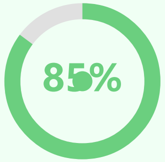
Uncertainty Estimation in Reward Models

Why Uncertainty Matters

In medical AI, knowing when the reward model is uncertain helps identify cases requiring additional expert review or model improvement.

Uncertainty Spectrum Visualization



High Confidence

✓ Deploy

Model is certain, safe to use

Medium Confidence

⚠️ Review

Uncertain, flag for expert

Low Confidence

🚫 Block

High uncertainty, do not use

Estimation Methods

Ensemble Methods

Train multiple reward models with different initializations, measure prediction variance

Bayesian Approaches

Model weight uncertainty with probability distributions (e.g., Bayesian Neural Networks)

Monte Carlo Dropout

Apply dropout during inference for variance estimation across multiple forward passes

Calibration

Ensure predicted confidence matches empirical accuracy using calibration techniques

Applications

- Active Learning: Query experts on high-uncertainty cases
- Safe Deployment: Flag uncertain predictions for human review
- Model Improvement: Identify areas needing more training data
- Confidence Intervals: Provide uncertainty bounds with predictions