

Reward Model Architecture

Model Structure

Reward models predict scalar scores for model outputs, learned from expert preference data to guide policy optimization.

Architecture Components

Input Layer

Tokenized Output Pairs (Output A, Output B)



Base Encoder (Transformer)

BERT, GPT, or Medical Domain Transformer



Pooling Layer

Aggregate Sequence Representations (Mean/Max/CLS)



Reward Head (Linear Layer)

**Loss Function**

Bradley-Terry / Ranking Loss

Training Process

1 Input

Pairs of outputs with preference labels

3 Loss Calculation

Penalize incorrect rankings

2 Forward Pass

Compute reward scores for both outputs

4 Optimization

Adam optimizer with LR scheduling