Lecture 11 - Contents

An overview of the main sections in this lecture.

Part 1
Medical Reward Modeling

Part 2
Policy Optimization

Part 3
RLHF in Clinical Practice

Hands-on
RLHF Pipeline

This outline is for guidance. Navigate the slides with the left/right arrow keys.

Lecture 11:

RLHF in Healthcare: Aligning AI with Medical Expertise

Ho-min Park

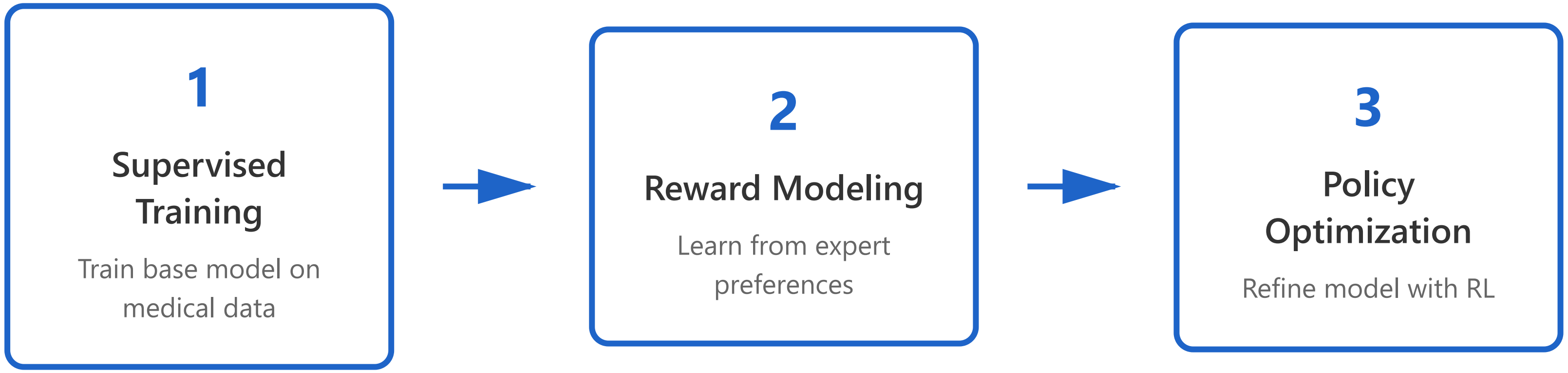
homin.park@ghent.ac.kr

powersimmani@gmail.com

RLHF Overview in Healthcare

What is RLHF?

Reinforcement Learning from Human Feedback aligns AI models with medical expertise through iterative feedback loops, ensuring outputs match clinical standards and safety requirements.



1

Supervised Training

Train base model on medical data

2

Reward Modeling

Learn from expert preferences

3

Policy Optimization

Refine model with RL

🎯 Clinical Alignment

Ensures AI decisions match medical standards

🛡️ Safety Enhancement

Reduces harmful or inappropriate outputs

📊 Continuous Improvement

Adapts to evolving medical knowledge

⚕️ Expert Integration

Incorporates physician judgment directly

Part 1/3:

Medical Reward Modeling

1. Clinical Preference Learning
2. Expert Feedback Collection
3. Annotation Interfaces
4. Preference Dataset Creation
5. Reward Model Architecture
6. Bradley-Terry Model
7. Uncertainty Estimation

Clinical Preference Learning

What is Clinical Preference Learning?

Process of capturing expert medical judgment through pairwise comparisons of model outputs, enabling the AI to learn what constitutes high-quality clinical responses.

Collection Methods

- Pairwise Comparisons: Experts choose between two model outputs
- Ranking Tasks: Order multiple outputs by quality
- Absolute Scoring: Rate individual outputs on fixed scales
- Natural Language Feedback: Detailed written critiques

Key Considerations

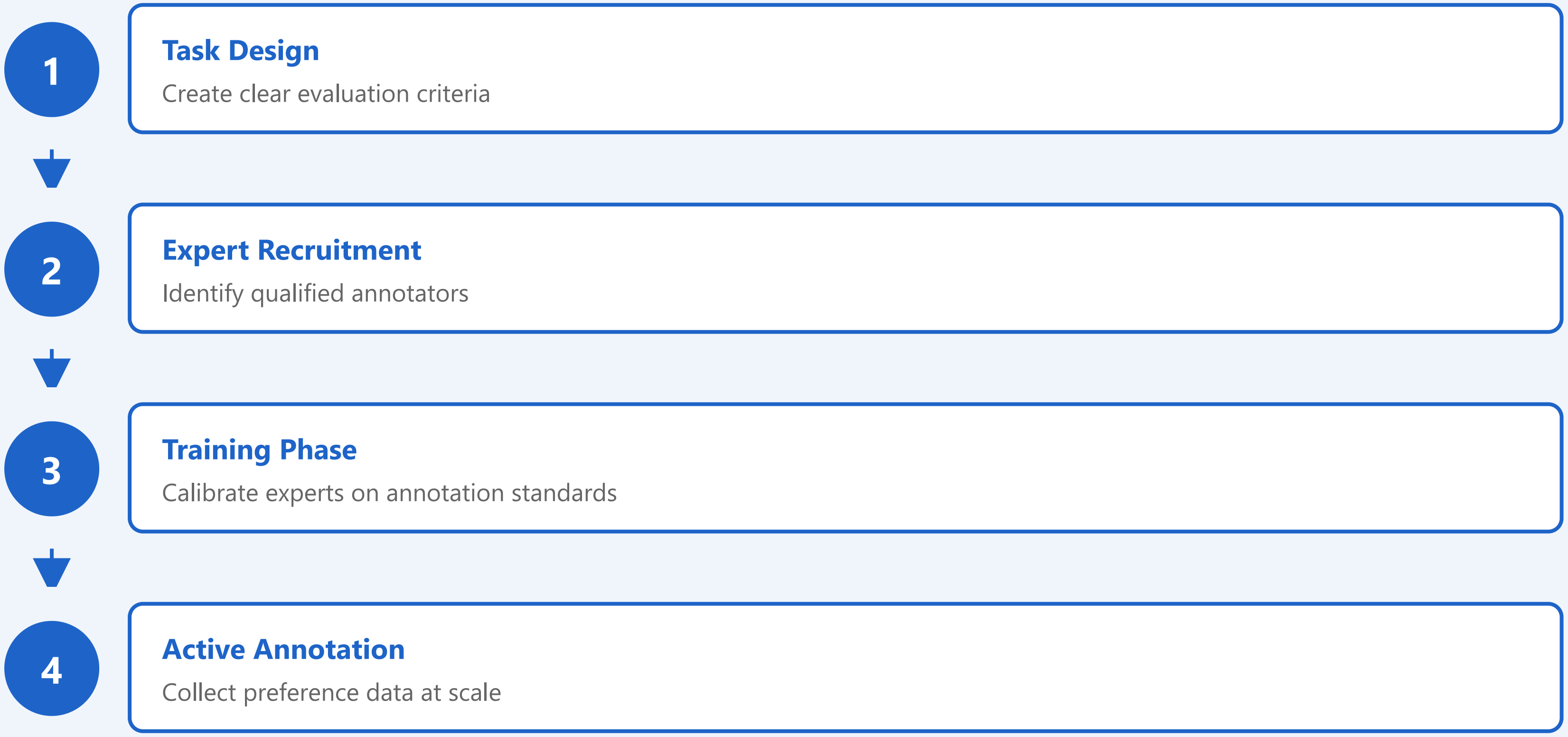
- Inter-annotator Agreement: Ensure consistency across experts
- Sample Diversity: Cover wide range of clinical scenarios
- Quality Control: Regular calibration sessions
- Expert Qualifications: Board-certified specialists in relevant domains

Expert Feedback Collection

Annotation Protocol

Structured guidelines ensure consistent, high-quality feedback from medical experts across all evaluation tasks.

Collection Pipeline

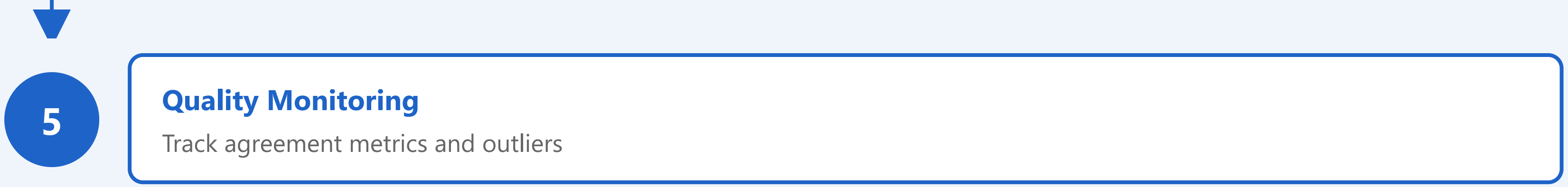


1 **Task Design**
Create clear evaluation criteria

2 **Expert Recruitment**
Identify qualified annotators

3 **Training Phase**
Calibrate experts on annotation standards

4 **Active Annotation**
Collect preference data at scale



5

Quality Monitoring

Track agreement metrics and outliers

Quality Assurance

- Gold Standard Examples: Pre-annotated cases for validation
- Inter-rater Reliability: Cohen's Kappa, Fleiss' Kappa
- Regular Audits: Review annotation quality periodically
- Feedback Loops: Discuss disagreements and edge cases

Annotation Interfaces

Interface Design Principles

User-friendly annotation tools maximize expert efficiency and reduce cognitive load during evaluation tasks.

Key Features

- Side-by-Side Comparison: View outputs simultaneously
- Contextual Information: Patient history, relevant guidelines
- Quick Actions: Keyboard shortcuts for common tasks
- Progress Tracking: Visual indicators of completion
- Comment System: Add detailed notes and rationale

Workflow Optimization

- Adaptive Sampling: Focus on uncertain cases
- Break Reminders: Prevent annotation fatigue
- Session Management: Save progress automatically
- Mobile Compatibility: Annotate on various devices

Preference Dataset Creation

Dataset Structure

Preference datasets contain pairs of model outputs with expert rankings, forming the foundation for reward model training.

Data Components

📝

1. Input Prompt

Clinical query or task description

A

2. Output A

"Patient presents with fever and cough. Recommend chest X-ray and CBC."

B

3. Output B

"Recommend immediate chest X-ray, CBC, and respiratory pathogen panel given symptoms."

✓

4. Preference Label

Output A → ✗ Output B ✓

Expert prefers B: More comprehensive diagnostic approach

⭐

5. Confidence Score (Optional)

Strength: High (0.9) - Clear preference for comprehensive testing

6. Rationale

Expert's reasoning: "Output B includes respiratory pathogen panel which is essential for differential diagnosis in current clinical context."

Dataset Statistics

10K-100K+
Preference Pairs

50/50
Balance (A vs B)

15+ Specialties
Clinical Coverage

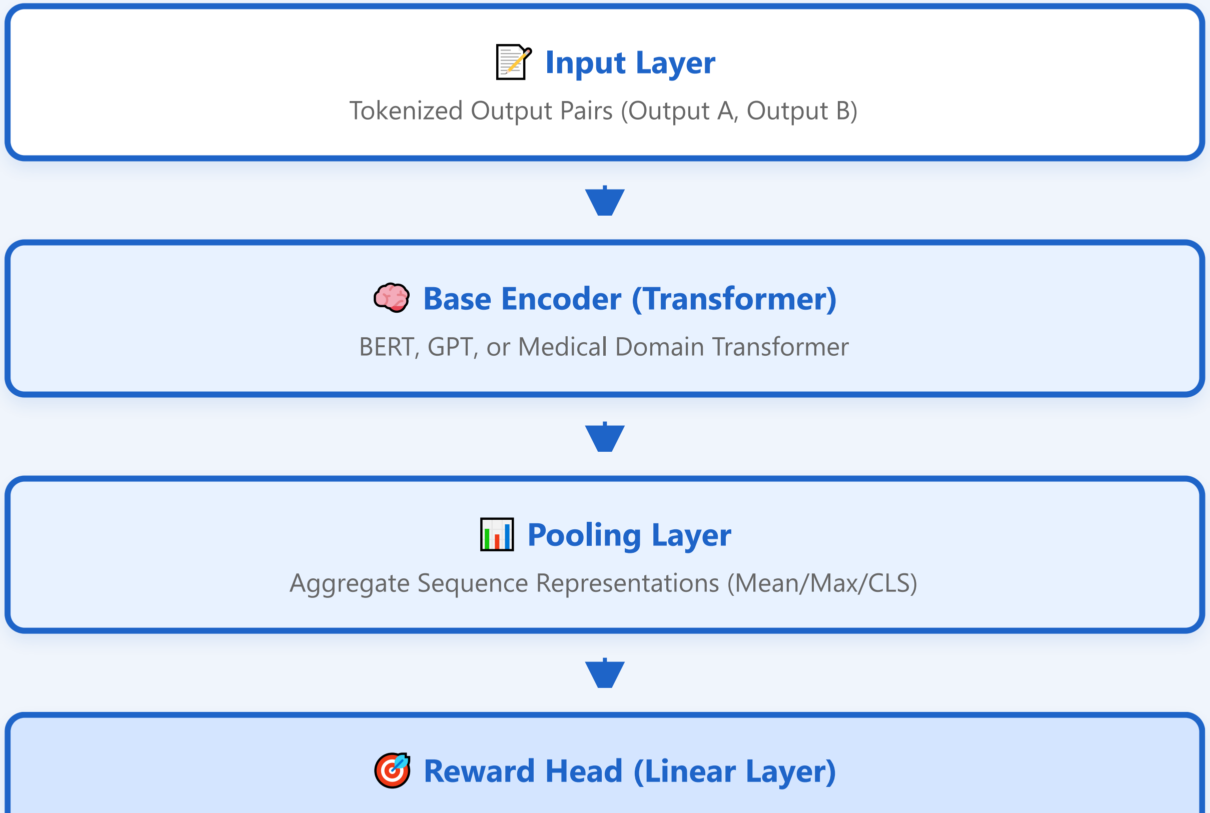
85-95%
Agreement Rate

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)

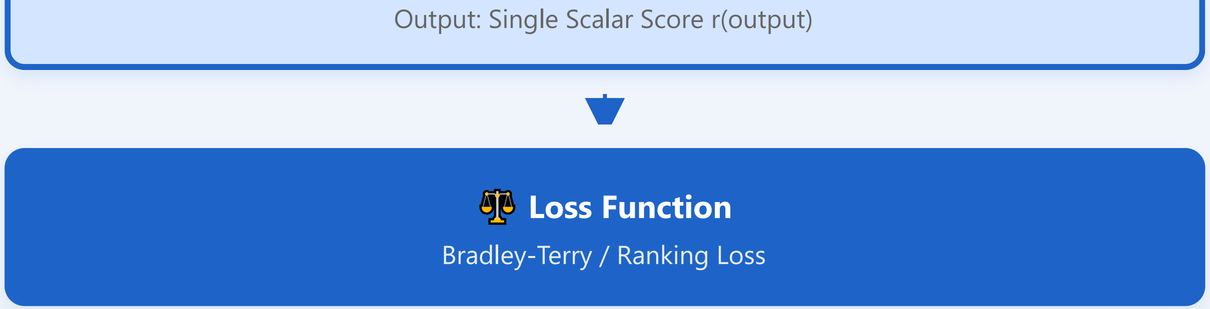
🧠 **Base Encoder (Transformer)**

BERT, GPT, or Medical Domain Transformer

📊 **Pooling Layer**

Aggregate Sequence Representations (Mean/Max/CLS)

🎯 **Reward Head (Linear Layer)**



Output: Single Scalar Score r(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

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

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

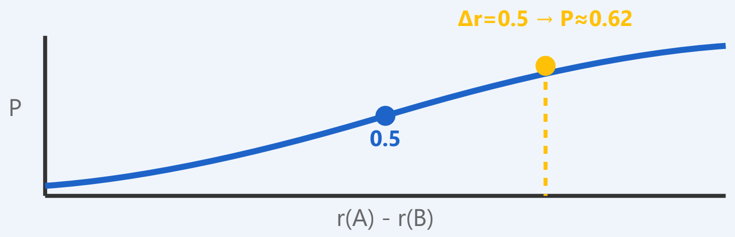
Output A

$r(A) = 0.8$

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

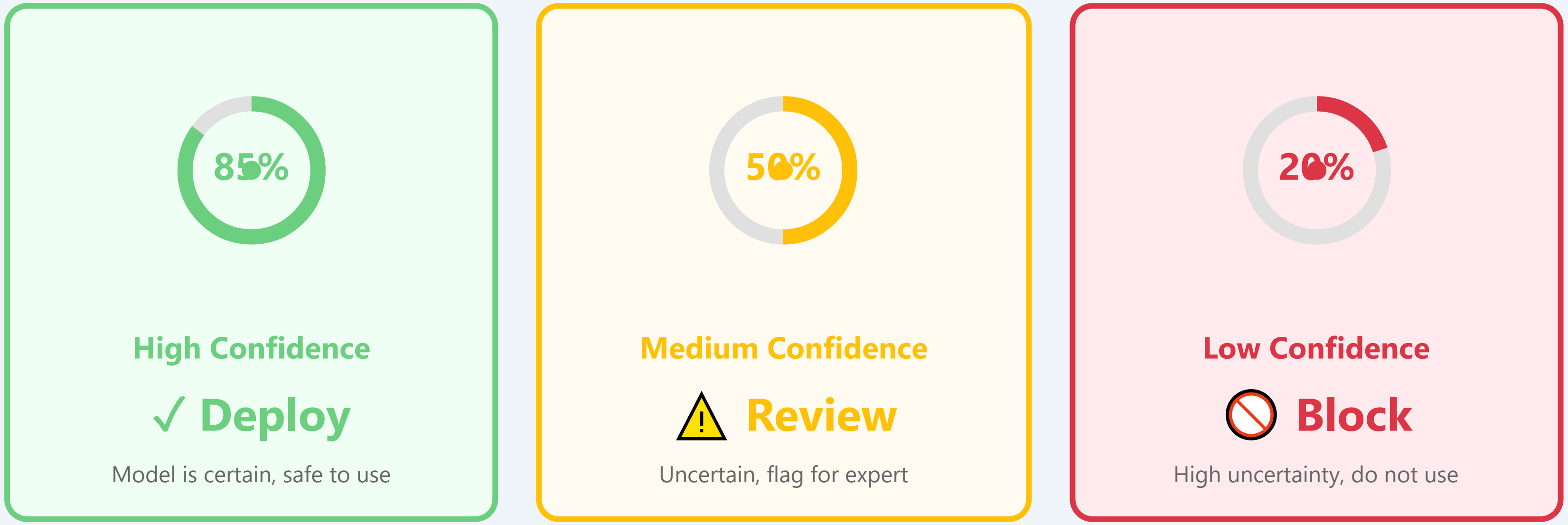
- Maximize log-likelihood of observed preferences
- Loss = -log(P(preferred > not_preferred))
- Gradient descent updates reward model parameters
- Converges to scores matching expert preferences

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

Part 2/3:

Policy Optimization in Medical AI

1. PPO in Medical Applications
2. Direct Preference Optimization (DPO)
3. Safety Constraints
4. KL Divergence Control
5. Exploration vs Exploitation
6. Online vs Offline RLHF

PPO in Medical Applications

Proximal Policy Optimization (PPO)

PPO is the most widely used algorithm for RLHF, providing stable updates that prevent catastrophic policy changes in medical AI systems.

Key Features

✂️ Clipped Objective

Limits policy updates to prevent drastic changes

🛡️ Trust Region

Maintains proximity to previous policy

📈 Stability

Reduces training instability vs vanilla PG

♻️ Sample Efficiency

Reuses data multiple times per iteration

Training Process (Iterative Loop)

1

2

3

Direct Preference Optimization (DPO)

What is DPO?

DPO directly optimizes the language model using preference data, eliminating the need for a separate reward model and RL training loop.

DPO vs PPO: Architecture Comparison

PPO

Two-Stage Process

- **Stage 1:** Train Reward Model
- 📊 Preference Data → r(output)
- **Stage 2:** RL Optimization
- 🔄 Actor-Critic + KL penalty
- ⚙️ Complex hyperparameters

VS

DPO

Single-Stage Process

- **Stage 1:** Direct Optimization
- ✓ Preference Data → Policy Update
- 🎯 No separate reward model
- ⚡ Simpler training pipeline
- 💾 Lower memory requirements

Safety Constraints in Medical RLHF

Why Safety Constraints?

Medical AI requires hard constraints to prevent harmful outputs that could endanger patients, regardless of reward optimization.

Multi-Layer Safety Architecture

🚫 **Layer 1: Hard Medical Rules**

Never violate established clinical guidelines (e.g., contraindications, age restrictions)

💊 **Layer 2: Dosage Limits**

Enforce safe medication dosing ranges based on patient factors (weight, age, renal function)

⚠ **Layer 3: Contraindication Checking**

Prevent dangerous drug interactions and allergic reactions

🎯 **Layer 4: Appropriate Scope**

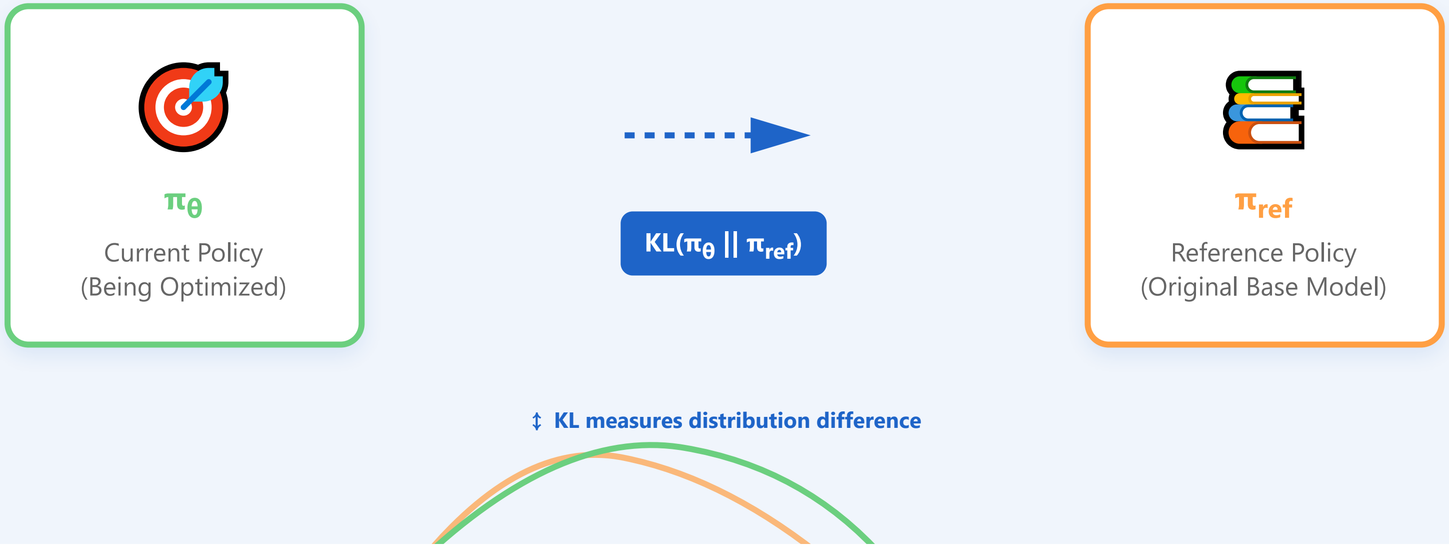
Stay within system's trained competency domain

KL Divergence Control

Purpose of KL Divergence

KL divergence measures how much the optimized policy deviates from the original base model, preventing harmful divergence from foundational knowledge.

Visual Representation



🎯

π_θ

Current Policy
(Being Optimized)

$KL(\pi_\theta \,||\, \pi_{ref})$

📚

π_{ref}

Reference Policy
(Original Base Model)

↕ KL measures distribution difference

Exploration vs Exploitation in Medical AI

The Dilemma

Balancing exploration (trying new approaches) with exploitation (using known-good approaches) is critical in medical AI where safety is paramount.

The Balance Concept

🔍 **Exploration**

Try new approaches
Discover better policies

⚖️

✅ **Exploitation**

Use proven methods
Maximize known rewards

Exploration Strategies

🎲 Entropy Bonus

🌡️ Temperature Sampling

Online vs Offline RLHF

Offline RLHF

Train on fixed datasets of preferences collected in advance. More common in medical AI due to safety requirements.

Offline Advantages

- Safety: All data reviewed before training
- Reproducibility: Fixed dataset enables consistent experiments
- Expert Efficiency: Collect preferences in batches
- Lower Risk: No live patient interaction during training

Online RLHF

Continuously collect new preferences during deployment and retrain. Riskier but more adaptive to evolving needs.

Online Advantages

- Adaptation: Quickly respond to new medical knowledge

Part 3/3:

RLHF in Clinical Practice

1. Clinical Outcome Rewards
2. Patient Satisfaction Metrics
3. Safety-Critical RLHF
4. Continuous Learning Systems
5. Feedback Loop Design
6. A/B Testing Framework
7. Performance Monitoring
8. Failure Mode Analysis
9. Case Study & Hands-On

Clinical Outcome Rewards

Outcome-Based Reward Design

Reward models can be trained not just on preferences, but on actual clinical outcomes to align AI with real-world treatment success.

Outcome Metrics

- Survival Rates: Long-term patient outcomes
- Readmission Rates: 30-day hospital readmissions
- Symptom Improvement: Patient-reported outcomes
- Complication Rates: Post-treatment adverse events
- Quality of Life: QALY, functional status measures

Implementation Challenges

- Long Feedback Loops: Outcomes take time to materialize
- Confounding Factors: Many variables affect outcomes
- Sample Size: Need large datasets for statistical power
- Ethical Considerations: Can't experiment with patient health

Patient Satisfaction Metrics

Why Patient Satisfaction Matters

Beyond clinical accuracy, AI systems should provide positive patient experiences, which correlate with treatment adherence and outcomes.

Satisfaction Dimensions

- Clarity: Was the explanation understandable?
- Empathy: Did the AI show appropriate compassion?
- Completeness: Were all questions answered?
- Trust: Did the patient feel confident in the advice?
- Efficiency: Was the interaction time-appropriate?

Collection Methods

- Post-Interaction Surveys: Rating scales, open feedback
- Implicit Signals: Time spent reading, follow-up questions
- Behavioral Data: Appointment adherence, prescription fills
- Sentiment Analysis: Analyze patient language patterns

Safety-Critical RLHF

Medical Error Prevention

RLHF systems in healthcare must have multiple safety layers to prevent errors that could harm patients.

Safety Mechanisms

- Pre-Deployment Testing: Extensive validation before clinical use
- Conservative Defaults: Err on side of caution when uncertain
- Human-in-the-Loop: Require expert confirmation for critical decisions
- Redundant Checks: Multiple independent safety verifications
- Fail-Safe Modes: Graceful degradation when errors detected

Critical Failure Modes

- Medication Errors: Wrong drug, dose, or timing
- Diagnostic Misses: Failing to identify serious conditions
- Inappropriate Reassurance: Downplaying concerning symptoms
- Scope Violations: Advising beyond system competency
- Harmful Content: Suggesting dangerous treatments

Continuous Learning Systems

Adaptive Medical AI

Continuous learning allows RLHF systems to adapt to new medical knowledge, evolving guidelines, and emerging conditions.

Learning Pipeline

- Data Collection: Continuously gather new preferences and outcomes
- Quality Control: Validate and filter incoming feedback
- Incremental Training: Periodically update models
- Validation: Test updated models before deployment
- Staged Rollout: Gradual deployment with monitoring

Update Frequency

- Critical Updates: Immediate for safety issues
- Guideline Changes: When major clinical guidelines update
- Performance Improvement: Monthly or quarterly cycles
- Seasonal Adjustments: For seasonal conditions (e.g., flu)

Feedback Loop Design

Closed-Loop Learning

Effective feedback loops ensure continuous improvement from deployment experience back into the training process.

Loop Components

- Deployment: Model serves real clinical interactions
- Monitoring: Log outputs, user interactions, outcomes
- Expert Review: Clinicians evaluate flagged cases
- Data Curation: Select valuable examples for training
- Model Update: Retrain with new feedback data
- Validation & Rollout: Test and deploy improved model

Feedback Triggers

- Low Confidence: Model uncertainty about output
- User Disagreement: User corrects or rejects output
- Adverse Events: Negative outcomes reported
- Novel Cases: Situations not seen during training
- Periodic Sampling: Random audit of routine cases

A/B Testing Framework

Experimental Design

A/B testing allows safe evaluation of RLHF model updates by comparing new versions against existing systems.

Test Setup

- Control Group: Existing model (baseline)
- Treatment Group: New RLHF model variant
- Randomization: Assign users/cases randomly to groups
- Sample Size: Calculate based on expected effect size
- Duration: Run until statistical significance achieved

Evaluation Metrics

- Primary: Clinical accuracy, safety events
- Secondary: User satisfaction, efficiency, cost
- Guardrail: Metrics that must not worsen (safety)
- Exploratory: Additional insights (e.g., bias, fairness)

Performance Monitoring

Continuous Surveillance

Real-time monitoring detects performance degradation, safety issues, and opportunities for improvement in deployed RLHF systems.

Key Performance Indicators (KPIs)

- Accuracy Metrics: Diagnostic accuracy, treatment appropriateness
- Safety Metrics: Error rates, adverse event reports
- Efficiency Metrics: Response time, query resolution rate
- User Metrics: Satisfaction scores, engagement rates
- Technical Metrics: Latency, uptime, system load

Monitoring Dashboard

- Real-Time Alerts: Immediate notification of critical issues
- Trend Visualization: Charts showing metric evolution over time
- Anomaly Detection: Automated identification of unusual patterns
- Drill-Down Analysis: Investigate specific cases or time periods
- Comparative Views: Compare across versions, time periods, demographics

Failure Mode Analysis

Understanding Failures

Systematic analysis of failure modes helps prevent future errors and improves RLHF system robustness.

Common Failure Modes

- Overconfidence: Model too certain about incorrect outputs
- Underspecification: Ambiguous queries lead to inappropriate responses
- Distribution Shift: Performance drops on out-of-distribution inputs
- Reward Hacking: Model exploits loopholes in reward function
- Bias Amplification: RLHF reinforces training data biases

Root Cause Analysis

- Data Issues: Insufficient or biased training data
- Model Limitations: Architecture can't capture necessary patterns
- Reward Misspecification: Reward doesn't capture true objectives
- Training Instability: Optimization issues during RLHF
- Deployment Mismatch: Different conditions than training

Case Study: Treatment Recommendation System

System Overview

A real-world example of RLHF applied to an AI system that recommends treatment options for chronic disease management.

Implementation Details

- Base Model: Fine-tuned medical LLM (e.g., Med-PaLM)
- Preference Data: 50,000 comparisons from 200 physicians
- Reward Model: Transformer-based classifier on treatment quality
- Policy Optimization: PPO with safety constraints
- Deployment: Staged rollout over 6 months

Results

- Accuracy: 15% improvement in treatment appropriateness
- Safety: 40% reduction in contraindication errors
- Satisfaction: 85% physician approval rating
- Efficiency: 30% reduction in time to formulate treatment plan
- Adherence: 12% increase in patient treatment adherence

Hands-On: Building an RLHF Pipeline

Practical Implementation

Step-by-step guide to implementing a basic RLHF pipeline for medical applications.

Setup & Prerequisites

- Python 3.8+, PyTorch, Transformers library
- Pre-trained medical language model (e.g., BioBERT, ClinicalBERT)
- Preference dataset (or use publicly available data)
- GPU with 16GB+ VRAM recommended

Pipeline Steps

- Step 1: Load and prepare preference dataset
- Step 2: Train reward model on preferences
- Step 3: Set up PPO or DPO training loop
- Step 4: Optimize policy with reward guidance
- Step 5: Evaluate on held-out test set
- Step 6: Analyze outputs and iterate

Ethical Considerations in Medical RLHF

Ethical Imperatives

RLHF in healthcare must navigate complex ethical terrain, balancing innovation with patient safety and equity.

Key Ethical Concerns

- Bias & Fairness: Ensure equitable performance across demographics
- Transparency: Make AI reasoning understandable to clinicians
- Accountability: Clear responsibility when errors occur
- Privacy: Protect patient data in feedback loops
- Autonomy: Preserve patient and physician decision-making
- Beneficence: Ensure AI improves outcomes for all patients

Bias Mitigation

- Diverse Annotators: Include experts from various backgrounds
- Stratified Evaluation: Test performance across demographic groups
- Fairness Metrics: Monitor for disparate impact
- Bias Audits: Regular third-party fairness assessments
- Corrective Action: Retrain on underrepresented groups

Thank you

Key Takeaways

✓ RLHF aligns AI with medical expertise through expert feedback

✓ Safety constraints are critical in healthcare applications

✓ Continuous monitoring ensures long-term performance

✓ Ethical considerations guide responsible deployment

Ho-min Park

homin.park@ghent.ac.kr

powersimmani@gmail.com