Bradley-Terry Preference Model

P(item+ > item- | user) = sigmoid(score+ - score-)

## DPO Objective Function

Single-stage training with theoretical guarantees
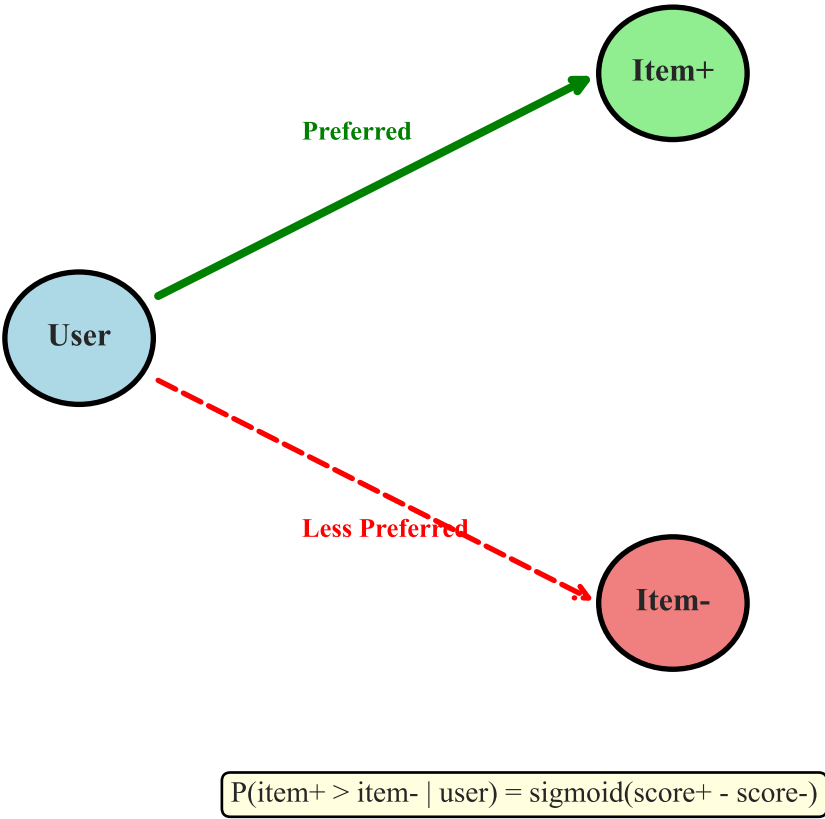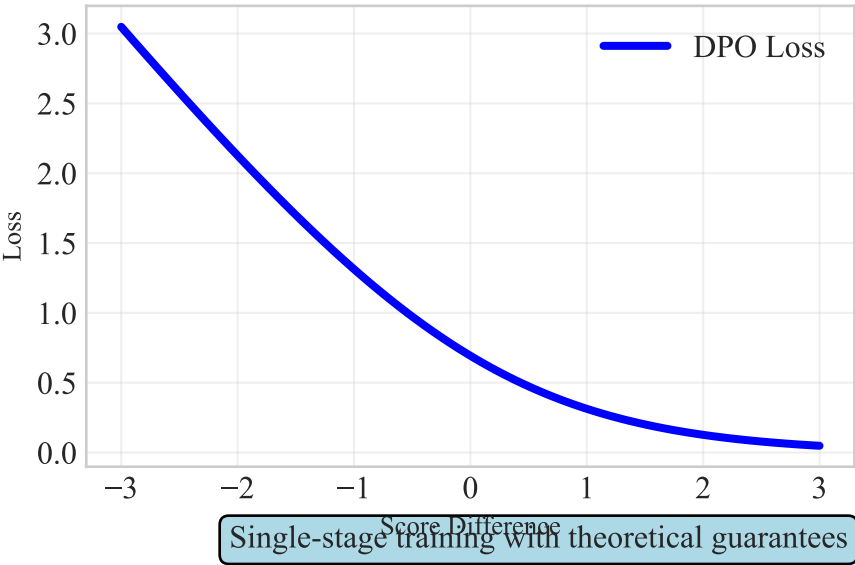
DPO Training Flow
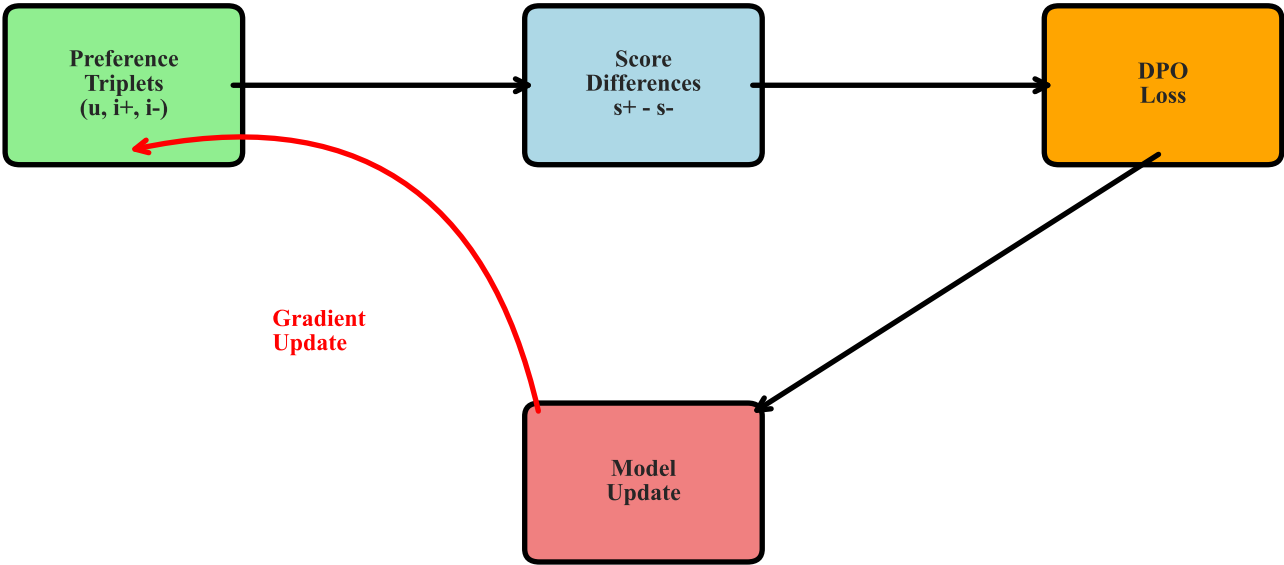
DPO vs Traditional Methods