- Report
  - Provide a brief description and comparison of DPO and ORPO.

    DPO (Data-Driven Policy Optimization)

    DPO 是一種基於數據驅動的策略優化方法，它利用已有的數據集來訓練語言模型，以優化特定的性能指標或任務目標。在 NLP 中，DPO 可以用來微調語言模型，使其在特定類型的問題上表現更好，如回答問題、生成文本等。DPO 通常會集中於如何更有效地從數據中學習策略，進而改善模型對於實際應用場景的適應性。

    ORPO (Off-policy Reinforcement Learning with Policy Optimization)

    ORPO 是一種結合了離策略（Off-policy）強化學習和策略優化技術的方法。離策略強化學習允許模型學習策略從與執行策略不同的數據中進行學習，這種方法使模型能夠有效利用過去的經驗進行學習，即使這些經驗來自於不同的行為策略。在 NLP 中，ORPO 可用於進行更復雜的交互式學習任務，如對話系統，其中模型需要從非同步反饋中學習。

    比較

    - **應用範圍**：DPO 著重於利用大規模數據集來改進特定任務的表現，而 ORPO 則較多用於需要從複雜環境或交互式任務中學習的場景。
    - **數據效率**：ORPO 通過離策略學習可以更有效地利用過往的學習經驗，相對於 DPO 可能在某些情況下更加數據效率。
    - **複雜性**：ORPO 的實現通常比 DPO 更複雜，因為它涉及到對策略的額外優化以及處理來自不同策略的數據。

    總的來說，選擇使用 DPO 還是 ORPO 取決於具體的應用需求和可用數據的性質。在某些情況下，結合兩者的優點可能會帶來更好的學習效果和模型表現。
  - Briefly describe LoRA

    是一種用於調整大型預訓練語言模型的技術。透過在模型的特定層中添加低秩矩陣來進行微調，而不是直接調整原有的大型權重矩陣。這種方法減少了微調過程中需要更新的參數數量，從而節省了計算資源，同時保持了模型的性能。LoRA 通過在模型的前饋網路中插入可學習的小型矩陣，實現了這一點。
  - Plot your training curve by W&B, including both loss and rewards
  - Comparison and analysis of results (before & after DPO & after ORPO)

    性能比較:

    基線對比 DPO：DPO 通常會專注於使用大量數據來優化特定的性能指標，所以在任務特定的性能上，您可能會觀察到明顯的提升。比如在一個特定類型的問題回答或者對話生成任務中，模型經過 DPO 後能

更準確地鑑別和回答問題。

DPO 對比 ORPO：由於 ORPO 涉及到更複雜的學習機制，它允許模型從非同策略的數據中學習，這可能進一步提高模型在更廣泛或更多樣化的情景下的適應性和性能。特別是在需要從用戶互動中動態學習的場景，ORPO 可能表現更好。

- **Extra Experiments**
  - Comparison and analysis of various hyperparameters (e.g., num_epochs, beta, etc.)
  - Comparison and analysis of various models, which GPT-4, GPT-4o, GPT-3.5, and etc. are allowed, for the variants of the required models please refer to here.