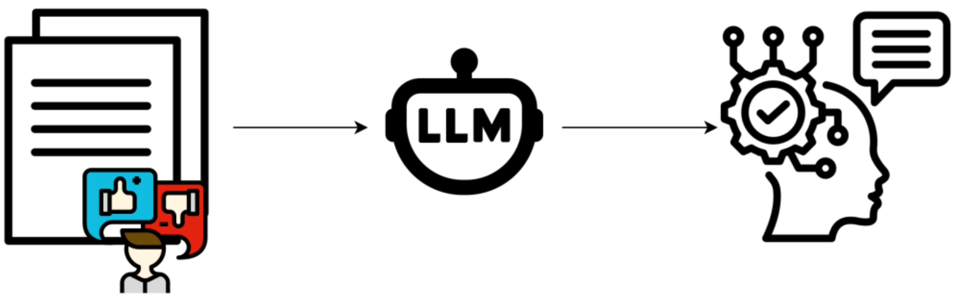
University of Stuttgart
Institute of Industrial Automation
and Software Engineering



Training LLMs on domain-specific knowledge with reinforcement learning based on preference data

Master Thesis Final Report

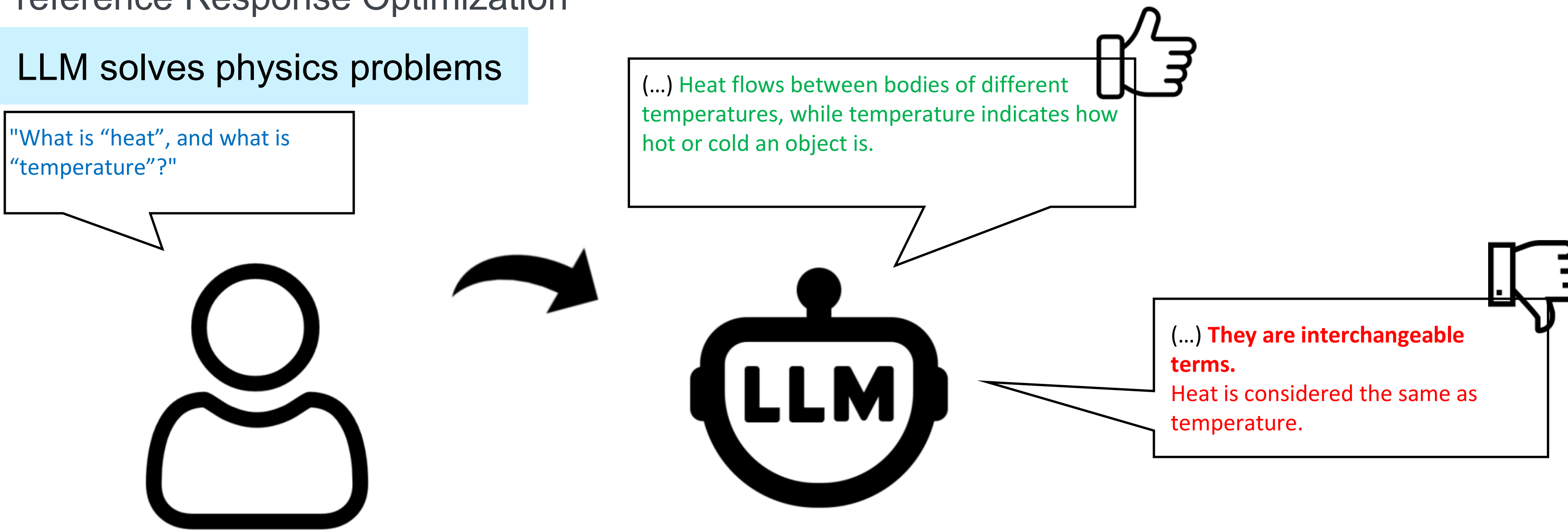
Presenter: Lun-Yu Yuan
Supervisor: Yuchen Xia
Examiner: Prof. Dr. Ing. Michael Weyrich

Contents

Introduction

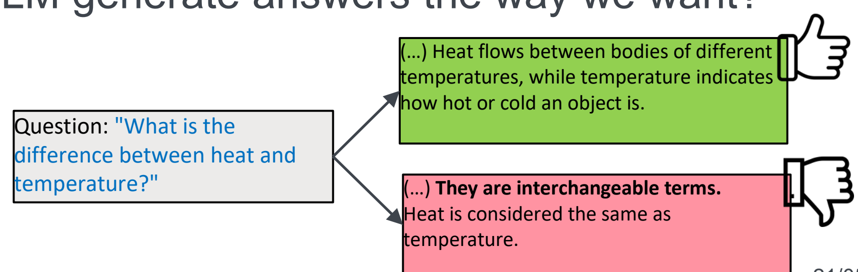
Preference Response Optimization

LLM solves physics problems



Question: How can we make an LLM generate answers the way we want?

Preference Learning



Introduction

Training Purpose

Method	Change the **Text Format**	Change the **Reasoning**

To find the **magnitude of the magnetic force** on a current-carrying wire placed in a magnetic field, we use the formula:

$$F = ILB\sin(\theta)$$

Where:

- F = magnetic force
- I = current = 5 A
- L = length of the wire = 2 m
- B = magnetic field strength = 0.3 T
- θ = angle between wire and magnetic field = 30°

Now, plug in the values:

$$F = 5 \times 2 \times 0.3 \times \sin(30^\circ)$$
$$\sin(30^\circ) = 0.5$$
$$F = 5 \times 2 \times 0.3 \times 0.5 = 1.5\,\mathrm{N}$$

✅ **Final Answer:**

The magnetic force on the wire is 1.5 newtons (N).

good **Text Format**

A wire of length 2 m carries a current of 5 A and is placed in a uniform magnetic field of 0.3 T. The wire makes an angle of 30° with the magnetic field. What is the magnitude of the magnetic force on the wire?

To find the magnitude of the magnetic force on a current-carrying wire placed in a magnetic field, we use the formula:

F = ILB sin(θ)

Where:
= magnetic force
= current = 5 A
L = length of the wire = 2 m
B = magnetic field strength = 0.3 T
θ = angle between wire and magnetic field = 30°

Now, plug in the values:

F = 5 × 2 × 0.3 × sin(30°)
sin(30°) = 0.5
F = 5 × 2 × 0.3 × 0.5 = 1.5 N

Final Answer:
The magnetic force on the wire is 1.5 newtons (N).

bad **Text Format**

To find the **magnitude of the magnetic force** on a current-carrying wire placed in a magnetic field, we use the formula:

$$\boxed{F = ILB\sin(\theta)}$$

Where:

- F = magnetic force
- I = current = 5 A
- L = length of the wire = 2 m
- B = magnetic field strength = 0.3 T
- θ = angle between wire and magnetic field = 30°

Now, plug in the values:

$$F = 5 \times 2 \times 0.3 \times \sin(30^\circ)$$
$$\sin(30^\circ) = 0.5$$
$$F = 5 \times 2 \times 0.3 \times 0.5 = 1.5\,\mathrm{N}$$

✅ **Final Answer:**

The magnetic force on the wire is 1.5 newtons (N).

correct **Reasoning**

To find the **magnitude of the magnetic force** on a current-carrying wire placed in a magnetic field, we use the formula:

$$\boxed{F = ILB\cos(\theta)}$$

Where:

- F = magnetic force
- I = current = 5 A
- L = length of the wire = 2 m
- B = magnetic field strength = 0.3 T
- θ = angle between wire and magnetic field = 30°

Now, plug in the values:

$$F = 5 \times 2 \times 0.3 \times \cos(30^\circ)$$
$$\cos(30^\circ) = 0.86$$
$$F = 5 \times 2 \times 0.3 \times 0.86 = 2.58\,\mathrm{N}$$

✅ **Final Answer:**

The magnetic force on the wire is 2.58 newtons (N).

incorrect **Reasoning**

Master thesis result from Mr. Lin MT 3806 (presented 30.04.2025)

LLMs	Dataset	Improved performance					
		trainset	testset	variants	Task	Task+prompt with knowledge	Task (MMLU)
GPT4o-mini	PDF_synthetic	5.86-->7.29 +14.3%	5.25-->5.96 +7.1%	5.88-->7.16 +12.8%	Success Rate: 58/87-->50/87	Success Rate:65/87	56.0-->51.8 -4.2%
	Knowledge_ distillation	5.86-->7.18 +13.2%	5.25-->7.29 +20.4%	5.88-->7.31 +14.3%	Success Rate: 58/87-->57/87	Success Rate: 70/87	56.0-->55.5 -0.5%

✓ Change of QA and the Text Format

✗ Problem solving the Reasoning

- Conclusion: training with supervised finetuning is effective (Text Format)

Overall Objective

Improve LLM generation ability

Introduction

Training Purpose

Method	Change the **Text Format**	Change the **Reasoning**
(MT3806) Supervised Fine-Tuning (SFT)	✓	✗
Preference Learning	?	?

To find the **magnitude of the magnetic force** on a current-carrying wire placed in a magnetic field, we use the formula:

$$F = ILB\sin(\theta)$$

Where:

- F = magnetic force
- I = current = 5 A
- L = length of the wire = 2 m
- B = magnetic field strength = 0.3 T
- θ = angle between wire and magnetic field = 30°

Now, plug in the values:

$$F = 5 \times 2 \times 0.3 \times \sin(30^\circ)$$
$$\sin(30^\circ) = 0.5$$
$$F = 5 \times 2 \times 0.3 \times 0.5 = 1.5\,\mathrm{N}$$

✅ **Final Answer:**

The magnetic force on the wire is 1.5 newtons (N).

good **Text Format**

A wire of length 2 m carries a current of 5 A and is placed in a uniform magnetic field of 0.3 T. The wire makes an angle of 30° with the magnetic field. What is the magnitude of the magnetic force on the wire?

To find the magnitude of the magnetic force on a current-carrying wire placed in a magnetic field, we use the formula:

F = ILB sin(θ)

Where:
F = magnetic force
I = current = 5 A
L = length of the wire = 2 m
B = magnetic field strength = 0.3 T
θ = angle between wire and magnetic field = 30°

Now, plug in the values:

F = 5 × 2 × 0.3 × sin(30°)
sin(30°) = 0.5
F = 5 × 2 × 0.3 × 0.5 = 1.5 N

Final Answer:
The magnetic force on the wire is 1.5 newtons (N).

bad **Text Format**

To find the **magnitude of the magnetic force** on a current-carrying wire placed in a magnetic field, we use the formula:

$$\boxed{F = ILB\sin(\theta)}$$

Where:

- F = magnetic force
- I = current = 5 A
- L = length of the wire = 2 m
- B = magnetic field strength = 0.3 T
- θ = angle between wire and magnetic field = 30°

Now, plug in the values:

$$F = 5 \times 2 \times 0.3 \times \sin(30^\circ)$$
$$\sin(30^\circ) = 0.5$$
$$F = 5 \times 2 \times 0.3 \times 0.5 = 1.5\,\mathrm{N}$$

✅ **Final Answer:**

The magnetic force on the wire is 1.5 newtons (N).

correct **Reasoning**

To find the **magnitude of the magnetic force** on a current-carrying wire placed in a magnetic field, we use the formula:

$$\boxed{F = ILB\cos(\theta)}$$

Where:

- F = magnetic force
- I = current = 5 A
- L = length of the wire = 2 m
- B = magnetic field strength = 0.3 T
- θ = angle between wire and magnetic field = 30°

Now, plug in the values:

$$F = 5 \times 2 \times 0.3 \times \cos(30^\circ)$$
$$\cos(30^\circ) = 0.86$$
$$F = 5 \times 2 \times 0.3 \times 0.86 = 2.58\,\mathrm{N}$$

✅ **Final Answer:**

The magnetic force on the wire is 2.58 newtons (N).

incorrect **Reasoning**

Basics

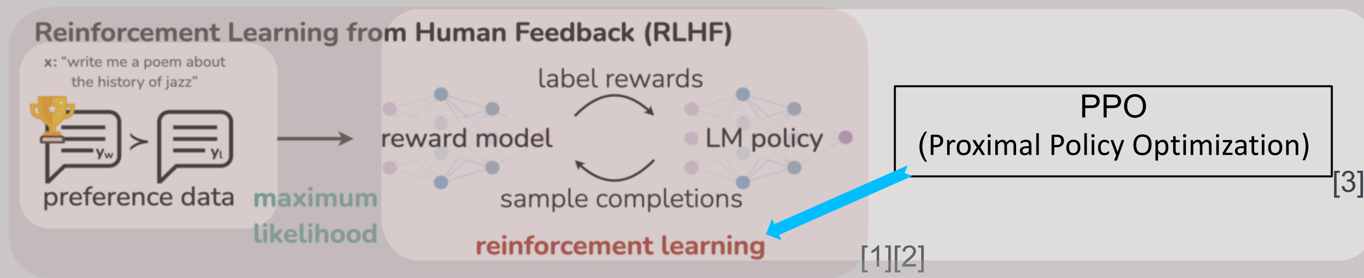
Basic Methods

- 2017: <u>R</u>einforcement <u>L</u>earning from <u>H</u>uman <u>F</u>eedback (RLHF)
- 2023: <u>D</u>irect <u>P</u>reference <u>O</u>ptimization (DPO)

Deep reinforcement learning from human preferences
PF Christiano, J Leike, T Brown… - Advances in neural …, 2017 - proceedings.neurips.cc
… For sophisticated **reinforcement learning** (RL) systems to interact usefully with real-world …
In this work, we explore goals defined in terms of (non-expert) **human preferences** between …
Save Cite Cited by 4018 Related articles All 14 versions

Basics

Reinforcement Learning from Human Feedback (RLHF) – 2017



[3]

[1][2]

- **Preference Data Collection:** Human annotators compare alternative responses to generate preference data.
- **Reward Model**: Learns from preference data to assign a 'reward' score to any given response.
- **Policy Fine-Tuning**: The LM policy is adjusted via reinforcement learning (e.g., PPO) using the reward model's scores.

Effectively Learned Preference

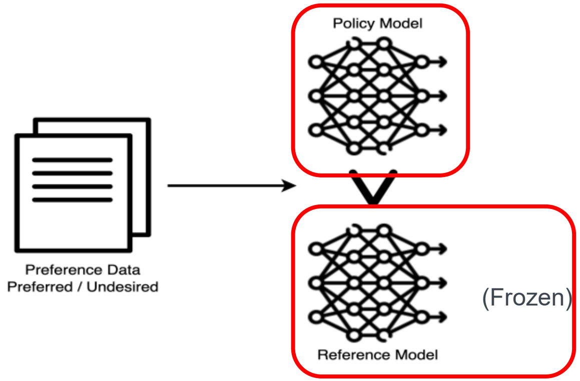
Disadvantages
- Complex pipeline
- Compute-intensive

Direct Preference Optimization (DPO)
[1]

Direct preference optimization: Your language model is secretly a reward model
R Rafailov, A Sharma, E Mitchell… - Advances in …, 2023 - proceedings.neurips.cc
… of human **preferences** and then use RL to **optimize** a language … **optimize** a language model to adhere to human **preferences**, … We propose **Direct Preference** Optimization (DPO), an …
Save Cite Cited by 3417 Related articles All 15 versions

Basics

Direct Preference Optimization (DPO) (2023)



- Purpose –
  Maximize the difference between preferred and rejected answers

Expected:

- Preferred answer’s score ↑ (policy model prefers)
- Rejected answer’s score ↓ (policy model rejects)

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta;\pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\boxed{\beta}\left[\boxed{\underbrace{\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\mathrm{ref}}(y_w\mid x)}}_{\text{Preferred answer's score}}} - \underbrace{\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\mathrm{ref}}(y_l\mid x)}}_{\text{Rejected answer's score}}\right]\right)\right]$$ [1]

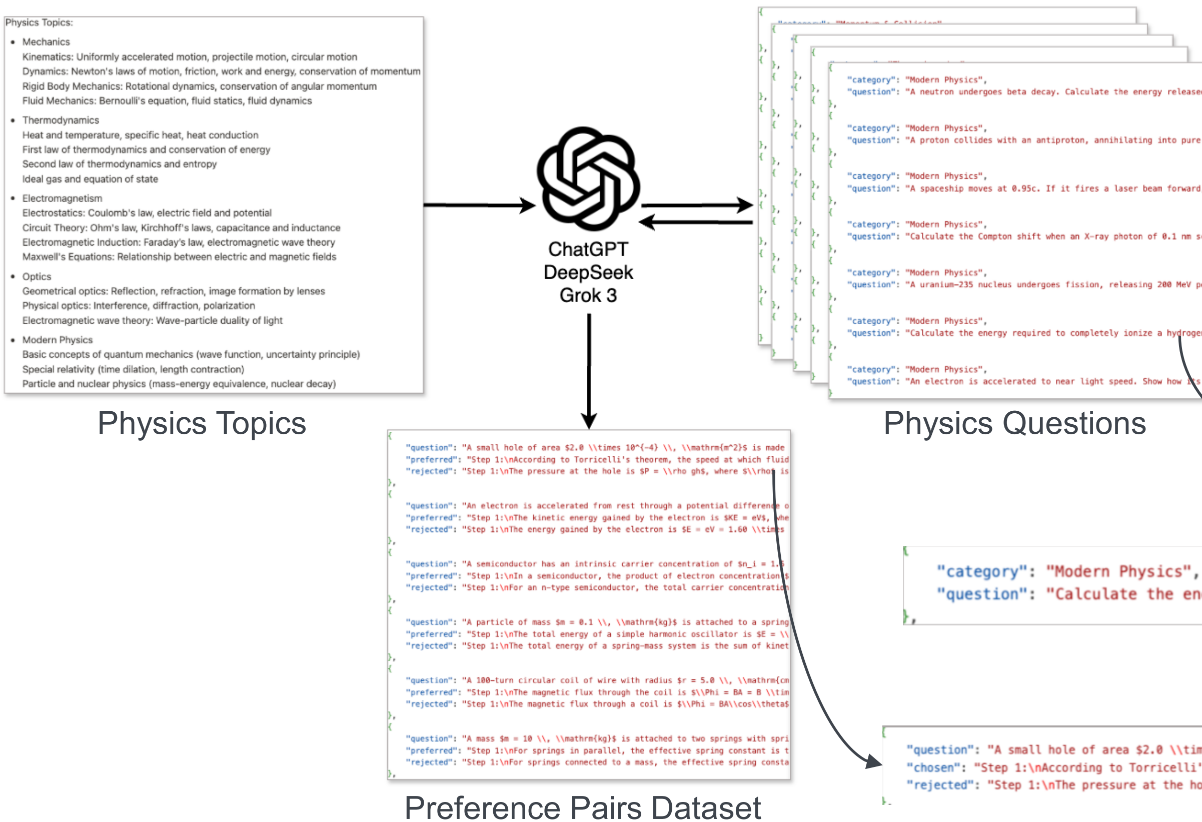
– $\pi_\theta(y|x)$: Probability of completion (y) given prompt (x) under the trainable model (Policy Model)
– $\pi_{ref}(y|x)$: Probability of completion (y) given prompt (x) under the fixed reference model (usually the SFT model)
– β: A temperature or scaling factor controlling how strongly you punish the rejected output
– $log\sigma(\cdot)$: A binary cross-entropy on the log-odds difference
(y_ω: Preferred answer, y_l: Rejected answer)

Training Experiments

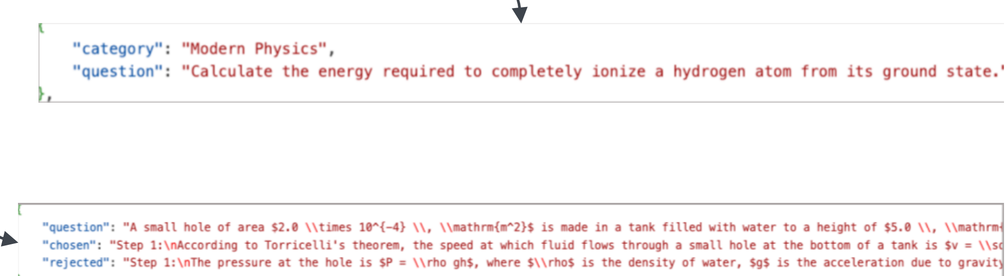
- Synthetic Training Dataset
- Model Selection and Training Resources
- Experiments Execution
  - one failed experiment
  - two extended experiments based on the failure
- Improving Direct Preference Optimization: Extended Approaches

Training Experiments on a Small Language Model – Llama-3.1-8B-Instruct

Synthetic Training Dataset(1/2)



- Collect physics topics
- Generate topic-based questions with LLM
- Produced **question and good/bad answer pairs** with LLM



Training Experiments

Synthetic Training Dataset(2/2)

question:
A 0.3kg bullet at 220m/s embeds in a 4kg block at rest. Final velocity?

preferred:
Step 1: Inelastic => $m_1v_1=(m_1+m_2)V$.
Step 2: Substitute $m_1=0.3$, $v_1=220$, $m_2=4$, $v_2=0$.
Step 3: Compute 0.3 * 220= (4.3)V => 66= 4.3V => V≈15.35m/s.
Answer: 15.35m/s.

rejected:
Step 1: Ignore block => final =220m/s.
Step 2: No momentum share.
Answer:220m/s.


```
{
    "question": "A small hole of area $2.0 \\times 10^{-4} \\, \\mathrm{m^2}$ is made
    "preferred": "Step 1:\nAccording to Torricelli's theorem, the speed at which fluid
    "rejected": "Step 1:\nThe pressure at the hole is $P = \\rho gh$, where $\\rho$ is
},
{
    "question": "An electron is accelerated from rest through a potential difference o
    "preferred": "Step 1:\nThe kinetic energy gained by the electron is $KE = eV$, whe
    "rejected": "Step 1:\nThe energy gained by the electron is $E = eV = 1.60 \\times
},
{
    "question": "A semiconductor has an intrinsic carrier concentration of $n_i = 1.5
    "preferred": "Step 1:\nIn a semiconductor, the product of electron concentration $
    "rejected": "Step 1:\nFor an n-type semiconductor, the total carrier concentration
},
{
    "question": "A particle of mass $m = 0.1 \\, \\mathrm{kg}$ is attached to a spring
    "preferred": "Step 1:\nThe total energy of a simple harmonic oscillator is $E = \\
    "rejected": "Step 1:\nThe total energy of a spring-mass system is the sum of kinet
},
{
    "question": "A 100-turn circular coil of wire with radius $r = 5.0 \\, \\mathrm{cm
    "preferred": "Step 1:\nThe magnetic flux through the coil is $\\Phi = BA = B \\tim
    "rejected": "Step 1:\nThe magnetic flux through a coil is $\\Phi = BA\\cos\\theta$
},
{
    "question": "A mass $m = 10 \\, \\mathrm{kg}$ is attached to two springs with spri
    "preferred": "Step 1:\nFor springs in parallel, the effective spring constant is t
    "rejected": "Step 1:\nFor springs connected to a mass, the effective spring consta
},
```


Preference-Pair Datasets

- Correct vs. Wrong

Contains Total 1004 data

- Training Dataset (804)
- Validation Dataset (100)
- Test Dataset (100)

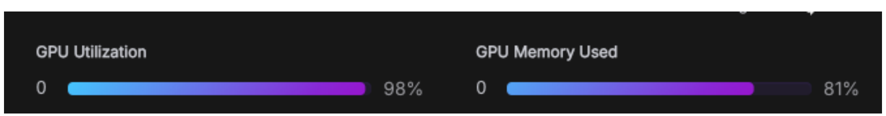
Training Experiments

Model Selection and Training Resource

	Llama-3.1-8B-Instruct
Developer	Meta (July 23, 2024)
Model Parameter Size	8B
Max Context Length	128k
Model Size	~16 GB (bfloat16)

➢ Cloud Computing Platform: **Runpod**

Hardware/Software	specs
Ubuntu	22.04
GPU	H200 SXM
vRAM	141 GB
pytorch version	2.7.0
python version	3.10
CUDA	11.8



- Used ~115GB vRAM during training

```
Collecting torch>=2.0.0
  Downloading torch-2.7.0-cp310-cp310-manylinux_2_28_x86_64.whl (865.2 MB)
     865.2/865.2 MB 6.8 MB/s eta 0:00:00
```

Training Experiments

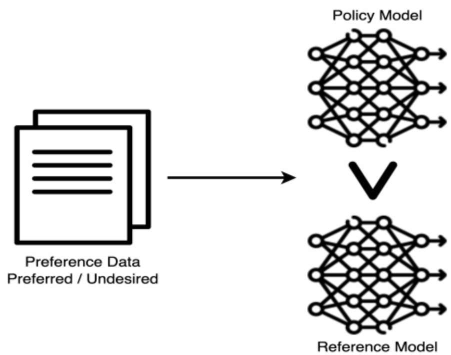
- **One Failed Experiment**

- Two Extended Experiments based on Failure

Direct preference optimization: Your language model is secretly a reward model
R Rafailov, A Sharma, E Mitchell… - Advances in …, 2023 - proceedings.neurips.cc
… of human preferences and then use RL to optimize a language … optimize a language model to adhere to human preferences, … We propose Direct Preference Optimization (DPO), an …
☆ Save 99 Cite Cited by 3417 Related articles All 15 versions

Basics

Direct Preference Optimization (DPO) (2023)



(Frozen)

- Purpose –
  Maximize the difference between preferred and rejected answers

Expected:

- Preferred answer's score ↑ (policy model prefers)
- Rejected answer's score ↓ (policy model rejects)

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\left[\underbrace{\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}}_{\text{Preferred answer's score}} - \underbrace{\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}}_{\text{Rejected answer's score}}\right]\right)\right]$$ [1]

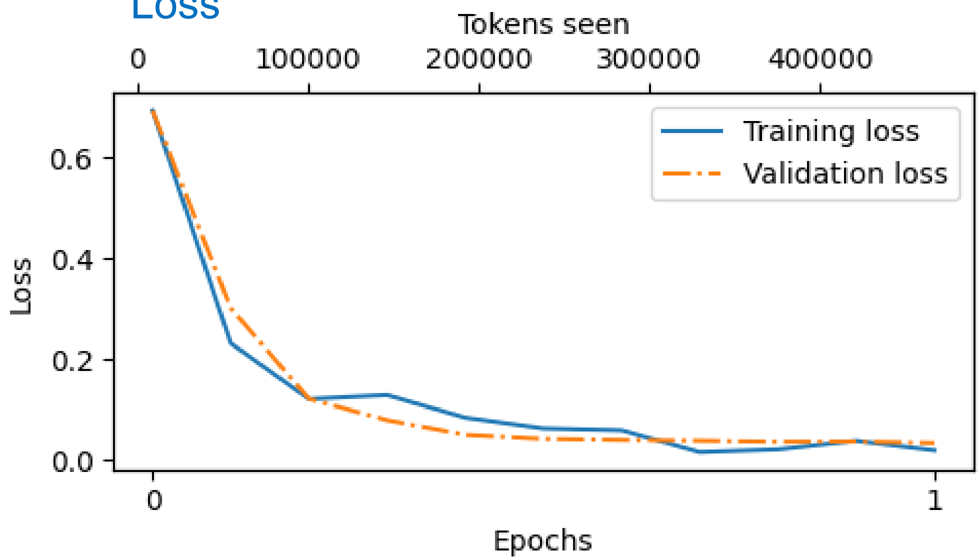
– $\pi_\theta(y|x)$: Probability of completion (y) given prompt (x) under the trainable model (Policy Model)
– $\pi_{ref}(y|x)$: Probability of completion (y) given prompt (x) under the fixed reference model (usually the SFT model)
– β: A temperature or scaling factor controlling how strongly you punish the rejected output
– $log\sigma(\cdot)$: A binary cross-entropy on the log-odds difference
(y_ω: Preferred answer, y_l: Rejected answer)

Training Experiment

Experiments execution

Method	DPO
Batch Size	2
gradient accumulation steps	8
Epoch	1
Learning Rate	3e-6
Beta	0.3

Loss



Model	Text Format	MMLU Benchmark (College Physics)
Original Model	3.0	29.4%
DPO FT Model	3.01	30.4%

Training Experiment

Experiments Execution

question: A wire of length 2m carries a current of 5A and is placed in a uniform magnetic field of 0.3T. The wire makes an angle of 30° with the magnetic field. What is the magnitude of the magnetic force on the wire?

Preference Pairs

preferred:
Step 1: The magnetic force on a current-carrying wire is given by $F = BIL\sin\vartheta$, where I is the current, L is the length, B is the magnetic field strength, and ϑ is the angle between the wire and the magnetic field.
Step 2: $F = 5A * 2m * 0.3T * \sin(30°) = 5 * 2 * 0.3 * 0.5 = 1.5N.$
Answer:1.5N

rejected:
Step 1: The magnetic force on a current-carrying wire is $F = BIL\cos\vartheta$
Step 2: $F = 5A * 2m * 0.3T * \cos(30°) = 5 * 2 * 0.3 *$ 0.866 = 2.6N.
Answer:2.6N

Identical

Original Model Response

Step 1: Recall the formula for the magnetic force on a current-carrying wire in a magnetic field. The magnetic force on a current-carrying wire in a magnetic field is given by the formula $F = BIL\sin\vartheta$, where F is the magnitude of the force, B is the magnitude of the magnetic field, I is the current in the wire, L is the length of the wire, and ϑ is the angle between the wire and the magnetic field.

Step 2: Plug in the given values into the formula. We are given that $B = 0.3T$, $I = 5A$, $L = 2m$, and $\vartheta = 30°$. Plugging these values into the formula, we get $F = 0.3T * 5A * 2m * \sin 30°$.

Step 3: Evaluate the expression. We know that $\sin 30° = 0.5$, so we can simplify the expression to F = 0.3T * 5A * 2m * 0.5.

Step 4: Perform the multiplication. Multiplying the numbers together, we get $F = 3N$.

The final answer is: 3"

DPO Fine-tuned Model Response

Step 1: Recall the formula for the magnetic force on a current-carrying wire in a magnetic field. The magnetic force on a current-carrying wire in a magnetic field is given by the formula $F = BIL\sin\vartheta$, where F is the magnitude of the force, B is the magnitude of the magnetic field, I is the current in the wire, L is the length of the wire, and ϑ is the angle between the wire and the magnetic field.

Step 2: Plug in the given values into the formula. We are given that $B = 0.3T$, $I = 5A$, $L = 2m$, and $\vartheta = 30°$. Plugging these values into the formula, we get $F = 0.3T * 5A * 2m * \sin 30°$.

Step 3: Evaluate the expression. Since $\sin 30° = 0.5$, we have 0.3T * 5A * 2m * 0.5 = 0.3 * 5 * 2 * 0.5 = 3N.

The final answer is: 3"

Reasoning Failed

Training Experiments

- One Failed Experiment

- **Two Extended Experiments based on Failure**

Training Experiments

Improving Direct Preference Optimization: Extended Approaches

Smaug: Fixing failure modes of preference optimisation with **dpo-positive**
A Pal, D Karkhanis, S Dooley, M Roberts… - arXiv preprint arXiv …, 2024 - arxiv.org
… Using these insights, we design **DPO-Positive** (DPOP), a new loss function and training procedure which avoids this failure mode. Surprisingly, we find that DPOP outperforms DPO and …
Save Cite Cited by 123 Related articles All 2 versions

- DPO-Positive (DPOP) (2024)

DPO

$$\mathcal{L}_{\text{DPOP}}(\pi_\theta;\pi_{\text{ref}}) = -E_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\left[\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} - \underbrace{\lambda\cdot\max(0,\log\frac{\pi_{\text{ref}}(y_w|x)}{\pi_\theta(y_w|x)})}_{Penalty\,Term}\right]\right)\right]$$ [4]

Penalty term activates **only when the reference scores the preferred response higher than the policy**, boosting its probability.

DPO-Shift: Shifting the Distribution of Direct Preference Optimization
X Yang, F Jiang, Q Zhang, L Zhao, X Li - arXiv preprint arXiv:2502.07599, 2025 - arxiv.org
… introduce **DPO-Shift** to controllably shift the distribution of the chosen probability. Then, we show that **DPO-Shift** … Furthermore, we demonstrate the superiority of **DPO-Shift** over DPO on …
Save Cite Related articles

- DPO-Shift (2025)

If λ_{shift} = 1 → DPO

$$\mathcal{L}_{\text{DPO-Shift}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\left[\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\text{ref}}(y_w\mid x)} - \underbrace{(\lambda_{\text{shift}})\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\text{ref}}(y_l\mid x)}}_{Shift\ Penalty}\right]\right)\right]$$ [5]

Reducing the impact of the rejected response in the loss calculation by scaling the term representing the rejected response's log-probability ratio (relative to the reference model) with a factor f(λ) less than 1

Training Experiments

Experiments Execution – Validation Loss

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta;\pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\left[\underbrace{\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\mathrm{ref}}(y_w\mid x)}}_{\text{Preferred answer's score}} - \underbrace{\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\mathrm{ref}}(y_l\mid x)}}_{\text{Dispreferred answer's score}}\right]\right)\right]$$

epoch = 1, learning rate = 3e-6

- The model under different methods show decreasing validation loss, indicating learning
- DPO-Shift (green) shows the lowest final validation loss.
- DPOP (orange) stabilizes at a higher loss.

Training Experiments

Reasoning Quality

question: A wire of length 2m carries a current of 5A and is placed in a uniform magnetic field of 0.3T. The wire makes an angle of 30° with the magnetic field. What is the magnitude of the magnetic force on the wire?

Preference Pairs

preferred:

Step 1: The magnetic force on a current-carrying wire is given by *F = BILsinϑ*, where *I* is the current, *L* is the length, *B* is the magnetic field strength, and *ϑ* is the angle between the wire and the magnetic field.

Step 2: *F = 5A * 2m * 0.3T * sin(30°)= 5 * 2 * 0.3 * 0.5 = 1.5N.*

Answer:1.5N

rejected:

Step 1: The magnetic force on a current-carrying wire is *F = BILcosϑ*

Step 2: *F = 5A * 2m * 0.3T * cos(30°) = 5 * 2 * 0.3 ** 0.866 = 2.6N.

Answer:2.6N

DPOP Fine-tuned Model Response (Beta=0.3, λ_{dpop}=50)

Step 1: To find the magnitude of the magnetic force on the wire, we can use the formula *F = BILsinϑ,* where *F* is the magnitude of the force, *B* is the magnitude of the magnetic field, *I* is the current in the wire, *L* is the length of the wire, and *ϑ* is the angle between the wire and the magnetic field.

Step 2: Given that the length of the wire is 2m, the current is 5A, the magnetic field strength is 0.3T, and the angle between the wire and the magnetic field is *30°,* we can substitute these values into the formula.

Step 3: Plugging in the given values, we get *F = 0.3T * 5A * 2m * sin30°*

Step 4: We know that *sin30° = 1/2*, so the equation becomes *F = 0.3T * 5A * 2m * ½*

Step 5: Simplifying this, we get *F = (0.3)(5)(2)(1/2)* = 1.5N

Step 6: Therefore, the magnitude of the magnetic force on the wire is 1.5N

The final answer is: 1.5

DPO-Shift Fine-tuned Model Response (Beta=0.3, λ_{shift}=0.75)

Step 1: The magnetic force on a current-carrying wire in a magnetic field is given by the equation

$F = IL \times B$

where I is the current, L is the length of the wire, and B is the magnetic field.

Step 2: Since the wire is placed in a uniform magnetic field, the force will be perpendicular to both the wire and the magnetic field. The magnitude of the force can be calculated using the formula *F = ILBsinϑ,* where *ϑ* is the angle between the wire and the magnetic field.

Step 3: Given that the length of the wire is 2m, the current is 5A, the magnetic field is 0.3T, and the angle between the wire and the magnetic field is *30°,* we can plug these values into the formula to find the magnitude of the force.

Step 4:

F = ILBsinϑ = (5A)(2m)(0.3T)*sin30°* = (5)(2)(0.3) (1/2) = 1.5N

Answer:1.5N

Evaluation and Quantitative Analysis

- Benchmark
- Perplexity Comparison
- Text Format Evaluation

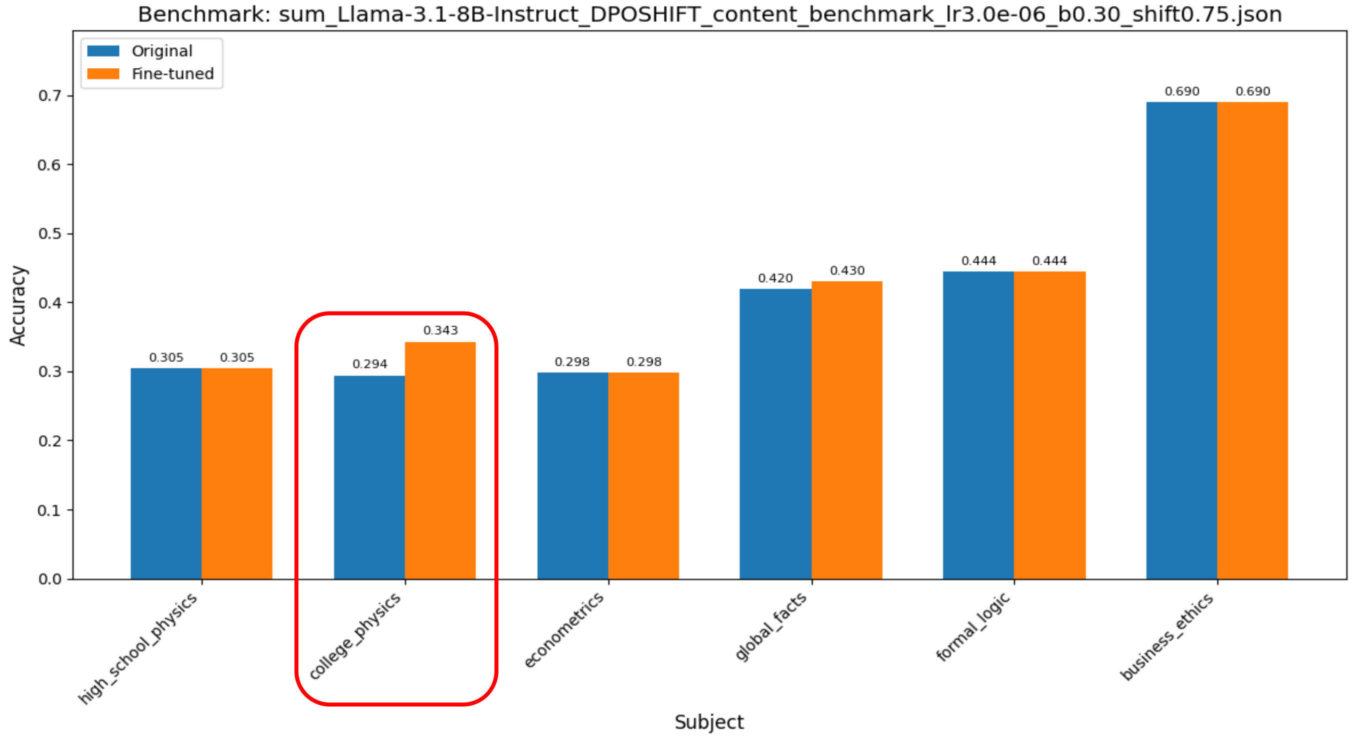
- high_school_physics: 173
- college_physics: 118
- econometrics: 131
- global_facts: 115
- formal_logic: 145
- business_ethics: 116

Evaluation and Analysis

Benchmark

TOTAL: 798 Benchmark Data

DPO-Shift



- The DPO-shift fine-tuned model does not suffer from catastrophic forgetting, and it achieves approximately a 5% marginally increase in accuracy on “college physics”.

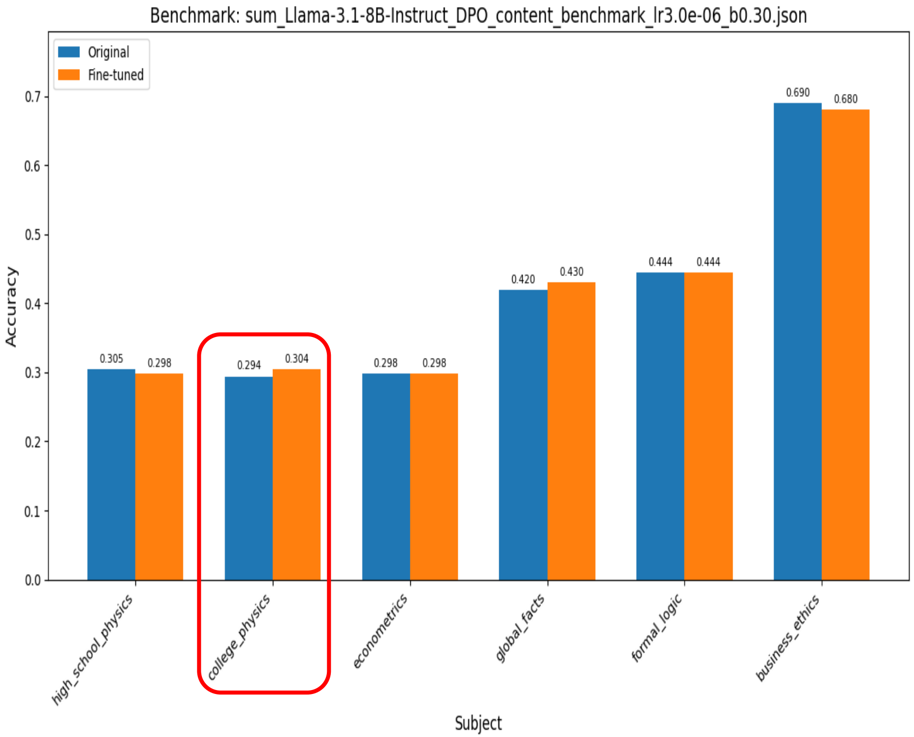
Evaluation and Analysis

Benchmark

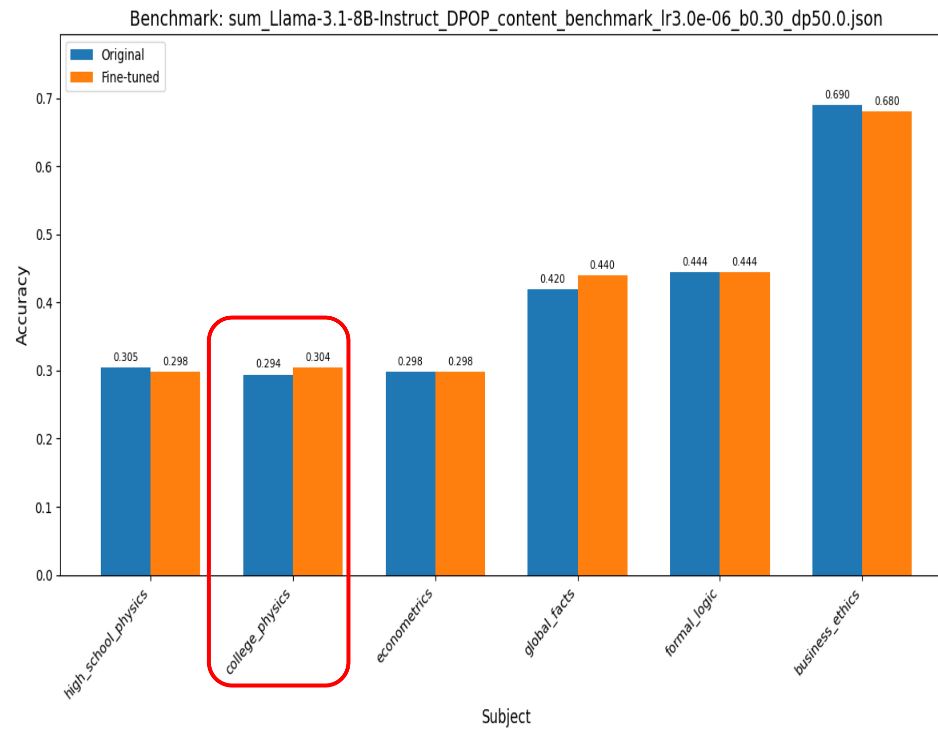
- high_school_physics: 173
- college_physics: 118
- econometrics: 131
- global_facts: 115
- formal_logic: 145
- business_ethics: 116

TOTAL: 798 Benchmark Data

DPO



DPOP



- Both DPO and DPOP fine-tuned models do not exhibit significant catastrophic forgetting.

Evaluation and Analysis

Index of Evaluation

Result Evaluation

- MMLU Benchmark Generalization and Catastrophic Forgetting
- Quantitative Metrics – **Perplexity (**How confident the model on the response)
- Model Response – **LLM-as-a-Judge on Text Format Evaluation**

➢ Model Response – LLM as a Judge

Gemini 2.5 Flash

Metrics	Description
Text Format (1-5)	how well the response's structure, wording, and presentation align with the expected answer

Evaluation and Analysis

Perplexity Comparison

$$\text{PPL}(W) = P(w_1, w_2, \ldots, w_N)^{-\frac{1}{N}}$$

- Perplexity measures how confident a language model is about a given text.
- Lower PPL = Higher Confidence → Better understanding and fluency.

	PPL of Self-generated	PPL of Preferred
Before Training (Reference Model)	2.3953 –	12.6434 –
After DPO Fine-Tuning (DPO Policy Model)	2.6235 ↑	14.4081 ↑
After DPOP Fine-Tuning (DPOP Policy Model)	2.7680 ↑	12.4052 ↓
After DPO-Shift Fine-Tuning (DPO-Shift Policy Model)	3.2302 ↑	9.5019 ↓

- **DPO-Shift** shows the **highest confidence** on preferred text (lowest PPL)
- **Trade-off:** All fine-tuning increases self-generated PPL, DPO-Shift increases the most. Improved preference alignment (especially DPO-Shift) may impact general generation in varied ways.

Evaluation and Analysis

Text Format Evaluation

- Text Format Score: 1-5

Model	Text Format
Original Model	3.0
DPO FT Model	3.01 ↑
DPOP FT Model	2.99 ↓
DPO-Shift FT Model	**3.83** ↑

- **DPO-Shift Excels in Response Text Formatting**
  - Achieves the highest score (3.83) in aligning response structure, wording, and presentation with desired formats
  - Significantly improves upon the original model (3.0) and other DPO methods

Overall Results

Overall Results

Method	$PPL_{preferred}$ ↓	$PPL_{Self\text{-}generated}$ ↓	LLM Text Format Score	MMLU College Physics
DPO	14.4081	2.6235	3.01 ↑	30.4%
DPOP	12.4052	2.7680	2.99 ↓	30.4%
DPO-Shift	9.5019	3.2302	3.83 ↑	34.3%

- **DPO-Shift Dominates Key Metrics:** Achieves best perplexity on preferred (9.5019), and the highest LLM Format score(4.14)
- **Key Trade-off:** Highest PPL on Self-generated for DPO-Shift
- **DPO-Shift's** higher MMLU College Physics score (34.3%) indicates it can better display its learned knowledge in general benchmarks that don't enforce strict output Text formats.

Findings and Outlook

Findings and Outlook

- **Findings**
  - **DPO-Shift** fine-tunes model's "**Text Format**" and "**MMLU Reasoning**" the best.
  - **DPO-Shift Leads in Quality & Alignment:** Achieves top LLM Judge scores and PPL on preferred responses, effectively learning "**Preferred Text Formats**" and "**Reasoning**."
  - **Preference Tuning: Key Trade-offs & Stability**
    - All DPO methods increase PPL on self-generated text (generality trade-off), especially DPO-Shift.

- **Outlook**
  - **Enhance Dataset Quality & Diversity**
    Enhance preference dataset quality (synthesis, verification, diversity) for improved real-world model performance. **Add randomness in rejected data.**
  - **SFT on specific domain first:** Using SFT for initial knowledge/pattern alignment before DPO-Shift to enhance domain-specific learning and expression.

Thank you!

Lun-Yu Yuan

e-mail st184762@stud.uni-stuttgart.de
phone +49 (0) 711 685-
fax +49 (0) 711 685-

University of Stuttgart
Institut of Industrial Automation and Software Engineering
Pfaffenwaldring 47, 70550 Stuttgart, Germany

Source

[1] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, “Direct Preference Optimization: Your Language Model Is Secretly a Reward Model,” Advances in Neural Information Processing Systems, vol. 36, pp. 53728–53741, 2023.

[2] F. Liu, “Learning to Summarize from Human Feedback,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA, 2020, pp. 583–592.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” arXiv preprint arXiv:1707.06347, 2017.

[4] A. Pal, D. Karkhanis, S. Dooley, M. Roberts, S. Naidu, and C. White, “Smaug: Fixing Failure Modes of Preference Optimisation with DPO-Positive,” arXiv preprint arXiv:2402.13228, 2024.

[5] X. Yang, F. Jiang, Q. Zhang, L. Zhao, and X. Li, “DPO-Shift: Shifting the Distribution of Direct Preference Optimization,” arXiv preprint arXiv:2502.07599, 2025.