## Provide a brief description and comparison of DPO and ORPO. (5%+5%)

RLHF 在 fine-tuned 階段會有最佳化的問題：

1. 需要 preference dataset 訓練 reward model,
2. 如何透過 RL 演算法 (PPO) 讓 LLM 透過 optimize reward 學習正確輸出。

DPO (Direct Preference Optimization):

主要採用直接優化策略的方式進行訓練，試圖解決 RLHF 訓練成本過高的問題，進而達到更好對於人類偏好回答的 alignment；主要的訓練方式是不透過 reinforcement learning，而是改用 supervised learning 及 preference dataset 進行訓練，不需建立 reward model，而是透過直接優化參數，避免 reinforcement learning 的複雜度以及不確定性並且提高模型的 performance。

ORPO (Odds Ratio Preference Optimization): 不同於 DPO，ORPO 對於 alignment 的優化方式透過結合 SFT 及 preference alignment，SFT 透過最大化合適答案的預測機率優化模型，但缺乏了對於 negative 的答案權重影響。ORPO 將 rejected 的 answer 透過增加一個 odd ratio loss 來去提高 choosen 的答案，降低 rejected 的答案（感覺有點像是 contrastive training）

## Briefly describe LoRA. (5%)

LoRA (Low-Rank Adaptation)

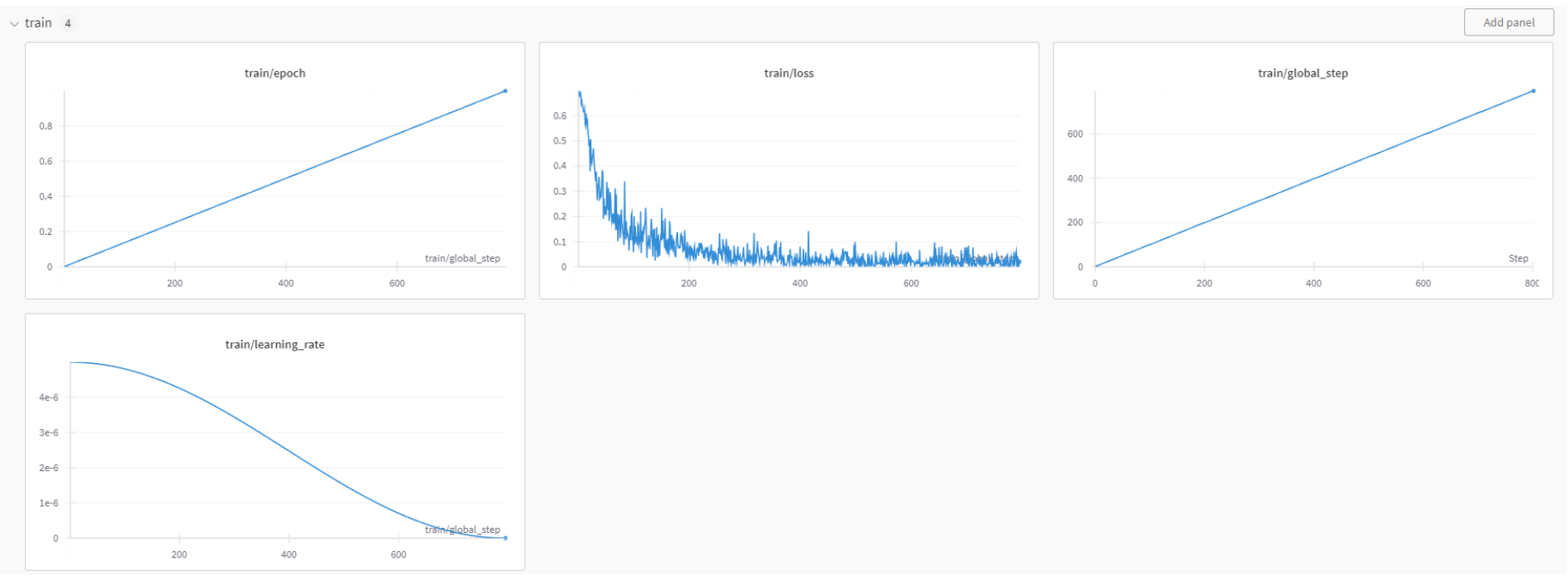
相較與幾年前的 LM (BERT-based model)，目前的 LLM 動輒便是幾個 b 的資料量，要進行訓練 (Fine-tuning) 或是 Inference 通常都需要非常可觀的 GPU，在 inference 的部分，通常可以透過 quantization 的機制去降低模型對於 GPU 的需求。而 Fine-tuning 則通常會使用 LoRA 作為微調的手段。

LoRA 的主要概念是透過凍結原始 pre-trained model 的 weight，並且在每一層都透過Low-rank approximation 去簡化矩陣，並且不會丟失大量的資訊，透過降低矩陣的秩，有效降低訓練參數量，也就是對模型進行很好的壓縮。LoRA 可以在訓練過程當中減少訓練時間及資源，並且能夠保存足夠的資訊量避免模型效能降低。

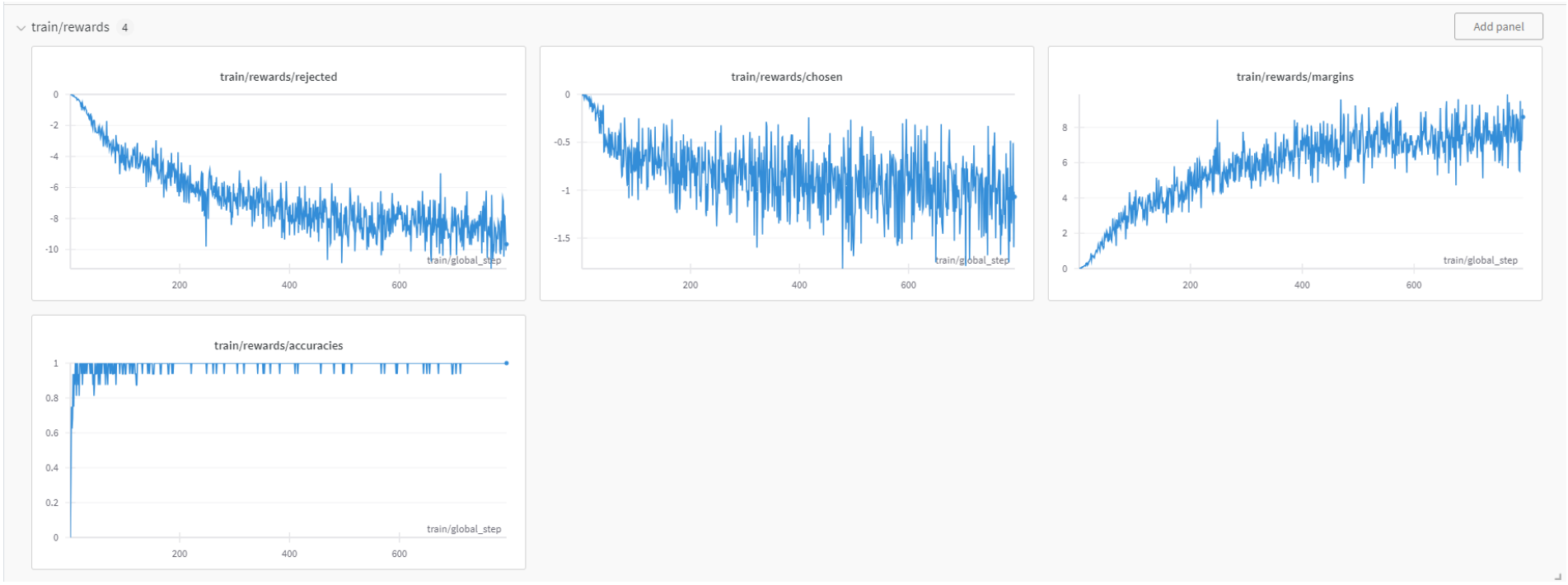
## Plot your training curve by W&B, including both loss and rewards. (5%)

1. DPO_mistral-7b-v0.3-bnb-4bit

   training loss:

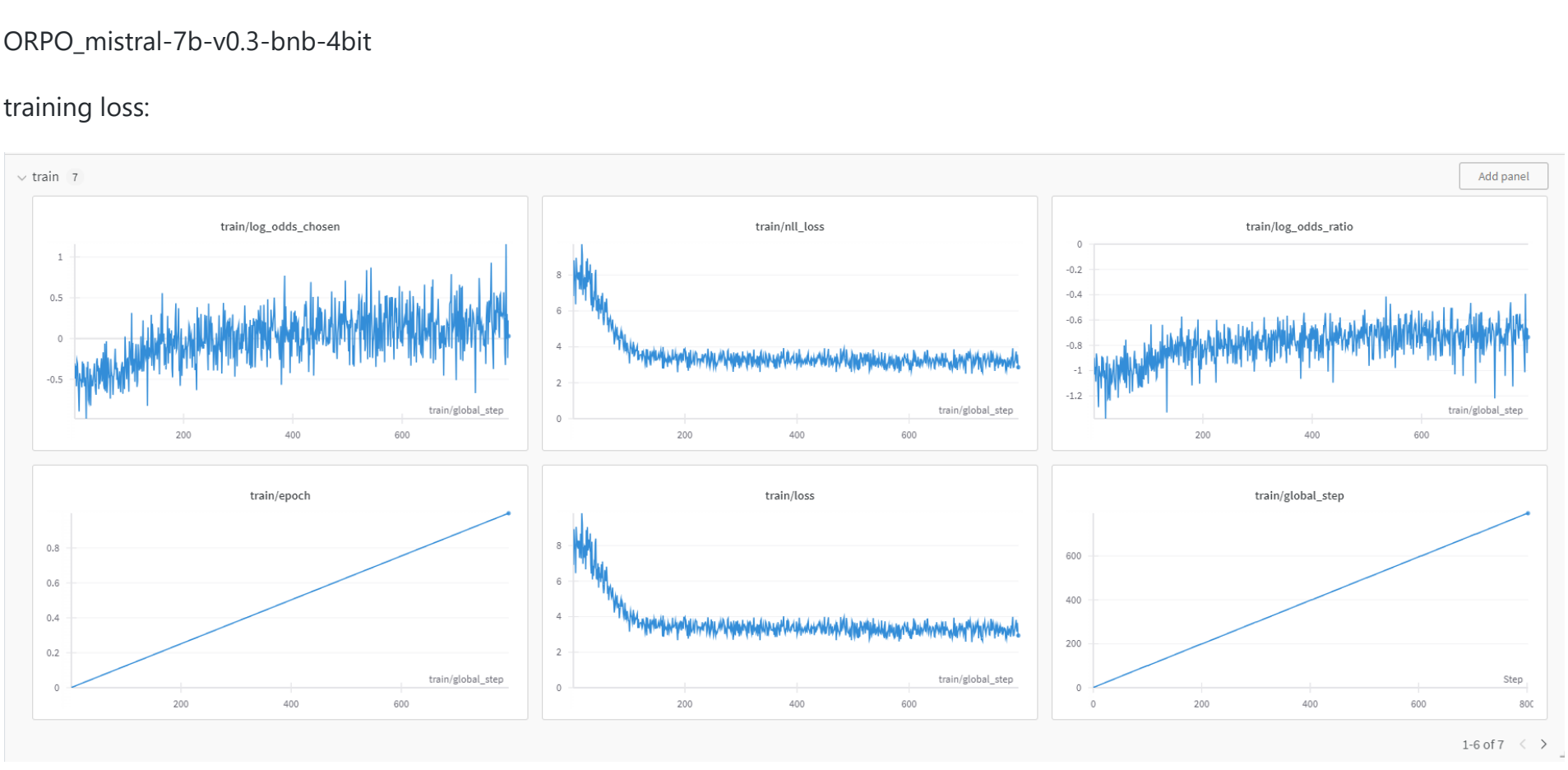


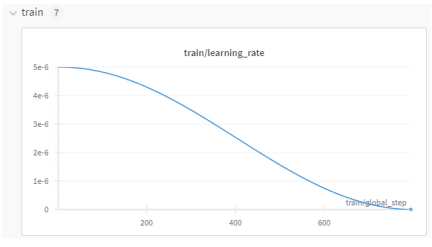

   rewards:



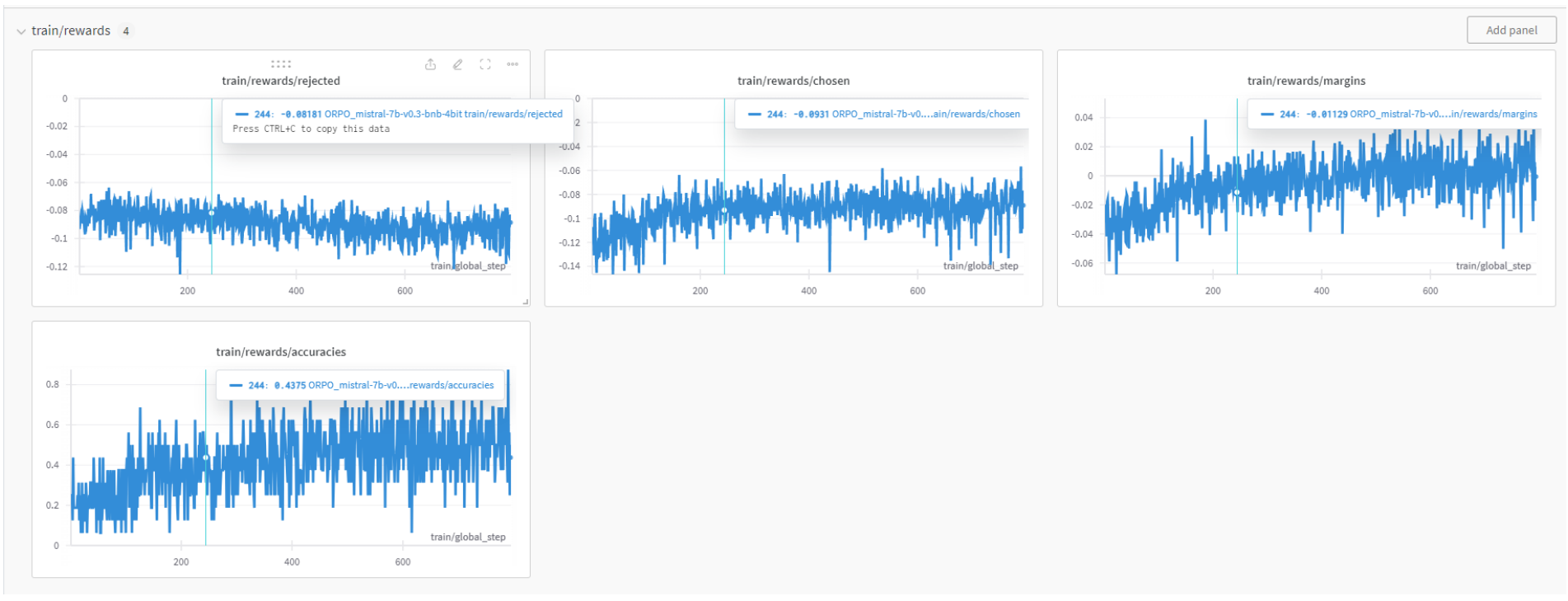

2. ORPO_mistral-7b-v0.3-bnb-4bit

training loss:




rewards:




## Comparison and analysis of results (before & after DPO & after ORPO) (5%)

針對 pre-trained model 以及 DPO, ORPO 生成的結果，我們逐一針對 10 題的生成結果進行相關比對及錯誤分析 (實驗僅就回應的合適性作為主要評估，對於部分回應的答案是否產生 hallucination 並無進一步分析)。

| id | pre-trained | DPO | ORPO | best models |
|---|---|---|---|---|
| 1 | 雖有正確回答，但未能終止回應，持續生成無關的發散內容 | 生成出答非所問的內容 | 正確回答 | ORPO |
| 2 | 生成不正確的數學解答，無法生成正確答案 | 生成答非所問的答案 | 雖有簡短的針對問題進行回覆，但是答案錯誤，正確答案應該是 10:115，卻回答成 10:75 | N/A |
| 3 | 雖有正確回答，但未能終止回應，持續生成無關的發散內容 | 雖有正確回答，但未能終止回應，持續生成無關的發散內容 | 正確回答相關內容 | ORPO |
| 4 | 正確回答相關內容 | 正確回答相關內容，但無條列陳述 | 正確回答相關內容 | pre-trained, ORPO |
| 5 | 雖有正確回答，但未能終止回應，持續生成無關的發散內容 | 雖有正確回答，但未能終止回應，持續生成無關的發散內容 | 正確回答相關內容 | ORPO |
| 6 | 生成出答非所問的內容 | 生成出答非所問的內容 | 正確回答相關內容 | ORPO |
| 7 | 正確回答相關內容 | 雖有正確回答，但未能終止回應，持續生成無關的發散內容 | 正確回答相關內容 | pre-trained, ORPO |
| 8 | 有生成類似正確的答案 (未求證) | 有生成類似正確的答案 (未求證) | 有生成類似正確的答案 (未求證) | pre-trained, DPO, ORPO |
| 9 | 生成出答非所問的內容 | 生成出答非所問的內容 | 有生成類似正確的答案 (未求證) | ORPO |
| 10 | 同第二題 | 同第二題 | 同第二題 | N/A |

上述觀察包含了以下假設

- 每一項目雖都透過 human 方式人工 review，但部分問題無法由人來直觀的確認答案的正確性，僅就答案的合適性進行評估。

藉由結果推論分析

1. Pre-trained model 及 DPO 都無法在有限的文字下回覆答案後即結束生成，大多問題都會發生不斷發散或是重複 input 及 response 的可能，導致生成的內容發散或是重複回答
2. DPO 雖然是一種能夠降低 RLHF 訓練成本的手段，但實務上，在 Mistral 的表現上即使 fine-tuned 後的表現仍較差，甚至比 pre-trained Mistral 的結果還要來得差。
3. ORPO 對於數學問題可能在計算上無法精確計算出答案
4. ORPO 的生成遠較於 DPO 及 pre-trained 模型來得好，表示基於 ORPO 的微調確實能夠有效的大幅幫助模型能夠更友善的回覆使用者問題

## Extra experiments

針對上述實驗的分析，額外進行了兩部分的實驗

1. 目前 LLM (例如: llama3, Mistral) 在發表時都會透過許多不同任務 (例如: MMLU 等) 進行效能的評估，並且多以 GPT 作為 benchmark 的對象，又或是一些競賽 (例如 Meta 近期的 Meta CRAG RAG 競賽) 也會將 GPT 的結果視為 ground truth 作為階段性的評審，主要原因在於 GPT 這樣資料量級的模型仍具備指標性，所以額外針對近期發表的 GPT-4o 將 test set 提供的 10 個 prompt 藉由 GPT 產生結果。此處的實驗設置較為單純，並不設定額外的 system prompt 以及 few shot 等內容，僅直接將 10 個 prompt 丟給 GPT-4o 紀錄其答案。(答案紀錄於 submission 資料夾中的 gpt4-o.json)
2. 另外在原有實驗分析中發現 ORPO 已獲得了較好的結果，而 DPO 則出乎意料並沒有更的效能提升，故嘗試對 DPO 的 hyper parameters 進行微調，主要微調內容如下：

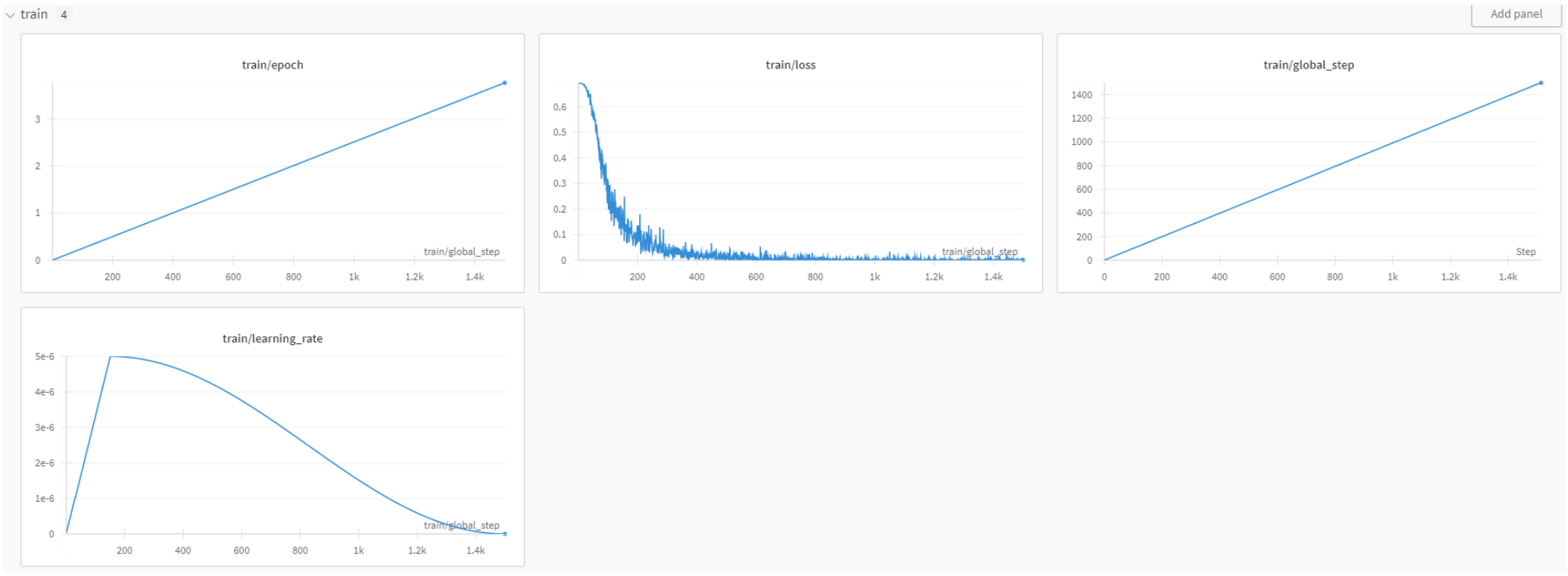
- **train_batch_size 和 eval_batch_size**：
  - **train_batch_size**: 增大批量大小嘗試讓模型在每個步驟中使用更多的數據進行更新，觀察是否能提升模型的穩定性和收斂速度。
  - **eval_batch_size**: 增大評估批量大小提高評估過程的效率和穩定性。
- **max_steps 和 num_epochs**：
  - **max_steps**: 增大最大訓練步數讓模型有更多的學習機會。
  - **num_epochs**: 增大訓練迭代次數可以讓模型學習得更充分。
- **weight_decay**：
  - 增加權重衰減防止過擬合。
- **warmup_ratio**：
  - 設置預熱比例，在訓練初期使用較低的學習率可以幫助模型穩定訓練。

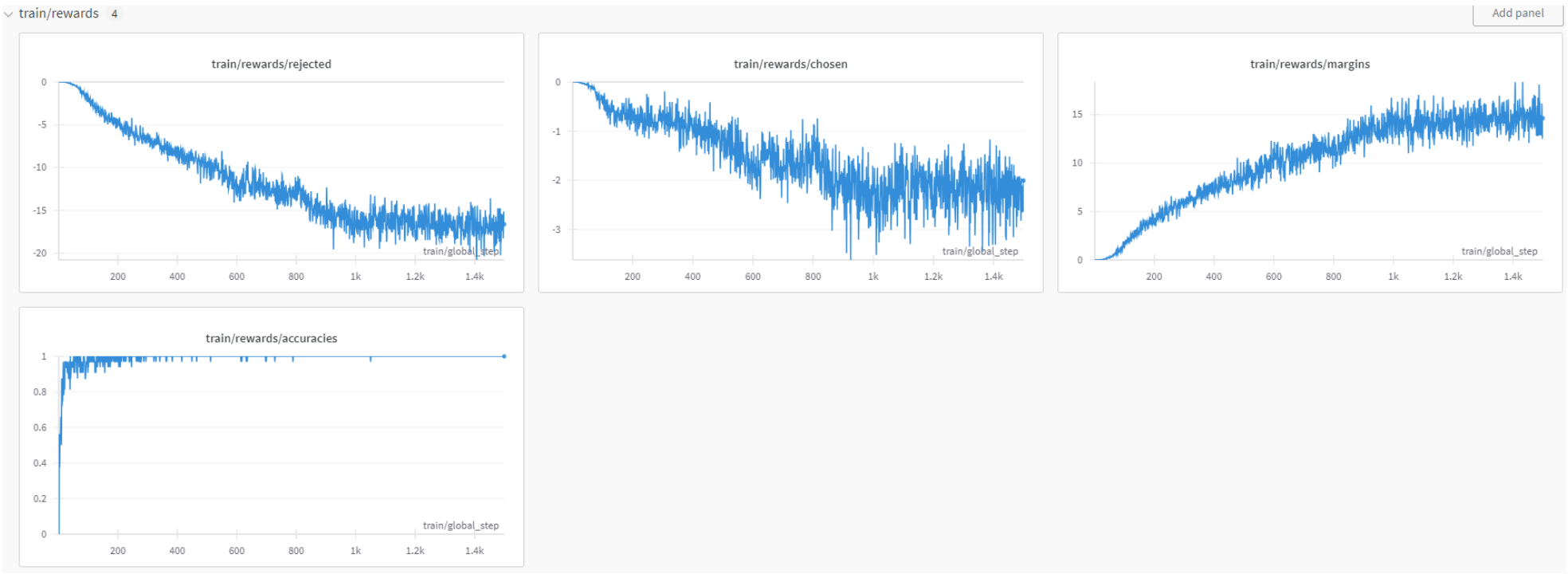
詳細調參如下：

| hyper parameter | default | fine-tuned |
|---|---|---|
| train_batch_size | 2 | 4 |
| eval_batch_size | 2 | 4 |
| max_steps | 0 | 1500 |
| num_epochs | 1 | 3 |
| weight_decay | 0 | 0.01 |
| warmup_ratio | 0 | 0.1 |

模型結果儲存於 submission 資料夾中的 DPO_mistral-7b-v0.3-bnb-4bit_finetuned.json

DPO 的 training loss 及 reward 如下:






實驗結果分析如下表 (實驗僅就回應的合適性作為主要評估，對於部分回應的答案是否產生 hallucination 並無進一步分析)

| id | GPT-4o | DPO_finetuned | best models |
|---|---|---|---|
| 1 | 正確回答相關內容 | 生成答非所問的答案 | GPT-4o |
| 2 | 正確回答相關內容 | 雖有針對問題回答，但答案錯誤，且未能終止回應，持續生成無關的發散內容 | GPT-4o |
| 3 | 正確回答相關內容 | 雖有正確回答，但未能終止回應，持續生成無關的發散內容 | GPT-4o |
| 4 | 正確回答相關內容 | 雖有針對問題回答，但答案錯誤，且未能終止回應，持續生成無關的發散內容 | GPT-4o |
| 5 | 正確回答相關內容 | 正確回答相關內容 | GPT-4o, DPO |
| 6 | 正確回答相關內容 | 生成答非所問的答案 | GPT-4o |
| 7 | 正確回答相關內容 | 生成答非所問的答案 | GPT-4o |
| 8 | 正確回答相關內容 | 生成答非所問的答案 | GPT-4o |
| 9 | 正確回答相關內容 | 生成答非所問的答案 | GPT-4o |
| 10 | 正確回答相關內容 | 生成答非所問的答案 | GPT-4o |

從生成結果來看，GPT-4o 壓倒性的在各類問題中 (數學、知識問題等) 都能提供精準、簡短、正確且符合人類適合的回應的內容，即使是數學相關問題也能正確計算答案後生成相關內容。而 DPO 即使透過相關參數調整，期望能藉由更多的 batch size 或是 epoch 的增加去強化其生成的答案，但在回應上仍然沒有獲得更好的成效，仍然有許多幻覺問題待解決。此部分考慮未來可以透過其他參數的調整 (beta, etc) 測試哪些參數的微調能夠提升 DPO 的準確率。