CS778: FOUNDATION OF MODERN AI

Course Project

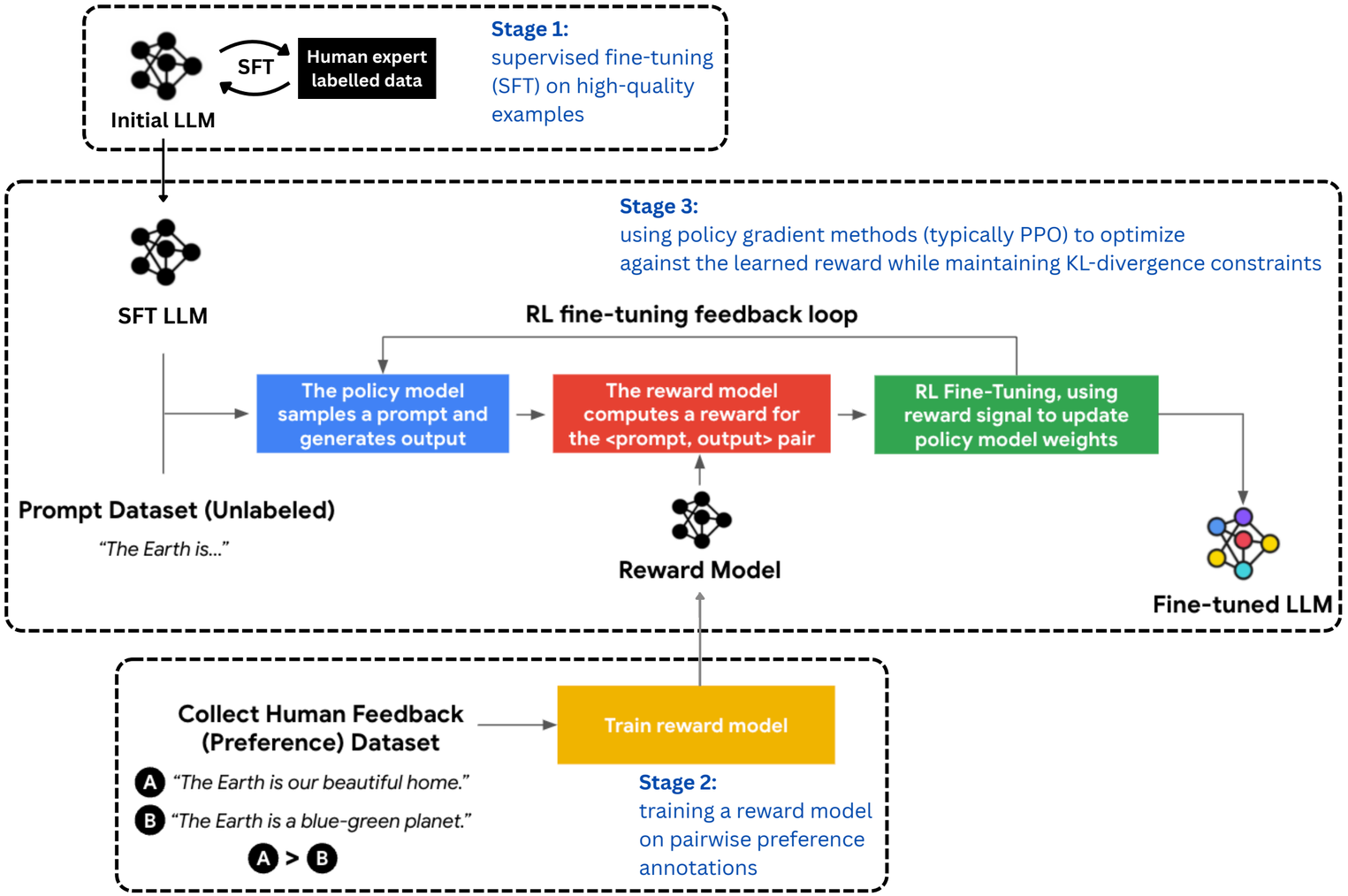
Direct Preference Optimization and Language Model Alignment

Aarsh Kaushik
220014, IIT Kanpur
aarshk22@iitk.ac.in

Keyansh Vaish
220525, IIT Kanpur
keyanshv22@iitk.ac.in

Tanmay Siddharth
221129, IIT Kanpur
tanmays22@iitk.ac.in

From RLHF to Direct Preference Optimization



The challenges with RLHF:

complex, brittle and computationally expensive

Offline and Off-policy learning

SFT πO preference data → Evolving policy π

Distribution shift degrades alignment performance and enables reward over-optimization

Reward Model Dependency

Reward Model r* — A separate discriminative model

Extract alignment signals directly from preferences without this intermediate stage

Convergence affected by sampling

Classical gradient-descent analyses do not directly address differences due to sampling in the preference optimization setting

How does the sampling distribution over response pairs affect convergence?

Review Aim and Scope

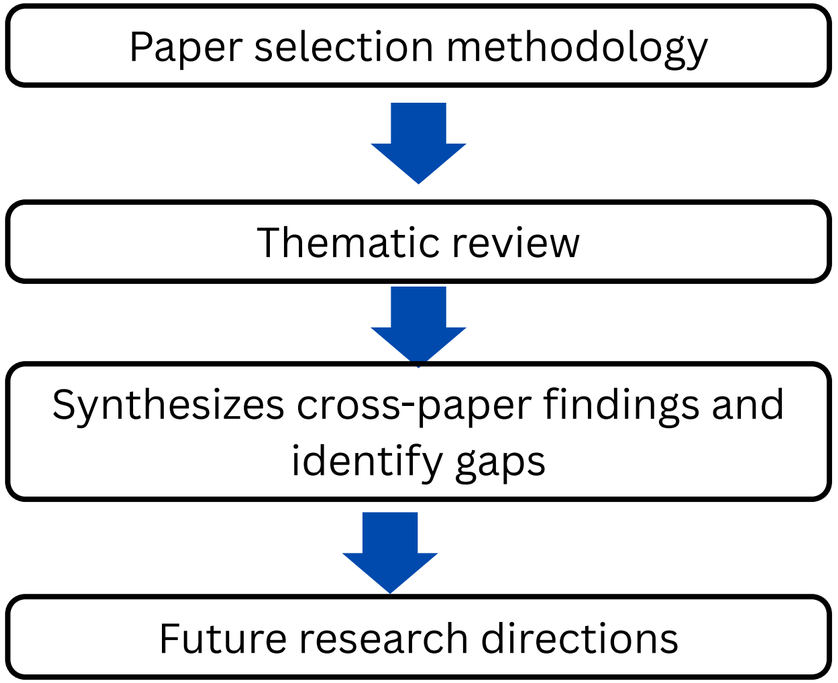
Aim

- Critically synthesizes these three papers to identify common themes, tensions, and unresolved questions
- We organize the discussion thematically; comparing theoretical framings, methods, empirical validation, and limitations

Scope

- Focus exclusively on the preference optimization paradigm for LLM alignment (DPO and its variants)
- Assume familiarity with basic concepts (policy gradient methods, KL divergence, preference models)

Roadmap of the review paper



Paper Selection

Search Strategy and Inclusion Criteria

Keywords:

"direct preference optimization", "DPO", "language model alignment", "online AI feedback", "samplers in DPO", "preference learning"

Inclusion criteria:

- Published or on arXiv between **2023–2024**
- Focus on **preference-based alignment for large language models**
- Novel methodological or theoretical contributions to **DPO**
- Empirical **validation** on standard language model benchmarks

Selected Papers

Foundational Idea:

Rafailov et al. (2023): ***Direct Preference Optimization***

RLHF as an optimization over policy parameters directly, bypassing explicit reward modeling

Practical Extension:

Guo et al. (2024): ***Direct Language Model Alignment from Online AI Feedback (OAIF)***

LLM annotator to generate preference feedback on-the-fly as the policy, thereby addressing a critical practical limitation of DPO-offline feedback

Theoretical Advancement:

Shi et al. (2024): ***The Crucial Role of Samplers in Online Direct Preference Optimization***

Optimization-theoretic analysis of DPO convergence rates under different sampling strategies

Theoretic and Comparative Review

Theoretical Frameworks and Problem Formulations

Probabilitic Preference Model

The Bradley–Terry Model

$$p(y_1 \succ y_2 \mid x) = \sigma\left(r(x, y_1) - r(x, y_2)\right)$$

RLHF Objective

$$\pi^* = \arg\max_{\pi} \; \mathbb{E}_{x\sim D,\, y\sim\pi(x)}\left[r(x,y)\right] - \beta\,\mathrm{KL}\left(\pi(x)\|\pi_{\mathrm{ref}}(x)\right)$$

DPO

Data Efficient solution of the objective

Optimal Policy

$$\pi^*(y|x) = \pi_{\mathrm{ref}}(y|x)\cdot Z(x)\cdot \exp\left(\frac{1}{\beta} r^*(x,y)\right)$$

Loss Function independent of reward model

$$L_{\mathrm{DPO}}(\pi_\theta) = \mathbb{E}_{(x,y_{\mathrm{w}},y_{\mathrm{l}})}\Big[-\log\,\sigma\left(\beta[\log\pi_\theta(y_{\mathrm{w}}|x) - \log\pi_{\mathrm{ref}}(y_{\mathrm{w}}|x)]\right.$$
$$\left. - \beta[\log\pi_\theta(y_{\mathrm{l}}|x) - \log\pi_{\mathrm{ref}}(y_{\mathrm{l}}|x)]\right)\Big]$$

Rafailov et al.

OAIF

Data regime

~~**Offline dataset {(x,y+, y-)}**~~

for t := 0 to T ***do***

1. x ~ D
2. **y1, y2 ~ πθt**
3. **LLM annotator to get preference pair y+, y-**
4. Update θ

Guo et al.

Sampler in DPO

Data Which samples are most valuable for optimization

DPO Formulation with specified sampling probability

$$L_{\mathrm{DPO}}^{\mathrm{exact}}(\theta) = \sum_{y,y'\in\mathcal{Y}} s(y,y')\cdot\rho(y,y';\theta)\cdot[-\log\sigma(\cdot)]$$

s(y, y′) is the sampling probability (uniform over *Y* or policy-dependent)

ρ quantifies violation of the preference ranking

Shi et al.

Theoretic and Comparative Review

Methods, Algorithms, and Data Regimes

----------Data Regimes: Offline vs. Online----------

Distribution shifts in offline feedback

- **Initial shift**: π0 ≠ πref
- **Gradual shift**: πt evolves during training, further diverging from the data generation distribution from π0

OAIF

By using online AI feedback, they ensure that the preference data is always generated by the current or recent policy, making training more on-policy.

----------Sampler Design----------

DPO-Unif

Uniform sampling over the action space Y

$O(\varepsilon^{-1})$ convergence

DPO-Mix-R

$s_1 \propto \exp(r(\cdot)) \quad s_2 \propto \exp(-r(\cdot))$

$O(\varepsilon^{-1/2})$ convergence

But r is unknown so impractical

DPO-Mix-P

$s_1 \propto \pi(\cdot)/\pi_{\text{ref}}(\cdot), s_2 \propto \pi_{\text{ref}}(\cdot)/\pi(\cdot)$

using the policy-difference as a proxy for reward

$O(\varepsilon^{-1/2})$ convergence

Practical Instantiation: logit mixing approximated DPO-Mix-P

$$\pi_{\text{mixed}}(y|x) \propto \lambda_1 \log \pi_\theta(y|x) + (1-\lambda_1) \log \pi_{\text{ref}}(y|x)$$

Theoretic and Comparative Review

Empirical Validation and Key Findings

Benchmarks & setup

- IMDb sentiment
- Reddit
- TLDR
- summarization,
- Anthropic HH dialogue

evaluate reward–KL trade-off, win rate, and human judgments

DPO Effectiveness vs PPO (Rafailov)

- Better reward–KL frontier
- Higher TLDR win rate (61% vs 57%)
- Improved Anthropic HH dialogue
- More robust to sampling temperature

Online AI feedback DPO vs Offline DPO (Guo)

- Higher win rates for TLDR(63.74% vs 7.69%) and helpfulness(58.60% vs 20.20%)
- Gave controllable behaviour (shorter responses with similar quality) and works even with smaller annotators.

Sampler's convergence improvement (Shi)

- New DPO-Mix samplers give faster, provably convergent optimization (quadratic vs linear)
- Explain / unify vanilla DPO, on-policy DPO, and GSHF

- **DPO generalizes at least as well as PPO on out of distribution data**
- **OAIF and DPO-Mix remain effective under annotator constraints and noisy gradients**

Theoretic and Comparative Review

Limitations and Quality Assessment

Limitation

Rafailov et al. (DPO):

- **Distribution shift** not fully addressed
- Can overfit to the **implicit reward**
- Limited evaluation scope due to **parameter size**

Guo et al. (OAIF)

- Computational cost due to **querying an LLM** at every training step
- Dependence on **annotator quality**
- **Controllability** depends on manual prompt design
- Distribution **shift in prompts**

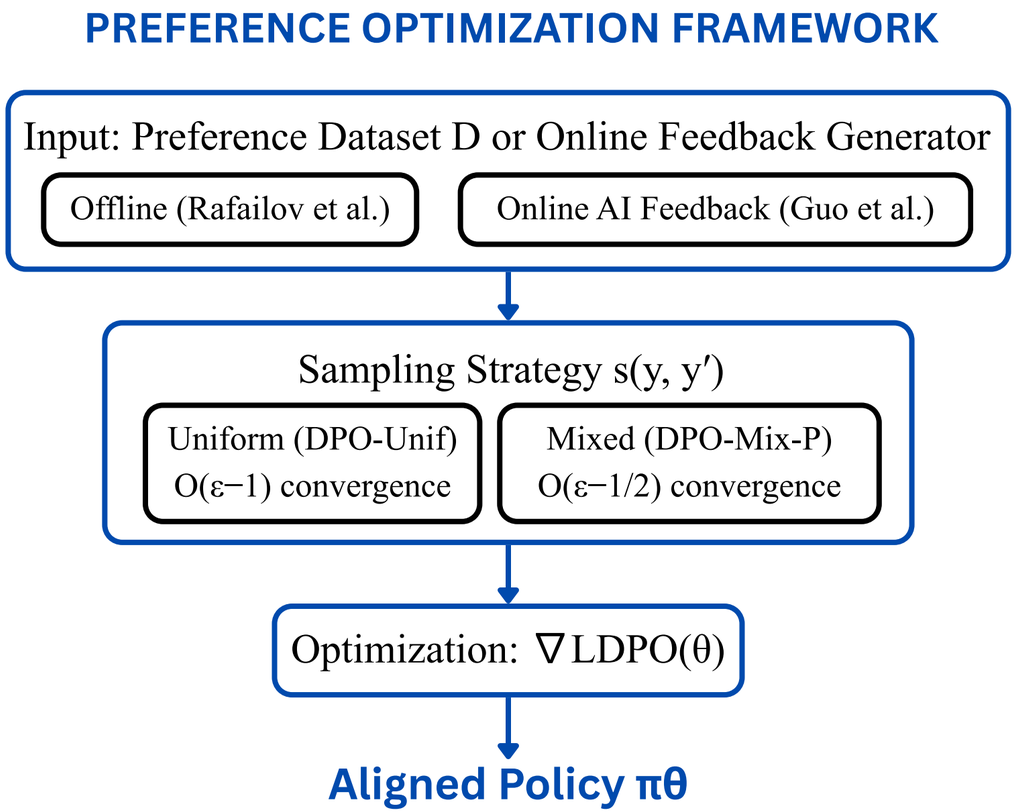
Shi et al. (Samplers in DPO)

- **Tabular** parametrization
- Exact **gradient assumption**
- **Rejection sampling** overhead adding computational cost and complexity
- **Posterior distribution** choice is not uniquely motivated

Quality Assement

Aspect	Rafailov et al. (DPO)	Guo et al. (OAIF)	Shi et al. (Samplers)
Theoretical rigor	Solid; closed-form solution elegant	Moderate; focuses on empirics	Very high; formal convergence theorems
Empirical validation	Strong; GPT-4 + human eval on 3 tasks	Very strong; extensive human eval across multiple conditions	Good; theory supported by experiments
Novelty	High; DPO is paradigm shifting	Moderate; straightforward online extension	High; new convergence insight fills gap
Reproducibility	Good; hyperparameters and baselines reported	Very good; detailed prompts, multiple seeds	Good; code promised; clear algorithms
Clarity	Excellent; well-structured paper	Good; clear motivation and algorithm	Good; theory heavy but well-explained

Integrative Framework: Feedback Regimes, Samplers, and Convergence



Quality Factors:

- Distribution Shift (offline → on-policy helps)
- Feedback Source (human → AI → LLM → controllable)
- Convergence Guarantees (theory validates practice)

Integrative Framework: Feedback Regimes, Samplers, and Convergence

Unresolved Tensions and Inconsistencies

Offline vs online DPO:

- Offline DPO still works surprisingly well despite distribution shift
- Online AI feedback gives clear gains, but **"why?"** is unresolved

Sampler theory vs practice:

Mixed samplers (DPO-Mix-P) have quadratic convergence in theory, but:

- Require an **ad-hoc mixing temperature**
- Give **only modest reward gains on real LM tasks** → unclear if the theoretical advantage really materializes

Bradley–Terry assumption:

- All papers assume BT preferences, but real **annotators show intransitivity, context/ fatigue noise, and length bias**
- The impact of this model mis-match on DPO/ OAIF is largely unexamined.

Missing Perspectives and Datasets

Narrow evaluation regime

All three works mostly test on TLDR + Anthropic HH (+ Safe-RLHF / Iterative-Prompt), with limited coverage of **instruction-following**, **math/reasoning**, **code**, **multilingual**, or **long-form generation tasks**, and weak comparisons to other direct methods

Unknown failure modes

Little analysis of where offline vs online DPO (or mixed samplers) **actually fail** or when one is systematically better than the other

Scaling uncertainty:

- Empirics are on ≤O(10B) models and ≤O(10^5) preference pairs
- behavior at 70B+ scale, 1M+ preferences, and under diverse/ conflicting preference populations is essentially unexplored.

Theory–practice gap

- Convergence results are for tabular softmax settings, **not neural network**
- We lack guarantees under function approximation and realistic training pipelines.

Implicit reward & preference model

The "implicit reward" optimized by DPO is not characterized, and robustness to violations of Bradley–Terry) is **not theoretically or empirically analyzed**

Integrative Framework: Feedback Regimes, Samplers, and Convergence

Conceptual Gaps and Future Directions

No integrated method yet

Current papers treat OAIF, DPO, and advanced samplers separately; no work combines online feedback, optimal samplers (DPO-Mix-P), and iterative annotator/sampler refinement in a single pipeline.

Opportunity: unified pipeline

Future work could:
(i) collect on-policy data via OAIF,
(ii) train with DPO + theoretically-motivated samplers, and
(iii) periodically update the annotator / sampling policy based on convergence diagnostics.

Unquantified cost–quality trade-offs:

Offline DPO is cheap but vulnerable to distribution shift; OAIF is alignment-strong but compute-heavy; mixed samplers promise faster convergence but add sampling overhead.

Open questions:

Q. *When is offline "good enough"?*
Q. *When is online feedback worth the cost?*
Q. *How does alignment quality scale with compute and sampler choice (full cost–benefit curve is missing)?*

Offline DPO: Training Dynamics and Metrics

These plots summarize training dynamics for the offline Direct Preference Optimization setup. Metrics include reward progression, logit differences, loss, accuracy, and the learning-rate schedule.

- Reward margins increase over training but with high variance.
- Logit gaps decrease, indicating weaker preference separation.
- Loss fluctuates, suggesting offline distribution mismatch.

Online DPO: Training Behavior and Performance Trends

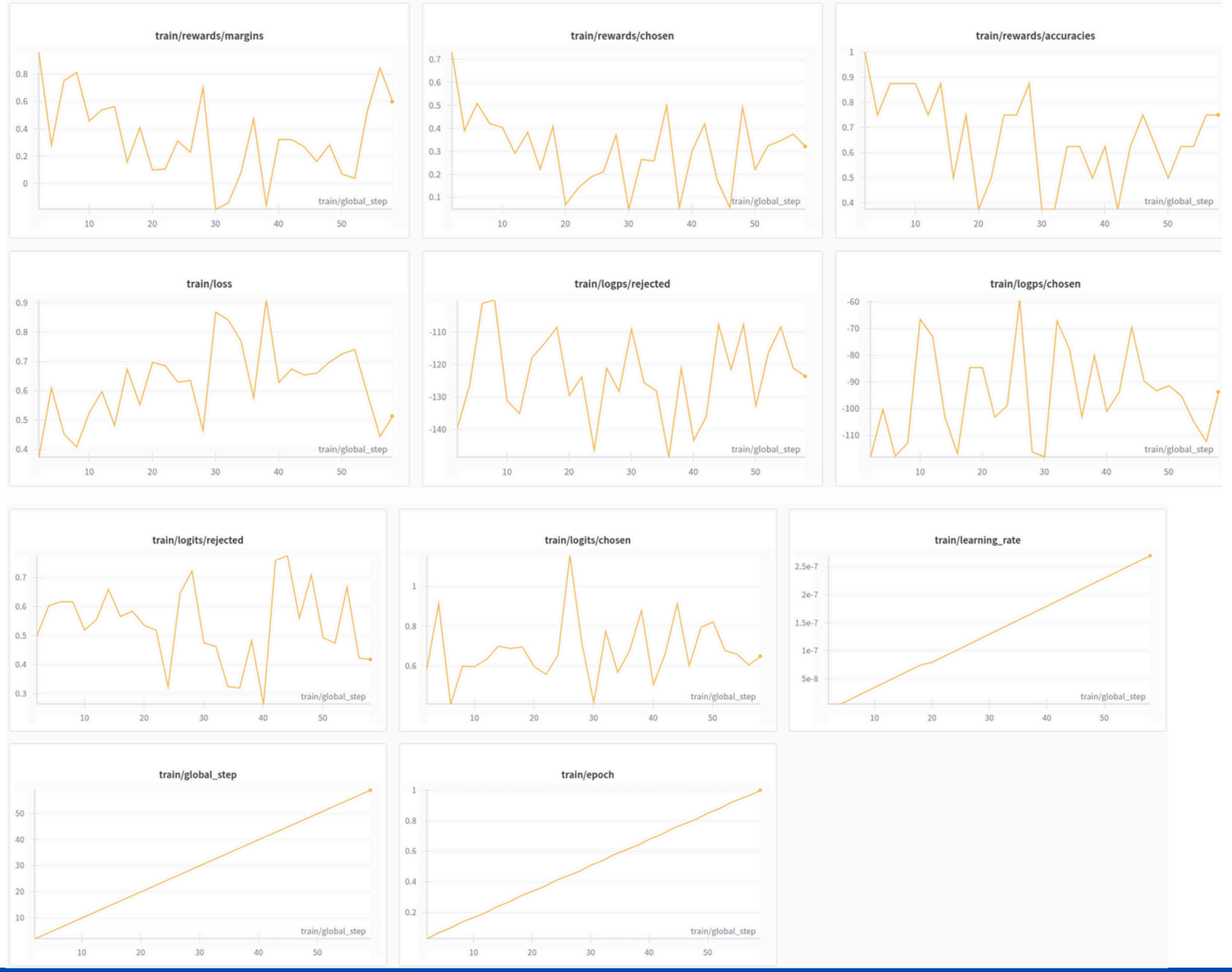
These plots illustrate training behavior for the online Direct Preference Optimization setup. New samples are generated each step, reducing distribution mismatch and providing more stable training signals.

- Reward margins increase steadily with lower variance compared to offline DPO.
- Logit differences between chosen and rejected responses show clearer separation.
- Loss trends are smoother, indicating more stable optimization from on-policy sampling.

Hybrid DPO: Mixed Training Setup and Metric Trends

This slide presents training behavior for the hybrid Direct Preference Optimization setup, where offline preference data and on-policy samples are combined. The metrics reflect reward progression, logit separation, accuracy, and loss across global steps.

- Reward margins and chosen rewards rise over time, showing effective preference strengthening.
- Logit differences for chosen vs. rejected responses remain well-separated.
- Loss trends are smoother than offline-only training and fully online training.

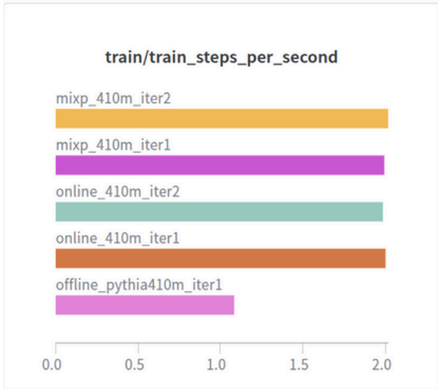


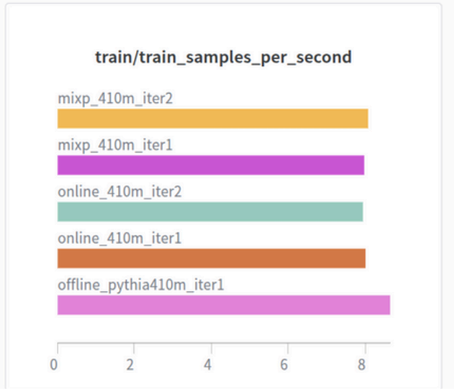
Final Results: Efficiency and Training Cost Metrics

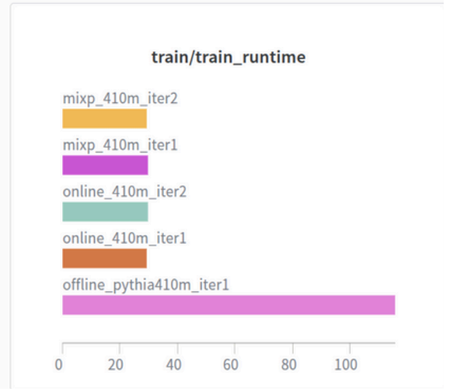
These aggregated metrics compare the computational performance of offline, online, and hybrid DPO setups. They show differences in throughput, runtime, and overall training cost across iterations.

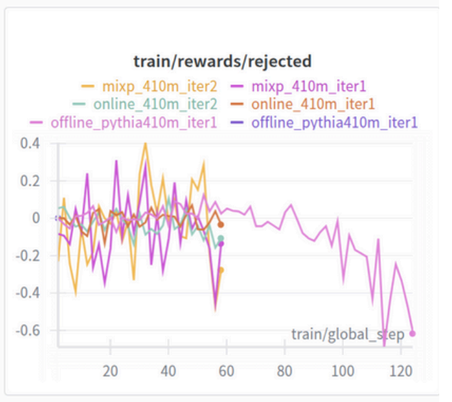
Three takeaways:

- Online DPO achieves the highest steps/s and samples/s due to continuous on-policy sampling.
- Hybrid DPO moderately trades off efficiency for better sample quality.
- Offline DPO is the slowest but most stable in runtime due to fixed datasets.









Final Results: Training Dynamics Across DPO Methods

These plots compare reward behaviors, logit separation, accuracy, and loss trajectories across offline, online, and hybrid DPO. Each method shows distinct stability and alignment patterns over global training steps.

- Online DPO exhibits fast reward margin growth but higher variance.
- Offline DPO shows stable logits but weaker reward separation due to static dataset bias.
- Hybrid DPO balances variance and stability, achieving intermediate but consistent improvements across metrics.

Thank you

You can find our work here at: https://github.com/Aarsh59/CS778-Foundations-Of-Modern-AI