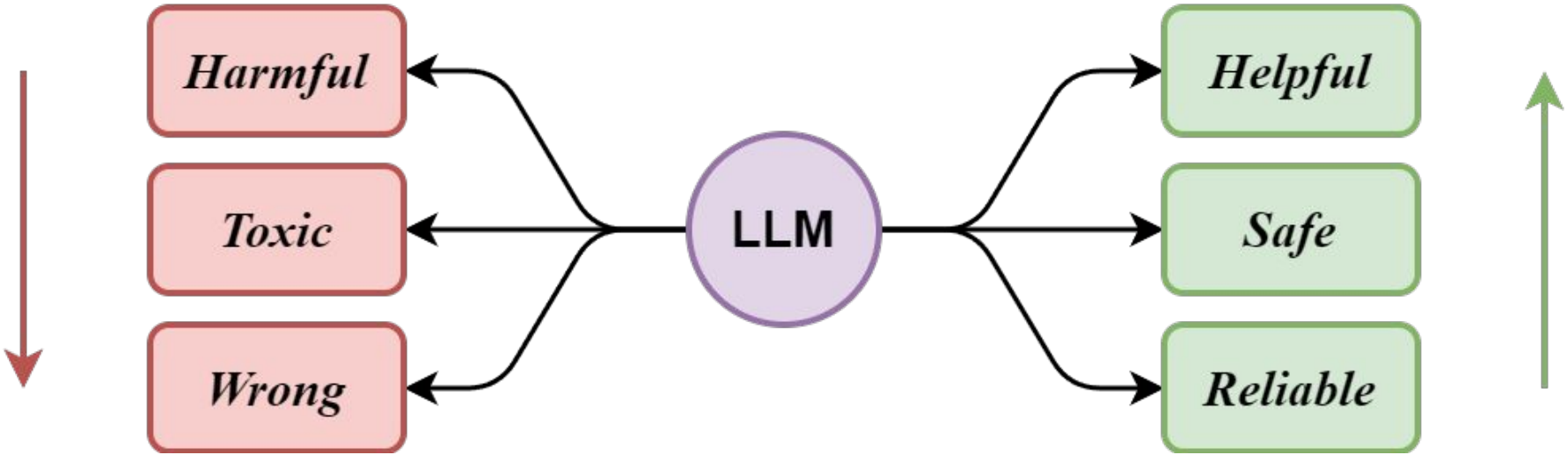
Would I Lie To You?
Inference Time Alignment of Language Models using Direct Preference Heads

Avelina Hadji-Kyriacou, Ognjen Arandjelovic

University of
St Andrews

What is alignment?



Alignment Increase Hallucinations

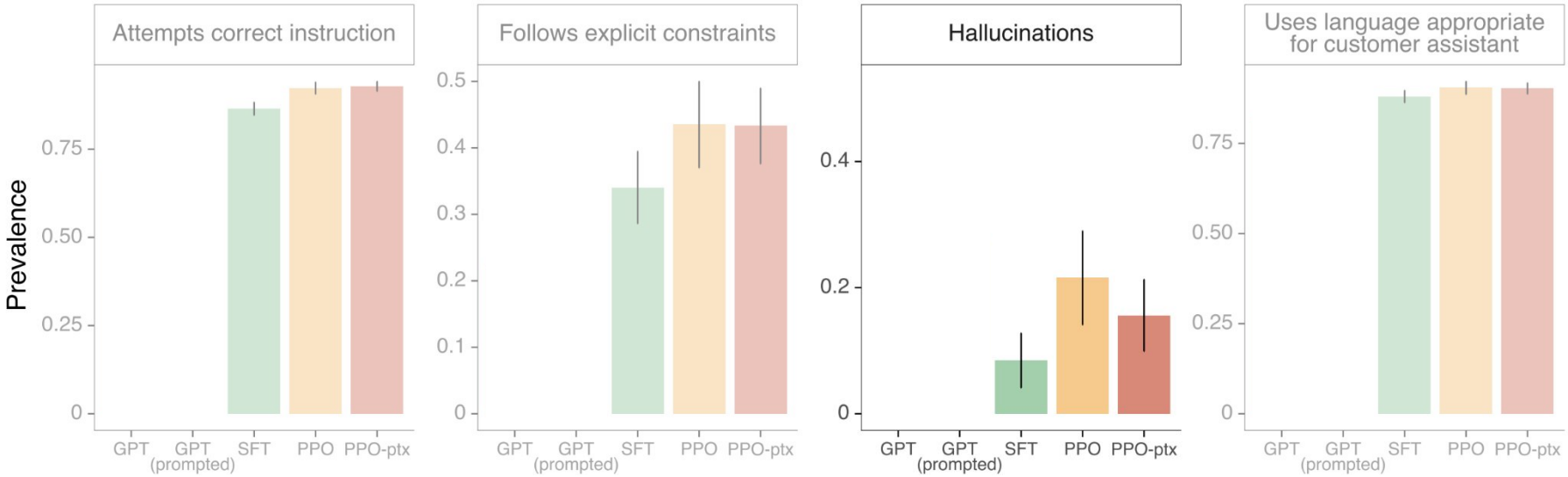


Figure 4 - Training language models to follow instructions with human feedback, Ouyang et al, 2022

Alignment Hurts Small Language Models

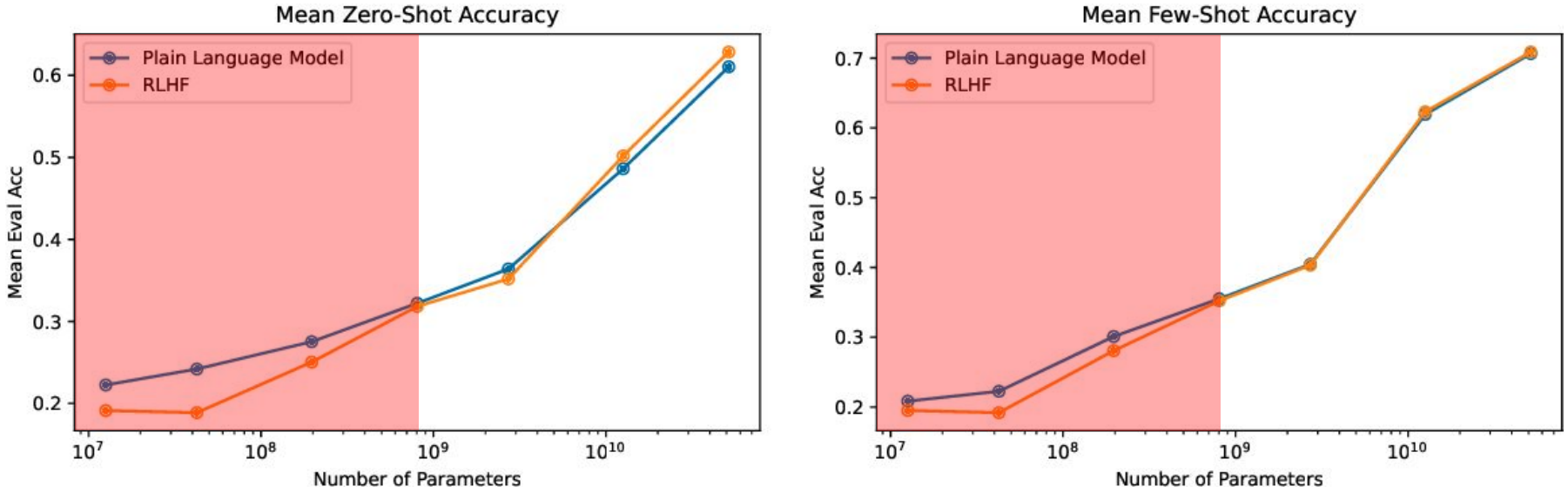
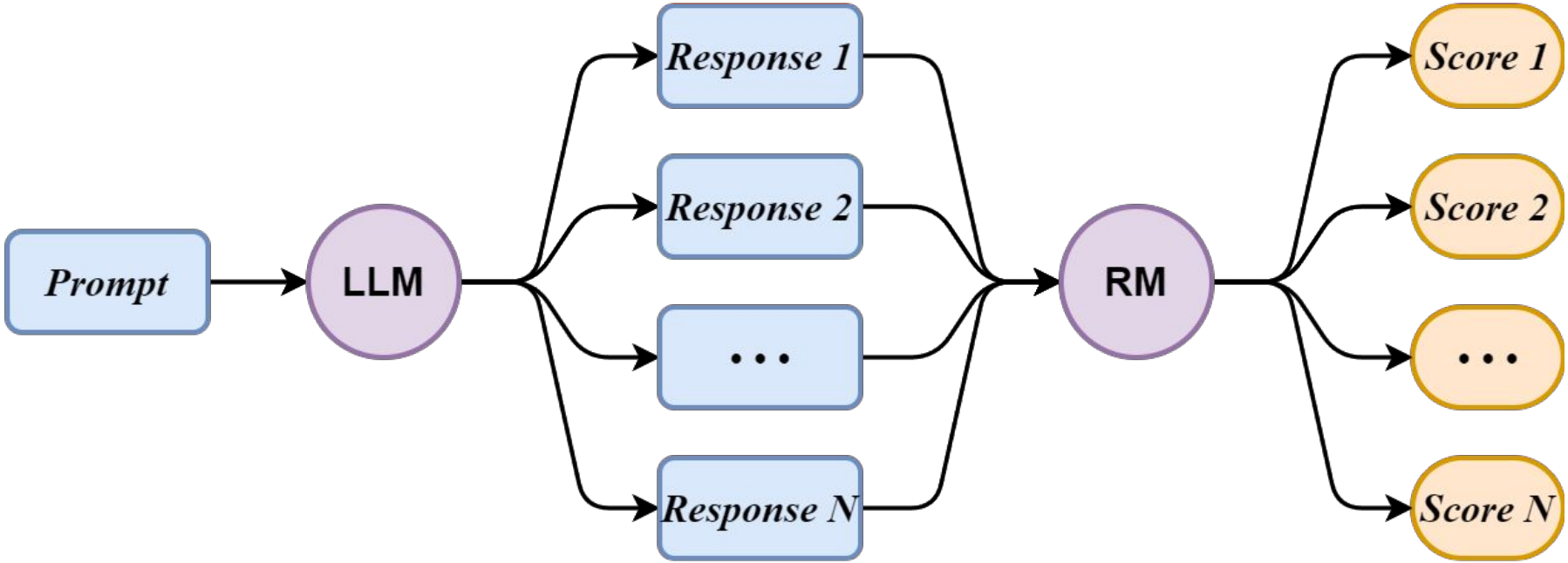


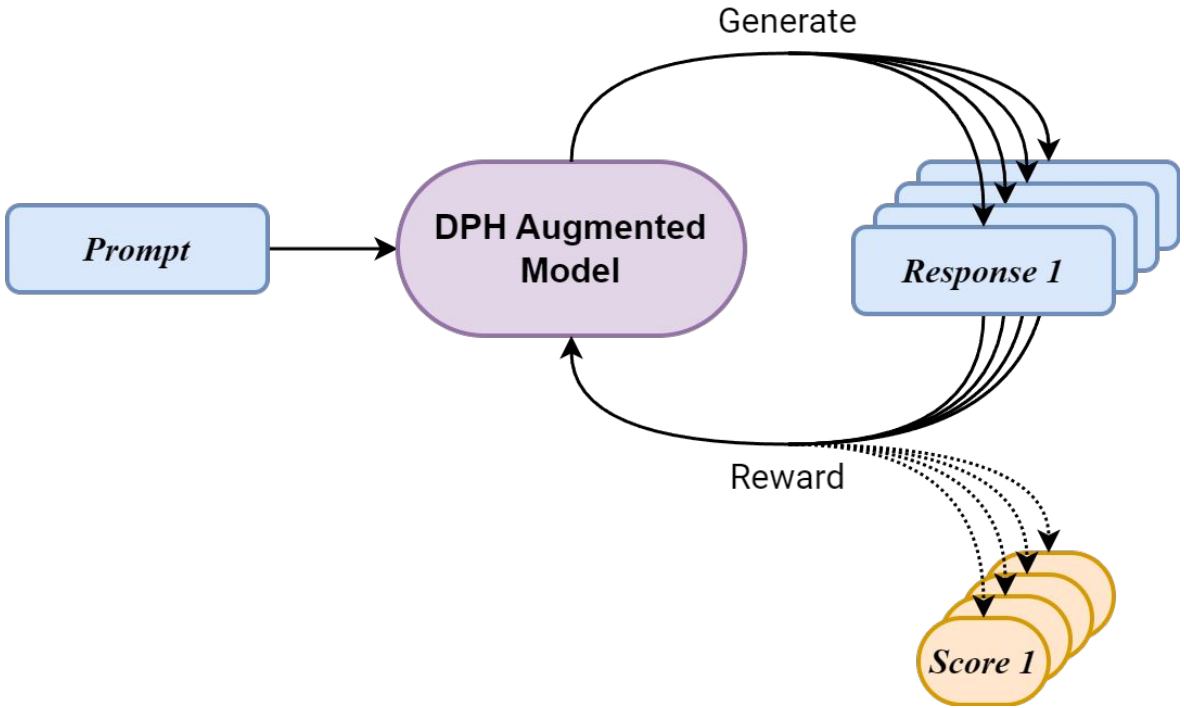
Figure 3 - Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, Bai et al, 2022

What is the alternative?

Inference Time Alignment



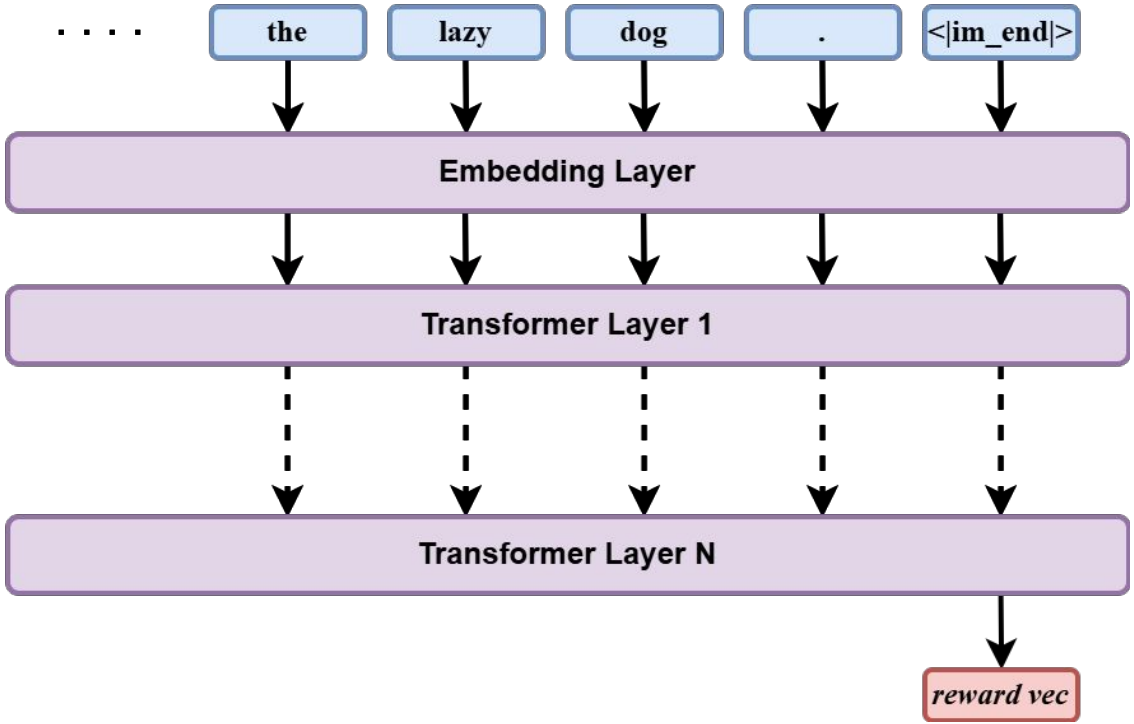
Inference Time Alignment with Direct Preference Heads



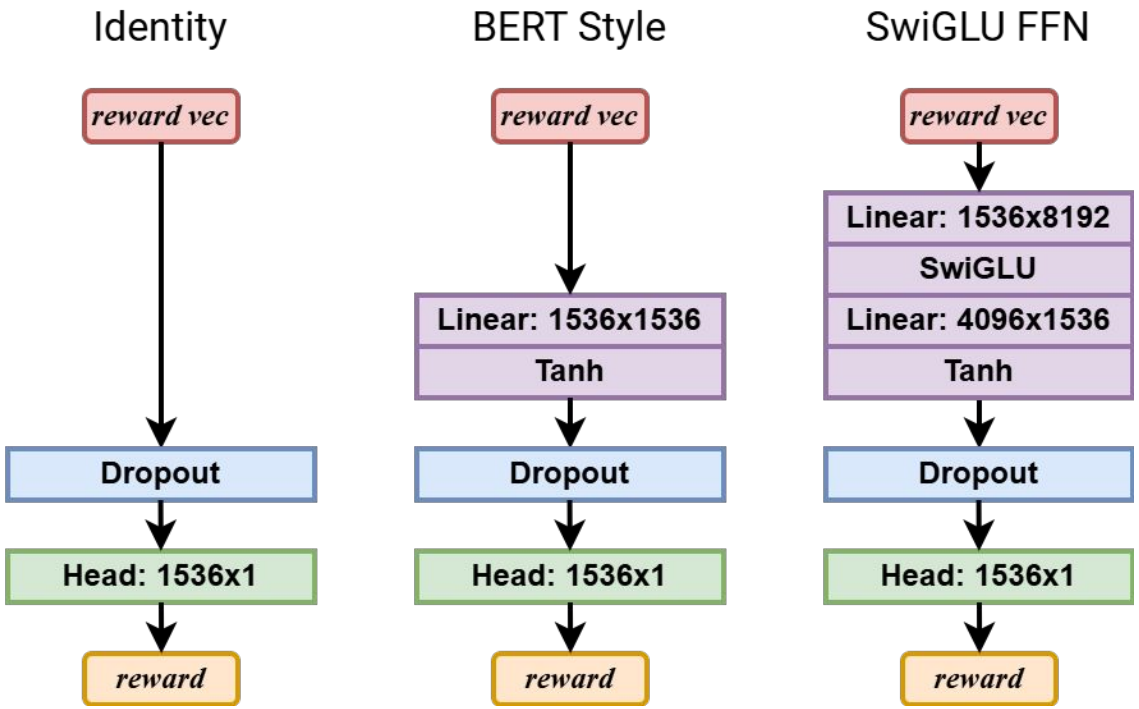
Direct Preference Heads

- Aggregation Function
  - **selects hidden representations** from the language model
  - combines them to **produce a single vector** (one per seq) or **sequence of vectors** (one per tok)
  - may be **learnable** or **parameter free**
- Pooling Function
  - transforms the aggregated representations to produce a **final representation vector** of size **d**
  - may be **learnable** or **parameter free**
- Reward Head
  - **learnable** d×1 projection
- Loss Function
  - **pairwise loss**, similar to DPO or ORPO

DPH Aggregation Function



DPH Pooling Function



DPH Loss Functions

$$\mathcal{L}_{\text{SepDPH}}(r_w, r_l) = -\underbrace{[(1-\epsilon)\log\sigma(r_w) + \epsilon\log\sigma(-r_w)]}_{\text{Positive BCE with label smoothing}} - \underbrace{[\epsilon\log\sigma(r_l) + (1-\epsilon)\log\sigma(-r_l)]}_{\text{Negative BCE with label smoothing}}$$

$$\mathcal{L}_{\text{ConDPH}}(r_w, r_l) = -\overbrace{(1-\epsilon)\log\sigma(r_w - r_l)}^{\text{Reward margin}} - \overbrace{\epsilon\log\sigma(r_l - r_w)}^{\text{Label smoothing}}$$

Evaluation

- **GPT4All** – Primary
  - **commonsense reasoning**
  - representative of model alignment across
- **GLUE** – Secondary
  - **natural language understanding**
  - tests for alignment degradation in classification tasks
- **RACE** – Secondary
  - **reading comprehension**
  - tests for alignment degradation in multiple-choice QA tasks

Results - Commonsense Reasoning

System	Tokens	Params	HellaSwag	OpenBookQA	WinoGrande	ARC-Challenge	ARC-Easy	BoolQ	PIQA	Average
$\text{Ours}_{\text{Vocab}}$	100B	551M	36.93	28.60	51.14	26.19	25.67	61.25	65.39	42.17
Ours_{SFT}	100B	551M	42.59	45.20	55.01	35.84	47.01	76.24	69.37	53.04
Ours_{DPO}	100B	551M	44.83	52.40	57.38	39.76	53.54	**79.08**	72.36	57.05
Ours_{DPH}	100B	+19M	**59.36**	**57.40**	**59.12**	**41.21**	**56.82**	78.81	68.77	**60.21**
Pythia-1.0B	300B	1.1B	47.16	31.40	53.43	27.05	48.99	60.83	69.21	48.30
Pythia-1.4B	300B	1.5B	52.01	33.20	57.38	28.50	54.00	63.27	70.95	51.33
TinyLlama	3T	1.1B	59.20	36.00	**59.12**	30.12	55.25	57.83	**73.29**	52.99

Results - Reading Comprehension

System	Tokens	Params	RACE-middle	RACE-high	Weighted Average
$\text{Ours}_{\text{Vocab}}$	100B	551M	26.0	24.6	25.0
Ours_{SFT}	100B	551M	56.1	52.9	53.8
Ours_{DPO}	100B	551M	65.9	59.8	61.6
Ours_{DPH}	100B	+19M	**66.9**	**60.6**	**62.5**
GPT-1	32B	117M	62.9	57.4	59.0
LLaMA 7B	1T	6.7B	61.1	46.9	51.0
LLaMA 13B	1T	13B	61.6	47.2	51.4

Results - Natural Language Understanding

System	Tokens	Params	MNLI m/mm	QQP F1/Acc	QNLI Acc	SST-2 Acc	CoLA M Corr	STS-B P/S Corr	MRPC F1/Acc	RTE Acc	Score w/o WNLI	WNLI Acc	Score w/ WNLI
$\text{Ours}_{\text{Vocab}}$	100B	551M	34.1/34.7	28.2/42.9	50.2	58.0	0.9	-0.9/99.2	69.4/57.4	50.9	42.8	34.9	41.9
Ours_{SFT}	100B	551M	73.6/75.0	59.1/82.8	81.4	90.8	22.7	80.6/92.4	80.6/75.2	71.4	72.0	38.4	68.2
Ours_{DPO}	100B	551M	78.8/80.2	65.6/85.6	87.0	93.3	36.5	83.7/94.4	83.9/79.1	73.9	77.0	37.7	72.7
Ours_{DPH}	100B	+19M	80.0/80.6	65.8/85.3	87.5	94.0	43.8	**85.3/93.0**	85.5/80.2	**75.3**	78.6	46.6	75.0
GPT-1	32B	117M	82.1/81.4	70.3/ -	87.4	91.3	45.4	82.0/80.0	82.3/ -	56.0	-	-	72.8
$\text{BERT}_{\text{Base}}$	128B	110M	84.6/83.4	71.2/ -	90.5	93.5	52.1	- /85.8	88.9/ -	66.4	-	-	78.3
$\text{BERT}_{\text{Large}}$	128B	340M	**86.7/85.9**	**72.1/89.3**	**92.7**	**94.9**	**60.5**	87.6/86.5	**89.3/85.4**	70.1	**82.5**	**65.1**	**80.5**

Results - Pooling Function and Loss Objective Ablations

Objective	Pooling Function	Add. Params	GLUE	GPT4All	RACE	HellaSwag	WinoGrande	PIQA
Separable	Identity	1536	75.06	56.86	56.54	46.63	53.20	65.29
Separable	BERT Style	2.4M	75.13	55.86	56.62	45.84	52.17	64.69
Separable	SwiGLU FFN	19M	**75.19**	57.14	**57.60**	48.72	53.35	64.96
Contrastive	Identity	1536	74.99	57.66	54.09	50.93	53.83	66.87
Contrastive	BERT Style	2.4M	73.91	57.07	55.89	49.98	54.62	67.30
Contrastive	SwiGLU FFN	19M	74.04	**58.28**	55.95	**51.38**	**55.80**	**67.57**

Where to next?

That's a wrap

Current Repo - github.com/Avelina9X/direct-preference-heads

Hugging Face - hf.co/collections/Avelina/direct-preference-heads-preprint-6612d8a6fa3843352943fd43

Nightly Repo - github.com/Avelina9X/memory-transformer-pt4/tree/new_pooler

Contact me - lhk3@st-andrews.ac.uk | avelina@avelina.io

NEURAL INFORMATION PROCESSING SYSTEMS

University of St Andrews