# RLHF

R12922A09
Yung-Hsiang Yang

## 1. DPO v.s. ORPO

### DPO (Direct Preference Optimization)

**方法：**直接最佳化模型的預測結果，使其盡可能符合用戶或系統的偏好。

**應用：**通常用於需要精細調整輸出結果以符合特定偏好或約束的情境，如推薦系統或個性化搜索。

**特點：**直接處理偏好數據，目標是最小化偏好錯誤，直接影響模型的預測結果。

**優點：**能夠精確對齊用戶或系統的偏好。通常能在偏好導向的任務中獲得更高的滿意度和效能。

### ORPO (Odds Ratio Preference Optimization)

**方法：**使用機率比（Odds Ratio）來量化和最佳化偏好，特別是針對二分類問題。

**應用：**常見於風險評估、醫療診斷等領域，需要考慮不同決策的相對風險或收益。

**特點：**透過最佳化機率比，平衡不同類別的偏好，特別適合於處理不平衡數據集或需要考慮錯誤代價的情境。

**優點：**能夠有效處理不平衡數據集，提供更穩健的預測。有助於理解和調整不同決策之間的相對風險或收益。

## LoRA

LoRA（Low Rank Adaptation）是一種在機器學習和深度學習中用於參數的方法。這種技術適用於大模型的微調（fine-tuning）。

LoRA 的概念是，既然 LLM 適用於不同任務，那代表模型對於不同任務會有不同的神經元/特徵來處理這件事，如果我們能從眾多特徵中找到適合那個下游任務的特徵，並對他們的特徵進行強化，那我們就能對特定任務有著更好的成果