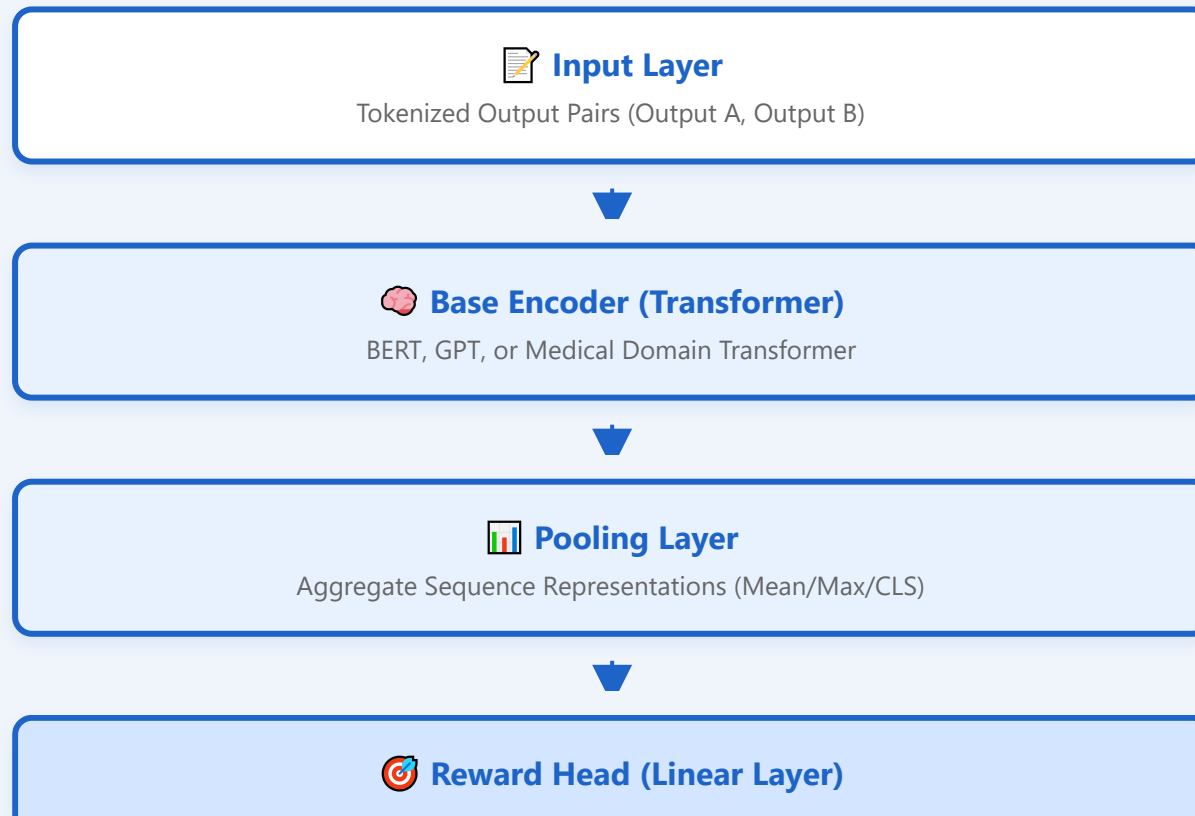


# Reward Model Architecture

## Model Structure

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

## Architecture Components



Output: Single Scalar Score  $r(\text{output})$



### Loss Function

Bradley-Terry / Ranking Loss

## Training Process

### 1 Input

Pairs of outputs with preference labels

### 2 Forward Pass

Compute reward scores for both outputs

### 3 Loss Calculation

Penalize incorrect rankings

### 4 Optimization

Adam optimizer with LR scheduling