Proposed Topic: Direct Preference Optimization (DPO) for LLM Training

Presentation outline

Objective: Provide a clear and rigorous overview of DPO as an alternative to RLHF for preference-based alignment of large language models.

1. **Introduction to Preference-Based Alignment**
    - Motivation: why align LLMs to human preferences?
    - Brief recap of RLHF pipeline (reward modeling + PPO fine-tuning)
2. **From RLHF to DPO**
    - Theoretical equivalence between DPO and RLHF under specific assumptions
    - Explanation of the DPO loss and training procedure
3. **Comparison: DPO vs RLHF**
    - Comparative table: efficiency, stability, compute requirements, alignment quality, complexity
    - Practical trade-offs and adoption in practice
4. **State of the Art**
    - What is used in practice today?
    - Recent extensions and improvements

Written Report

1. **Implementation**
    - Implementation of DPO training pipeline on a preference dataset
    - *Optional:* simplified RLHF pipeline for comparison
2. **Evaluation Framework**
    - Metrics: win-rate against base or reference model, empirical quality evaluation
    - Study of hyperparameters (e.g., effect of β)
    - Qualitative examples illustrating behavioral differences