Bradley-Terry Model

Mathematical Foundation

The Bradley-Terry model converts reward scores into probabilities for pairwise comparisons, providing a principled approach to preference learning.

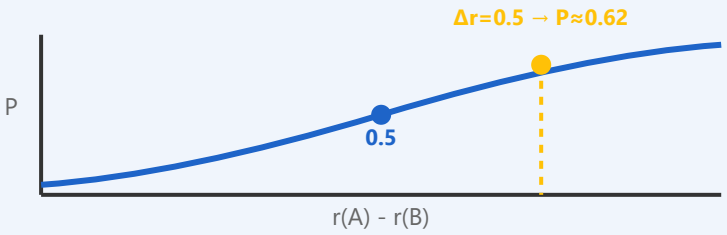
Model Formula & Visualization

$$P(A > B) = \sigma(r(A) - r(B)) = 1 / (1 + \exp(-(r(A) - r(B))))$$

Output A
r(A) = 0.8

VS

Output B
r(B) = 0.3



Higher Reward Difference → Stronger Preference Probability

💡 σ (sigmoid function) ensures probability output between 0 and 1

Training Objective

- Maximize log-likelihood of observed preferences
- Loss = -log(P(preferred > not_preferred))
- Gradient descent updates reward model parameters
- Converges to scores matching expert preferences