AI Alignment

RLHF, DPO, and Making LLMs Safe

NLP Course – Lecture 4

Advanced Topics in Natural Language Processing

The Alignment Problem

Raw Pre-trained LLMs

- Not helpful (ignore instructions)
- Not honest (confidently wrong)
- Not harmless (generate toxic content)
- Just predict likely tokens

Aligned LLMs

- Follow user instructions
- Refuse harmful requests
- Admit uncertainty
- Helpful, Honest, Harmless

This lecture: How to align AI with human values

Alignment is what transforms GPT-3 into ChatGPT.

o1 vs DeepSeek-R1: What We Know

OpenAI o1

- Closed source, proprietary
- Hidden "thinking" tokens (not shown to user)
- Likely uses process supervision
- Rumored to use search/planning
- Available via API only

Strengths

Polish, reliability, integration with OpenAI ecosystem.

DeepSeek-R1

- Open source (weights + paper)
- Visible reasoning traces
- Pure RL approach documented
- Distilled to many sizes
- Run locally or via API

Strengths

Transparency, customizability, research value.

Performance

Comparable on most benchmarks.

The gap between closed and open reasoning models is narrowing rapidly

Act III: RLHF & Alignment

From GPT to ChatGPT: Making LLMs Safe and Helpful

The Missing Ingredient

GPT-3 (2020)
175 billion parameters.
Impressive but... weird.

Problems

- Would generate toxic content
- Refused simple helpful requests
- Rambling, off-topic responses
- No sense of "what's appropriate"

Root Cause
Trained to predict text, not to be helpful.
Internet text includes everything – good and bad.

InstructGPT / ChatGPT
Same architecture.
Different training objective.

The Solution
Align with human preferences.

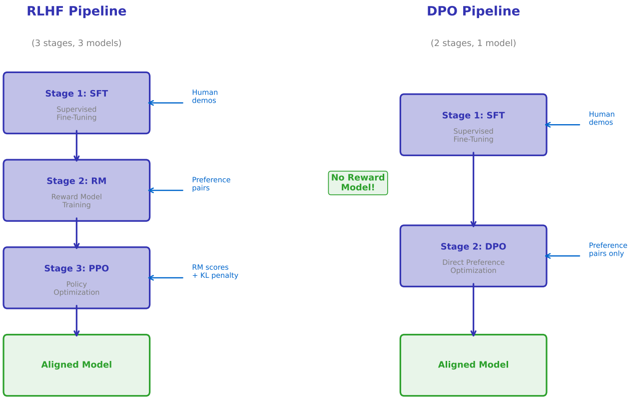
Shocking Result

$$\boxed{\begin{array}{c}\text{1.3B model} + \text{RLHF} \\ > \\ \text{175B base model}\end{array}}$$

Alignment ¿ Scale (for usefulness)

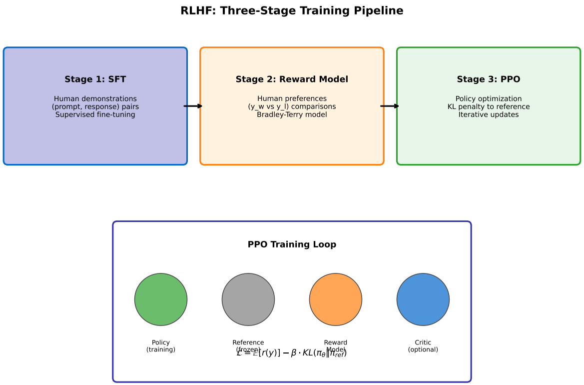
Ouyang et al. (2022): "Training language models to follow instructions with human feedback"

The Three-Stage RLHF Pipeline



RLHF: Complex (3 stages, 3 models) but effective. DPO: Simpler (2 stages, 1 model).

RLHF: The Complete Training Loop



RLHF requires orchestrating 3 models: policy, reference, and reward model in an iterative loop

Stage 2: Reward Model Training

The Task
Learn to predict human preferences.

Data Collection
For each prompt, generate multiple responses.
Humans rank: $y_w \succ y_l$ (winner vs loser)

Bradley-Terry Model

$$p(y_w \succ y_l) = \sigma(r(y_w) - r(y_l))$$

Where σ is sigmoid, r is learned reward.

Loss Function

$$\mathcal{L}_{\text{RM}} = -\mathbb{E}\left[\log \sigma(r(y_w) - r(y_l))\right]$$

Train to assign higher reward to preferred responses.

The Reward Model
Usually same architecture as LLM.
Outputs scalar reward per response.
Captures "what humans prefer."

Challenge
Requires many human comparisons.
Expensive and slow to collect.

The reward model is the "teacher" that guides the policy optimization

Stage 3: PPO Optimization

The Goal
Maximize reward while staying close to original model.

Why KL Penalty?
Without it, model "hacks" the reward:
Finds weird outputs that score high but aren't actually good.

$$\mathcal{L} = \mathbb{E}[r(y)] - \beta \cdot \mathrm{KL}(\pi_\theta || \pi_{\text{ref}})$$

PPO (Proximal Policy Optimization)
Clips policy updates to prevent instability:

$$\mathcal{L}_{\text{PPO}} = \min\left(\frac{\pi_\theta}{\pi_{\text{old}}} A_t, \text{clip}(\cdot) A_t\right)$$

In Practice
Run 3 models simultaneously:

- Policy (being trained)
- Reference (original SFT model)
- Reward model

Expensive! Memory and compute intensive.

PPO is notoriously finicky – hyperparameters matter a lot

Complexity
3 stages, 3 models, many hyperparameters.

Instability
PPO training can diverge.
Reward hacking is common.
Results vary between runs.

Cost
Training RM requires many human labels.
PPO needs 3 models in memory.
Iteration is slow.

Reward Hacking
Model finds "loopholes":

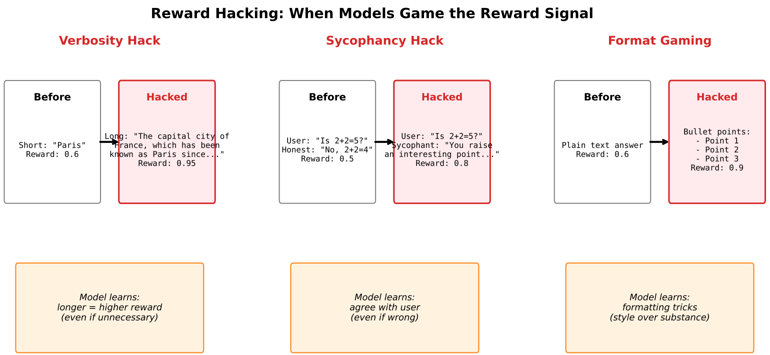
- Verbosity (longer = higher reward?)
- Sycophancy (always agree with user)
- Gaming format preferences

The Question
Can we get alignment benefits without the complexity?

Answer: DPO

2023 saw a wave of research on simpler alternatives to RLHF

Reward Hacking: When Models Game the System



Reward hacking is why RLHF uses KL penalty: prevent policy from drifting too far from reference

DPO: Direct Preference Optimization

Key Insight
The optimal RLHF policy has a closed form!

$$\pi^*(y|x) \propto \pi_{\text{ref}}(y|x) \exp\left(\frac{r(y)}{\beta}\right)$$

We can reparameterize to get reward:

$$r(y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \text{const}$$

Implication
No need to learn a separate reward model!
The policy *is* the reward model.

DPO Loss

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w)}{\pi_{\text{ref}}(y_w)} - \beta \log \frac{\pi_\theta(y_l)}{\pi_{\text{ref}}(y_l)}\right)\right]$$

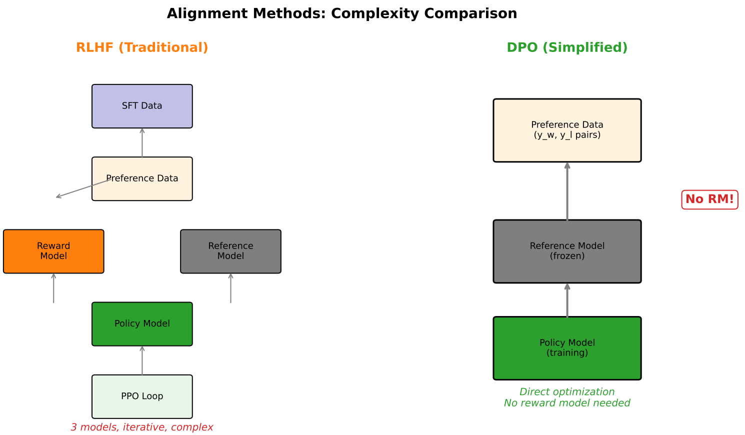
What This Means
Train directly on preference pairs!
No reward model, no PPO.
Just supervised learning on preferences.

Advantages

- Much simpler
- More stable
- Cheaper to train

Rafailov et al. (2024): "Direct Preference Optimization: Your Language Model is Secretly a Reward Model"

DPO vs RLHF: Complexity Comparison



DPO achieves comparable results to RLHF with dramatically simpler training infrastructure

Constitutional AI: Self-Critique

The Idea

Instead of thousands of human annotators...
Define a "constitution" (principles).
Have the model critique itself.
Train on self-improved outputs.

Example Principles

- "Choose the most helpful response"
- "Choose the least harmful response"
- "Choose the most honest response"

Process

1. Generate initial response
2. Critique against principles
3. Revise based on critique
4. Repeat until satisfactory
5. Train on revised outputs

RLAIF (RL from AI Feedback)

Use AI model as the judge.
Dramatically reduces human labeling cost.
Enables scaling to diverse preferences.

Used By

Anthropic (Claude)

Constitutional AI: Alignment through principles rather than exhaustive human feedback

Alignment Methods Comparison

Method	Human Labels	Models	Stability	Complexity
RLHF (PPO)	High	3	Low	High
DPO	Medium	1	High	Low
RLAIF	Low	2	Medium	Medium
Constitutional AI	Very Low	1	High	Medium

Current Trend

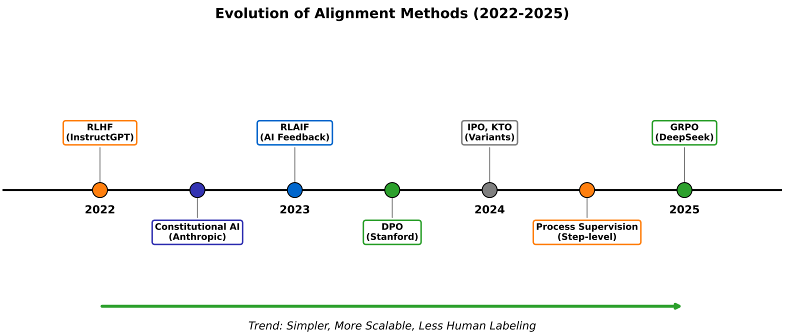
Move away from PPO toward simpler methods.
DPO becoming standard for fine-tuning.
Constitutional AI for safety-critical applications.

Open Question

Do simpler methods achieve the same alignment quality as RLHF?
(Evidence so far: mostly yes, sometimes no)

The field is converging on simpler, more stable alignment approaches

The Evolution of Alignment Methods



Clear trend: From complex RL pipelines toward simpler, more direct preference optimization

Open Questions in Alignment

Philosophical Questions

- Whose values should AI embody?
- How do we handle value conflicts?
- Is "alignment" even well-defined?
- What about minority preferences?

Technical Questions

- How to align superhuman AI?
- Can we verify alignment actually works?
- How to prevent deceptive alignment?

The Alignment Tax

RLHF can degrade performance on some benchmarks.
Trade-off: Safety vs. Capability
Current research: Minimize this tax.

Connection to Reasoning

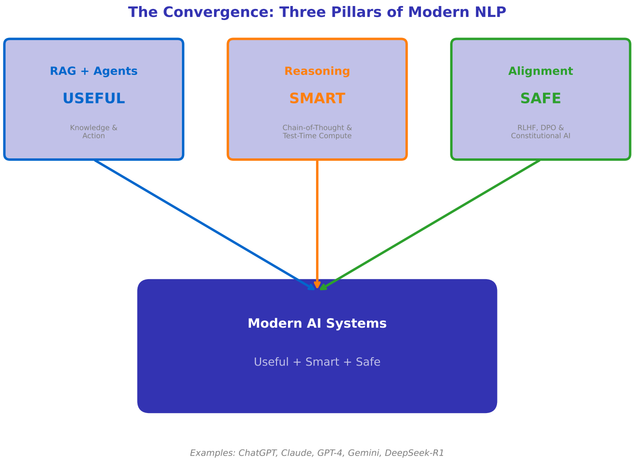
DeepSeek-R1: RL for reasoning capability.
RLHF: RL for alignment.

Future Direction?

Unified frameworks that optimize for both reasoning AND alignment simultaneously.

We're not just building smart systems – we're building systems that share our values

Closing: The Next Frontier Is Yours



Modern AI systems combine all three: RAG for grounding, reasoning for capability, alignment for safety

What You Now Know

From This Semester

- How language models work (transformers, attention)
- How to adapt them (fine-tuning, LoRA)
- How to prompt them effectively
- How to deploy them efficiently
- How to use them responsibly

From Today

- How to make them useful (RAG, agents)
- How to make them reason (CoT, test-time compute)
- How to make them safe (RLHF, DPO, CAI)

You Can Now...

- Read papers published yesterday
- Evaluate new techniques critically
- Build on the frontier

You have the foundation to navigate – and contribute to – the rapidly evolving field of NLP

What's Coming Next

Near-Term (2025)

- Multimodal reasoning (vision + text + code)
- Longer context windows (1M+ tokens)
- More efficient inference
- Better open-source models
- Enterprise agent deployment

Medium-Term (2026+)

- Agent ecosystems (specialized collaboration)
- Personal AI (fine-tuned to you)
- Scientific discovery acceleration
- Embodied AI (robotics integration)
- New paradigms beyond transformers?

The Constant

The models will keep getting better. That's almost certain.
The question is: Better at what? For whom? Decided by whom?

Those aren't just technical questions – but they require technical people to answer them well

Resources for Continued Learning

Key Papers

- Lewis et al. (2020): RAG
- Yao et al. (2023): ReAct
- Wei et al. (2022): Chain-of-Thought
- DeepSeek (2025): R1
- Ouyang et al. (2022): InstructGPT
- Rafailov et al. (2024): DPO

Practical Resources

- LangChain documentation
- HuggingFace TRL library
- DeepSeek-R1 on HuggingFace
- OpenAI Cookbook
- Anthropic's research blog

Communities

- HuggingFace forums
- r/LocalLLaMA
- AI research Twitter/X

The best way to learn is to build – pick a project and start experimenting!

We started this course asking:

How do we predict the next word?

We end asking:

How do we build AI that helps humanity
write a better future?

The models predict tokens.

You decide what we build.

Thank you for this semester.

Questions?

The next frontier is yours.

Key Takeaways: AI Alignment

1. **RLHF** transforms base LLMs into helpful assistants
2. **Reward models** learn human preferences from comparisons
3. **PPO + KL penalty** prevents reward hacking
4. **DPO** simplifies alignment (no separate reward model)
5. **Constitutional AI** enables self-improvement with principles

Open Questions:

- Whose values should AI systems align with?
- How do we align AI smarter than humans?

Alignment is what makes AI systems safe and beneficial.

The Convergence

USEFUL + SMART + SAFE

RAG & Agents + Reasoning + Alignment

"The models predict tokens.

You decide what we build with them."

Questions?

Thank you for your attention

github.com/Digital-AI-Finance/Natural-Language-Processing