Modifying the Behaviour of Pre-trained Models

Multi-task Capabilities



**PLM - Pretrained Language Model*

Literature Review

Patching Editing

Modifying **Alignment**

Debugging

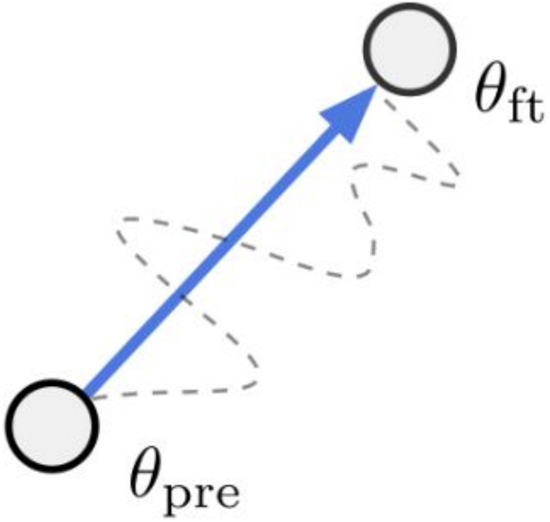
Steering

Common Methods - Fine-tuning[1,2], Sparse Parameter Tuning[3,4], Reinforcement Learning through Human Feedback (RLHF)[5,6]

Limitations - Efficiency, Catastrophic Forgetting, Hard to Add/Remove Tasks

**References in last slide*

Main Idea: Play with Arithmetic Operations in Weight Space



$$\tau = \underline{\theta_{ft} - \theta_{pre}}$$

Element-Wise Difference

τ → **"Task Vector"**

θ_{pre} → **Pre-trained Weights**

θ_{ft} → **Fine-tuned Weights**

Main Idea: Play with Arithmetic Operations in Weight Space

$$\theta_{new} = \underbrace{\theta + \lambda\tau}_{\text{Element-Wise Addition}}$$

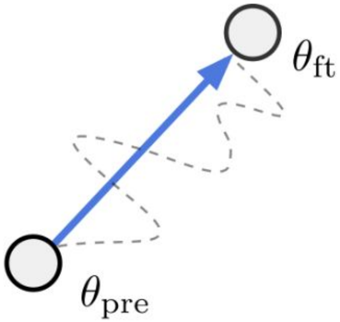
λ → **Scalar Hyperparameter**

θ → **Weights (Same Architecture)**

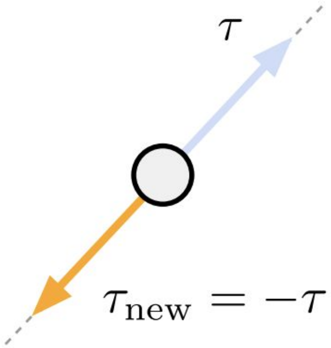
θ_{new} → **Task Added Weights**

Application: Forgetting via Negation

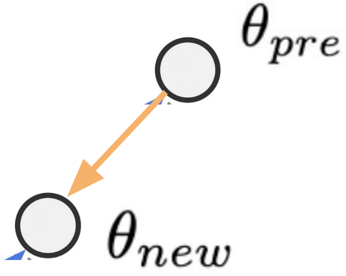
Goal : Unlearning Undesired Biases of Pre-trained Model



1. Find Task Vector



2. Negate Task Vector



3. Add to Pre-trained Weights

Results: Forgetting via Negation

Negative Target Task: Making Language Model Produce Toxic Content

	Negative Target Tasks (↓)		Control Task
Method	% toxic generations (↓)	Avg. toxicity score (↓)	WikiText-103 perplexity (↓)
Pre-trained	$\boxed{4.8}$	$\boxed{0.06}$	$\boxed{16.4}$
Fine-tuned	57	0.56	16.6
Gradient ascent	0.0	0.45	$>10^{10}$
Fine-tuned on non-toxic	1.8	0.03	17.2
Random vector	4.8	0.06	16.4
Negative task vector	$\boxed{0.8}$	$\boxed{0.01}$	$\boxed{16.9}$

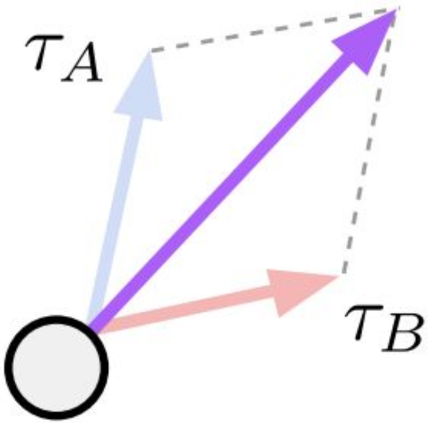
6x Reduction in Toxic Content

~ Similar Perplexity

Application: Learning Via Addition

Goal : Multi-task Learning

$$\tau_{\text{new}} = \tau_A + \tau_B$$



$$\theta_{new} = \theta_{pre} + \lambda \tau_{new}$$

Application: Multi-task Learning

Build a model that can

(a) Classify Digits Images : 0,1, 2 ….. 9
(b) Classify Car Images: Mercedes, Tesla, Toyota

$$\tau_{multi} = \tau_{Digits} + \tau_{Car}$$

$$\theta_{multi} = \theta_{pre} + \lambda\tau_{multi}$$

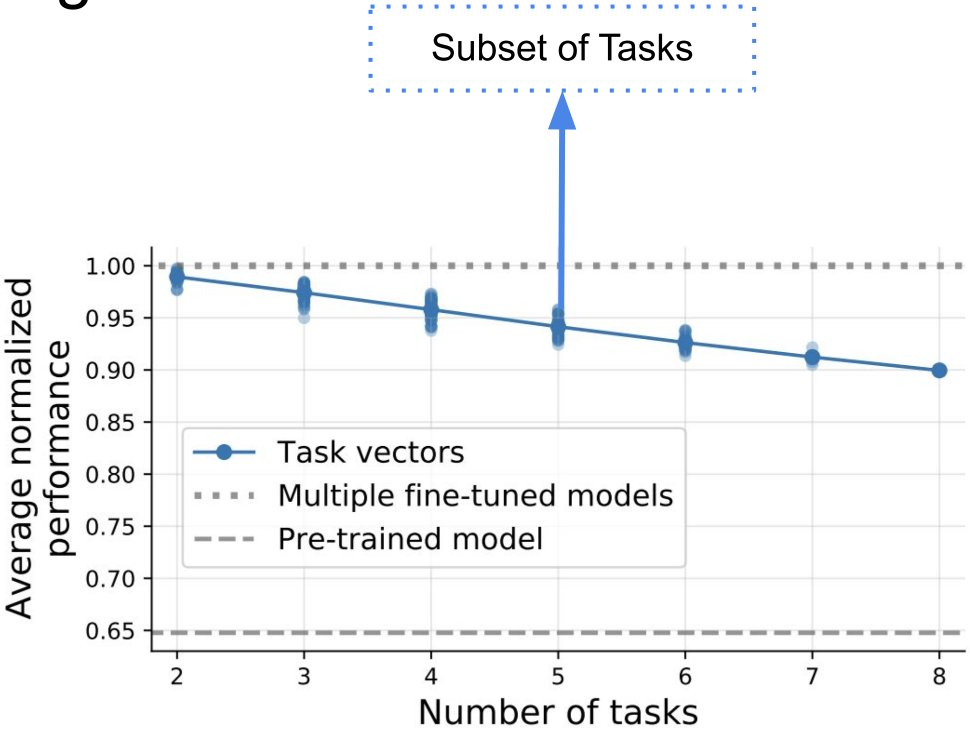
Results: Multi-task Learning

Adding Task Vectors for Multi-task Learning

Average Performance
Of Tasks

Average Normalized Performance

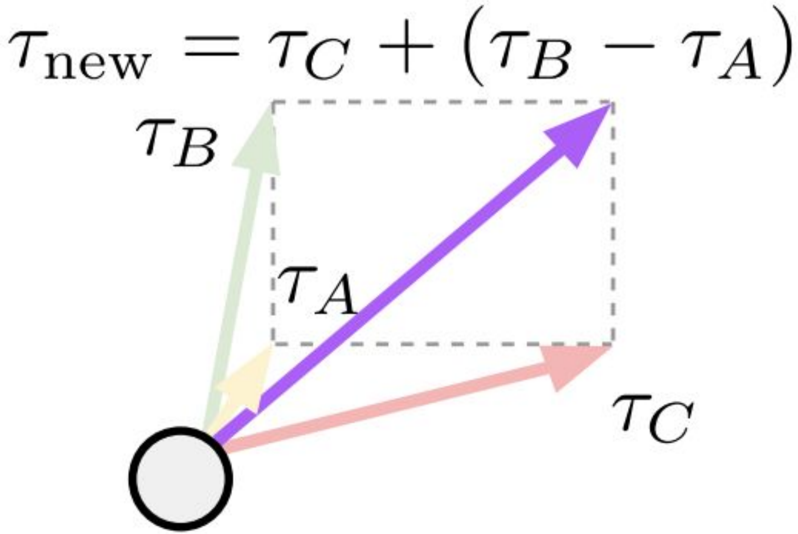
Normalized by corresponding
Fine-tuned Model Single
Task Performance



Different Image Classification Tasks

Application: Task Analogies

A is to B as C is to (D?)



Results: Domain Adaptation

Fine-tuned on Auxiliary Data for Language Modeling

$$\hat{\tau}_{\text{target; sent}} = \boxed{\tau_{\text{target; lm}}} + (\tau_{\text{auxiliary; sent}} - \boxed{\tau_{\text{auxiliary; lm}}})$$

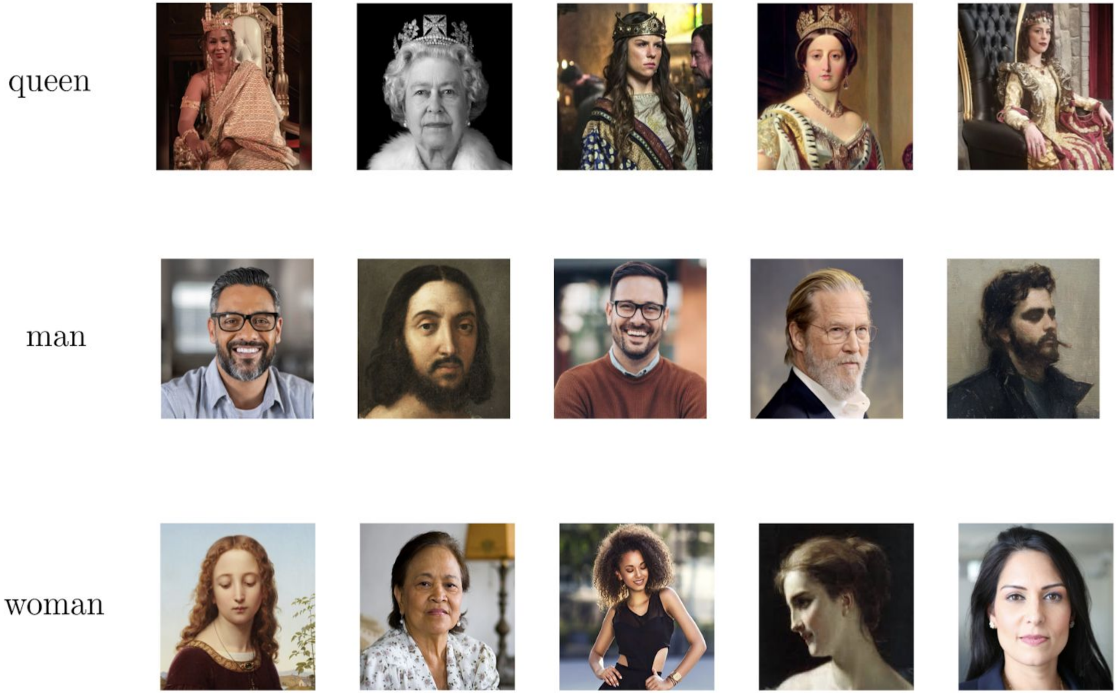
Fine-tuned on Target Dataset for Language Modelling

	target = Yelp			target = Amazon		
Method	T5-small	T5-base	T5-large	T5-small	T5-base	T5-large
Fine-tuned on auxiliary	88.6	92.3	95.0	87.9	90.8	94.8
Task analogies	89.9	93.0	95.1	89.0	92.7	95.2
Fine-tuned on target	91.1	93.4	95.5	90.2	93.2	95.5

Detour: An Observation on the Results

Method	target = Yelp			target = Amazon		
	T5-small	T5-base	T5-large	T5-small	T5-base	T5-large
Fine-tuned on auxiliary	88.6	92.3	95.0	87.9	90.8	94.8
Task analogies	89.9	93.0	95.1	89.0	92.7	95.2
Fine-tuned on target	91.1	93.4	95.5	90.2	93.2	95.5

Initial Gap is Not Large !!!



$$\tau_{\text{queen}} + (\tau_{\text{man}} - \tau_{\text{woman}}) \quad = \quad ?$$

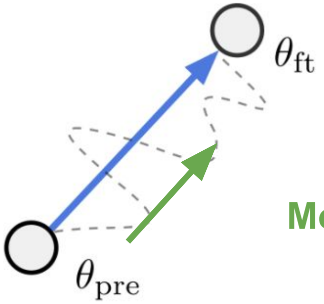
Results: Learning through Analogies

$$\tau_{\text{queen}} + (\tau_{\text{man}} - \tau_{\text{woman}}) = \hat{\tau}_{\text{king}}$$

Intuition

Hypothesis is based related empirical work of interpolation of weights -

a) Results of Ensembling Weights ~ Results of Ensembling Predictions[1]



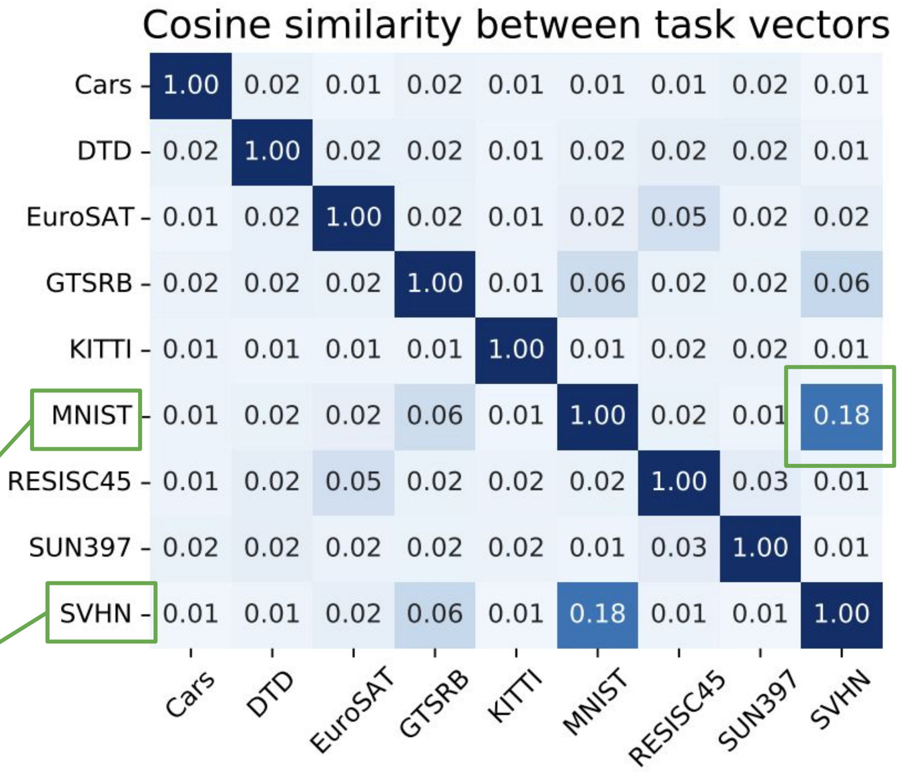
Movement in this direction => Performance Gain

b) Performance improves linearly when fine-tuning

[1]*Robust fine-tuning of zero-shot models (CVPR 2022)*

Why Task Arithmetic works well?

- Task Vectors are close to Orthogonal
- Combining Multiple Task Vectors ~ Minimal Interference



Similar Tasks ~ Higher Similarity

Strengths

- Efficient : Only Element Wise Operations b/w matrices
- Modular : Add/Remove Abilities to Models
- Retain Control Task Performance
- Strong Empirical Results

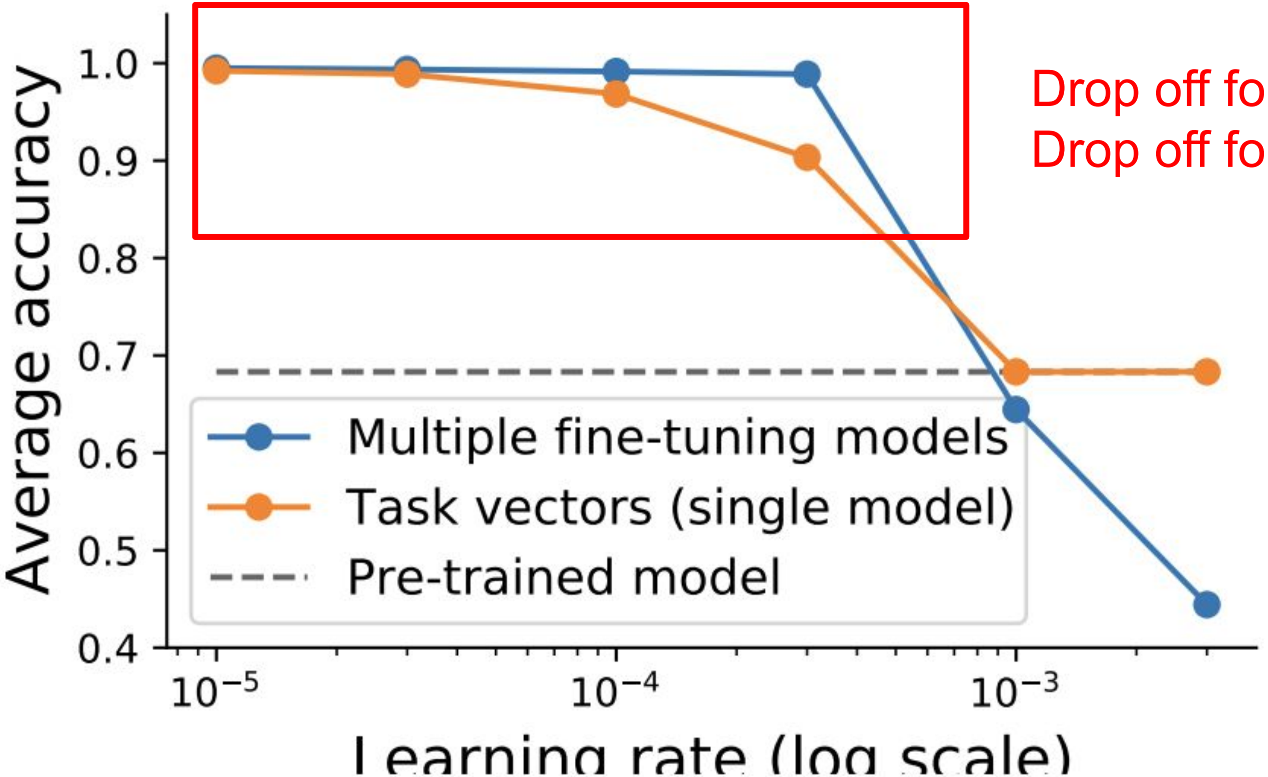
Limitations

- Architecture Restrictions
- Sensitive to High Learning Rate
- Negative Interference in Multi-task Learning

Limitations: Architecture Restricted

- Element-Wise Operation: Restricted to same Architecture
- All experiments on same pre-trained initialization
- Only works for the fine-tuning regime

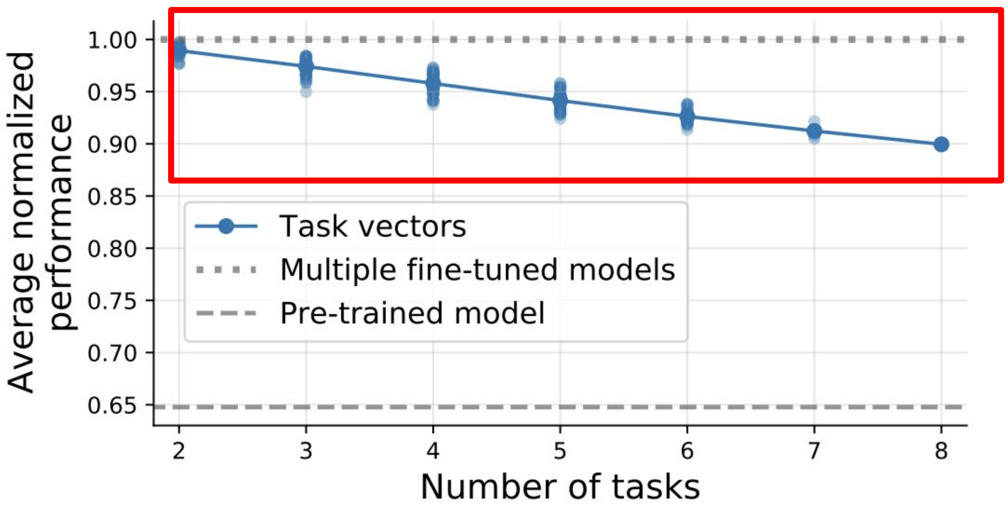
Limitation: Sensitive to High Learning Rate



Drop off for Task Vectors >
Drop off for Fine-tuning

Limitation: Multi-task Learning

Still Room For Improvement!



Different Image Classification Tasks

Future Directions

- Expanding this framework:
  - Architecture-Invariant
  - Multi-Modal Architectures
- Exploring the Weight Space of Models in Depth[2]

[1] *Knowledge is a Region in Weight Space for Fine-tuned Language Models (ICLR 2023)*

Questions?

Thanks!

Literature Review References

[1] Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification." arXiv preprint arXiv:1801.06146 (2018).

[2] Fu, Zihao, et al. "On the Effectiveness of Parameter-Efficient Fine-Tuning." arXiv preprint arXiv:2211.15583 (2022).

[3] Karan Goel, Albert Gu, Yixuan Li, and Christopher Re. Model patching: Closing the sub- group performance gap with data augmentation, 2020. https://arxiv.org/abs/2008. 06775.

[4] Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. In Advances in Neural Information Processing Systems (NeurIPS), 2021. https: //arxiv.org/abs/2111.09839.

[5] Christiano, Paul F., et al. "Deep reinforcement learning from human preferences." Advances in neural information processing systems 30 (2017).

[6] Marco Tulio Ribeiro and Scott Lundberg. Adaptive testing and debugging of nlp models. In Annual Meeting of the Association for Computational Linguistics (ACL), 2022. https: //aclanthology.org/2022.acl-long.230/.

Image References

1) *University of Washington Image*
2) *Microsoft Research*
3) *AI-2 Image*
4) *Insider ChatGPT Image*
5) *TechCruch GPT3 Image*

Rest of the Images are from the paper