MIXTURE OF EXPERTS-DIRECT PREFERENCE OPTIMIZATION: ALIGNING DIRECTLY TO HETEROGENEOUS HUMAN EXPECTATIONS

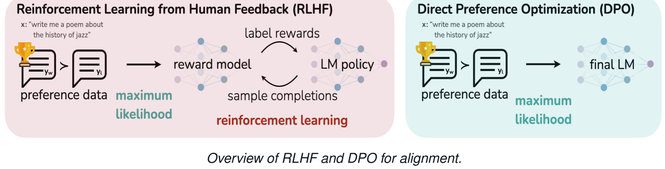
Jason Bohne[1,2], Paweł Polak[1], Brian Bloniarz[2], Gary Kazantsev [2], David Rosenberg[2]

[1]Stony Brook University [2] Bloomberg

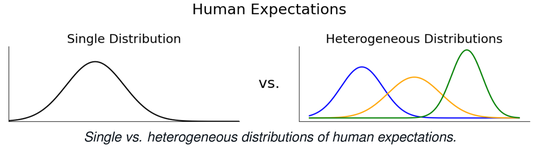
Direct Alignment Methods

Alignment is a crucial step for building safe LLMs that satisfy human expectations and values.

- Alignment with **Reinforcement Learning from Human Feedback** (RLHF) is costly due to sampling, and reinforcement learning is unstable.
- **Direct Preference Optimization** (DPO) is efficient and stable, avoiding both the reward model and the RL loop.



Overview of RLHF and DPO for alignment.



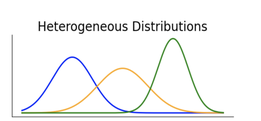
Single vs. heterogeneous distributions of human expectations.

Current direct methods are limited to a **single distribution** of human expectations, failing to capture heterogeneity across expectations, e.g., alignment to **different regulatory standards** or **investor types**.

Latent Conditional Bradley-Terry Model

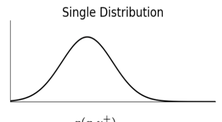
Model human preference likelihoods conditionally on a **latent class** $z = k$, offering a richer approach to modeling pairwise preferences than the classical Bradley–Terry model [2].

$$\mathbb{P}(y^+ \succ y^- \mid x, z = 1)$$
$$\mathbb{P}(y^+ \succ y^- \mid x, z = 2)$$
$$\mathbb{P}(y^+ \succ y^- \mid x, z = 3)$$



$$\mathbb{P}(y^+ \succ y^- \mid x, z = k) = \frac{e^{r_k(x,y^+)}}{e^{r_k(x,y^+)} + e^{r_k(x,y^-)}}$$

$$\mathbb{P}(y^+ \succ y^- \mid x)$$



$$\mathbb{P}(y^+ \succ y^- \mid x) = \frac{e^{r(x,y^+)}}{e^{r(x,y^+)} + e^{r(x,y^-)}}$$

Direct Alignment for our Mixture of Experts Model

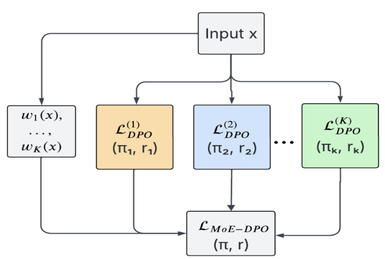
To better capture this heterogeneity, we structure our policy and reward to be a **Mixture of Experts** model.

Theorem (Informal)[a]**.** Optimal policy for k-th expert of the KL-regularized reward mixture has the closed form:

$$\pi_k^*(y \mid x) = \frac{1}{Z_k'(x)} \pi_{\mathrm{ref}(k)}(y \mid x) \exp\left(\frac{1}{\beta} r_k(x, y)\right)$$

Corollary (Informal). Using this in the latent Bradley–Terry model and aggregating the K experts, the alignment loss[b] is:

$$\mathcal{L}_{MoE-DPO}(x, y^+, y^-) \propto \sum_{k=1}^{K} w_k(x) \mathcal{L}_{DPO}^{(k)}(x, y^+, y^-)$$



Schema of our MoE Models.

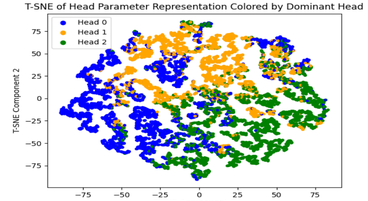
[a]Proof for the theorem on the closed form of the optimal policy can be found in [3].
[b]Here $w_k(x)$ denotes the input-dependent mixture weights, where $\sum_{k=1}^{K} w_k(x) = 1$.

Experiment: Multi-Objective Alignment

Dataset: IMDB Movie Reviews, **Model:** GPT-2, **Task:** Open-ended movie review generation.
Human Annotation: Preferences generated for ∼25,000 prompts with three reward functions: sentiment (DistilBERT), informativeness (spaCy), and grammar (spaCy).
Evaluation: On a 10,000-sample test set, the GPT-2 mixture aligned with MoE-DPO achieves **higher reward scores** than the DPO-aligned model, with clear **expert specialization** across components.

Table 1: Reward scores (Mean ± SE) for Mix-DPO on IMDb test set.

Model	Sentiment	Informativeness	Grammar
Baseline DPO	0.610 ± 0.004	0.363 ± 0.008	0.216 ± 0.001
Case 1 (Mixture)	0.654 ± 0.004	0.326 ±0.007	0.241 ± 0.001
Case 1 (Sparse)			
Head 0	0.616 ± 0.02	**0.396** ± 0.007	0.263 ± 0.001
Head 1	**0.720** ± 0.003	0.394 ± 0.008	0.213 ± 0.001
Head 2	0.632 ± 0.004	0.342± 0.007	**0.267** ± 0.001



Experiment: Multi-User Alignment

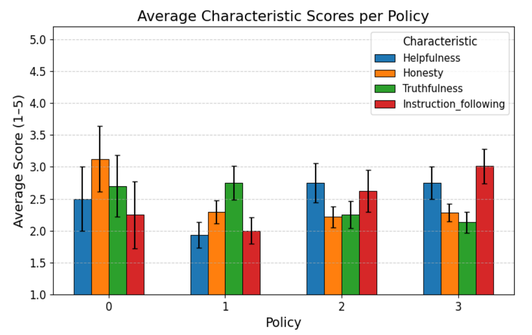
Dataset: UltraFeedback-Personalized [2], **Model:** Llama 3.2 1B, **Task:** Open-ended text generation.
Human Annotation: Preferences are generated for 10,000 prompts and response pairs with respect to four user types of helpfulness, honesty, truthfulness, and instruction-following, using GPT-4 as annotator.
Evaluation: MoE-DPO achieves **higher win rates** and **expert specialization**, as judged by GPT-4o.

Table 3: Pairwise Win Rates for MoE-DPO

Model	Evaluation	Win Rate
MoE-DPO (K=4)	Mixture — vs DPO	54.8% ± 2.7%
	Mixture — vs Reference	56.9% ± 2.8%
	Sparse — vs DPO	55.1% ± 3.2%
	Sparse — vs Reference	56.5% ± 2.8%

Note. Reference policy is the pretrained Llama 3.2 1B.

Each point represents a user encoded by a vector of reward differences which are computed with the learned MoE policy over (prompt, context response) triples.



Conclusions and Extensions

- Using a latent Bradley–Terry model, we introduce direct alignment to MoE models.
- Show improved generalization that offers interpretability through expert specialization.
- Well-suited for financial applications, such as alignment across regulatory standards or investor types.

[1] Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." Advances in neural information processing systems 2023.
[2] Poddar, Sriyash, et al. "Personalizing reinforcement learning from human feedback with variational preference learning." NeurIPS Spotlight, 2024.
[3] Bohne, Jason, Polak, Pawel, Rosenberg, David, Bloniarz, Brian, and Kazanstev, Gary. "Mix and MoE-DPO: Variational Inference Approach to Direct Preference Optimization" 2025.