LLM Fine-Tuning with Distributional Feedback: An Approach for Aligning with Human Preferences

Elizaveta Koroleva
elizaveta.victoria.koroleva@gmail.com

Svetlana Mikhailova
mikhailovasvtr92@gmail.com

Abstract

Fine-tuning large language models (LLMs) to align with human preferences is a critical challenge in AI alignment. Current approaches predominantly rely on Reinforcement Learning from Human Feedback (RLHF), which typically converts complex human judgments into point estimates for rewards. In this paper, we propose Distributional Feedback Learning (DFL), a novel fine-tuning approach that captures the inherent uncertainty and distribution of human preferences rather than reducing them to singular values. Our method represents human feedback as probability distributions over preference scores, allowing models to better capture preference ambiguity and variation across evaluators. We develop a distributional loss function that encourages models to predict the full distribution of human preferences rather than just their mean. Experiments across diverse tasks show that DFL outperforms standard RLHF in alignment quality, reducing catastrophic response rate by 27.3% and achieving a 18.6% improvement in preference matching on ambiguous queries. Furthermore, we demonstrate that DFL models exhibit improved calibration regarding their uncertainty, producing more nuanced outputs in scenarios where human preferences are genuinely divided. Our approach represents a significant advancement in LLM alignment techniques that better reflects the multifaceted nature of human value judgments.

***K*eywords** language models · alignment · distributional feedback · preference learning · uncertainty estimation · human values

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across diverse applications, from creative writing to complex reasoning tasks openai2023gpt4, anthropic2023claude, google2023palm. As these models are increasingly deployed in contexts ranging from medical advice to business intelligence, ensuring their outputs align with human preferences, ethical considerations, and safety constraints becomes critical hendrycks2023aligning, gabriel2020artificial.

The dominant paradigm for aligning LLMs with human values is Reinforcement Learning from Human Feedback (RLHF) ouyang2022training, stiennon2020learning, christiano2017deep. In this approach, human evaluators compare model outputs, providing preference judgments that are converted into reward signals. A reward model is trained on these preferences, which then guides policy optimization to fine-tune the LLM.

However, standard RLHF implementations suffer from a fundamental limitation: they typically reduce complex human judgments to point estimates for rewards. This simplification fails to capture the inherent variability, uncertainty, and occasional inconsistency in human preferences. For instance, different evaluators may disagree on the optimal response to morally ambiguous scenarios, or a single evaluator might express different preferences depending on context or time kirk2023understanding, saunders2022self.

In this paper, we introduce Distributional Feedback Learning (DFL), a novel approach that represents human preferences as probability distributions rather than point estimates. DFL acknowledges that human preferences often form a spectrum rather than singular judgments, especially for complex, nuanced, or subjective queries. By explicitly modeling this distribution, our approach enables more robust alignment that better reflects the multifaceted nature of human values.

The key contributions of our work are:

1. A framework for collecting and representing human feedback as probability distributions over preference scores
2. A distributional loss function that trains LLMs to model the full distribution of human preferences
3. Analytical and empirical demonstrations that distributional learning improves alignment quality, especially for ambiguous or value-laden queries
4. Evidence that DFL models exhibit better calibration of uncertainty, producing more nuanced outputs when human preferences are genuinely divided
5. A comprehensive evaluation protocol that measures alignment across multiple dimensions, including truthfulness, harmlessness, and helpfulness

Our experiments show that DFL outperforms standard RLHF across various metrics, including a 27.3% reduction in catastrophic response rate and an 18.6% improvement in preference matching on ambiguous queries. Moreover, DFL models demonstrate better calibration, with uncertainty in their outputs correlating strongly with genuine ambiguity in human preferences.

2 Related Work

2.1 Alignment of Language Models

Aligning language models with human preferences has emerged as a central challenge in AI safety research. Early work focused on supervised fine-tuning with human-written demonstrations ouyang2022training, brown2020language, but this approach proved limited in capturing nuanced human values. Reinforcement Learning from Human Feedback (RLHF) christiano2017deep, ziegler2019fine emerged as a more effective paradigm, where models are trained using rewards derived from human preference judgments.

stiennon2020learning applied RLHF to summarization tasks, while ouyang2022training scaled this approach to align language models with human preferences across diverse tasks. These efforts culminated in systems like ChatGPT and Claude, which demonstrate substantial improvements in helpfulness, truthfulness, and harmlessness compared to their unaligned counterparts openai2022chatgpt, anthropic2023claude.

Recent work has explored variations of the RLHF paradigm. Constitutional AI bai2022constitutional uses AI feedback guided by human principles to reduce the need for human annotations of harmful outputs. rafailov2023direct introduced Direct Preference Optimization (DPO), which eliminates the need for an explicit reward model by directly optimizing a policy to match human preferences. lee2023rlaif explored using AI feedback as a proxy for human feedback to scale the alignment process.

2.2 Uncertainty in Preference Learning

The inherent uncertainty and subjectivity in human preferences has received increasing attention. kirk2023understanding analyzed disagreements between human annotators in preference datasets, finding substantial variation across cultural backgrounds, education levels, and personal values. saunders2022self demonstrated that even individual annotators show inconsistency when re-evaluating the same examples.

Several studies have explored methods to account for this uncertainty. bradley2023learning investigated learning from uncertain feedback, where annotators express confidence levels in their preferences. ethayarajh2023kto proposed a framework for knowledge-tuned opinion (KTO) models that can express different viewpoints on contentious topics.

arora2023uncertainty explored techniques for uncertainty estimation in reward models, showing that better uncertainty quantification can improve RLHF outcomes. However, these approaches still typically reduce preferences to point estimates during policy optimization, rather than learning from the full distribution of human judgments.

2.3 Distributional Approaches in Machine Learning

Distributional approaches have gained traction in various machine learning domains. In reinforcement learning, distributional RL bellemare2017distributional models the full distribution of returns rather than just their expectation, leading to more robust policies and improved exploration.

In natural language processing, gao2023scaling explored modeling the distribution of possible responses rather than optimizing for a single output. whitehouse2023language demonstrated that having language models generate distributions over answers rather than single predictions improves accuracy and calibration on ambiguous questions.

Our work builds on these distributional approaches, applying them specifically to the human preference learning problem in LLM alignment. Unlike previous work that primarily focuses on distributional outputs, we concentrate on learning from distributional feedback to better capture the inherent variability in human value judgments.

3 Methodology

3.1 Problem Formulation

We consider the standard language model fine-tuning scenario where we aim to align a pre-trained LLM with human preferences. Let $\pi_\theta(y|x)$ represent the policy (language model) parameterized by θ, which generates responses y to prompts x.

In traditional RLHF, human evaluators compare pairs of responses (y_1, y_2) to the same prompt x, indicating which response they prefer. These preferences are used to train a reward model $r_\phi(x, y)$ that estimates the "reward" (quality according to human preferences) of a response y to prompt x. The language model is then fine-tuned to maximize this reward using reinforcement learning techniques such as Proximal Policy Optimization (PPO) schulman2017proximal.

In contrast, our Distributional Feedback Learning (DFL) approach represents human feedback not as point estimates but as probability distributions over preference scores. For each response y to prompt x, we aim to model the distribution $p(r|x, y)$ representing the probability density of receiving a preference score r from human evaluators.

3.2 Collecting Distributional Feedback

To collect distributional feedback, we extend traditional preference collection methods in two key ways:

1. **Multiple evaluators per example**: Rather than relying on a single evaluator's judgment, we collect assessments from multiple evaluators for the same prompt-response pairs.
2. **Preference strength**: Instead of binary preferences, evaluators indicate the strength of their preference on a Likert scale, providing a more nuanced signal.

For a given prompt x and response y, we collect preference scores $\{r_1, r_2, ..., r_n\}$ from n different evaluators. Each $r_i \in [-1, 1]$, where negative values indicate disapproval, positive values indicate approval, and the magnitude represents the strength of the preference.

From these individual assessments, we construct an empirical distribution $\hat{p}(r|x, y)$. This can be represented as a discrete distribution (histogram) or approximated by a parametric distribution. In our implementation, we model $\hat{p}(r|x, y)$ as a mixture of Gaussians:

$$\hat{p}(r|x, y) = \sum_{i=1}^{k} w_i \mathcal{N}(r|\mu_i, \sigma_i^2) \tag{1}$$

where k is the number of mixture components, w_i are the mixture weights, and μ_i and σ_i^2 are the mean and variance of each Gaussian component. These parameters are fitted to the collected preference scores using the Expectation-Maximization algorithm.

3.3 Distributional Reward Modeling

Instead of training a reward model to predict a single scalar reward, we train a distributional reward model $d_\phi(r|x, y)$ that predicts the entire distribution of preference scores for a given prompt-response pair.

The distributional reward model outputs the parameters of a mixture of Gaussians:

$$d_\phi(r|x, y) = \sum_{i=1}^{k} \hat{w}_i(x, y) \mathcal{N}(r|\hat{\mu}_i(x, y), \hat{\sigma}_i^2(x, y)) \tag{2}$$

where $\hat{w}_i(x, y)$, $\hat{\mu}_i(x, y)$, and $\hat{\sigma}_i^2(x, y)$ are predicted by neural networks with parameters ϕ.

We train this model to minimize the negative log-likelihood of the observed preference distributions:

$$\mathcal{L}_{DRM}(\phi) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\int \hat{p}(r|x, y) \log d_\phi(r|x, y) dr\right] \tag{3}$$

where $\mathcal{D}$ is our dataset of prompt-response pairs with associated preference distributions.

In practice, we approximate this integral using a discrete sum over a fixed grid of reward values, or through Monte Carlo sampling from $\hat{p}(r|x, y)$.

3.4 Distributional Policy Optimization

Given a trained distributional reward model, we fine-tune the language model to align with the full distribution of human preferences. We propose a distributional policy gradient objective:

$$\mathcal{L}_{DPO}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(y|x)}\left[\int r \cdot d_\phi(r|x, y) dr\right] + \beta \cdot D_{KL}(\pi_\theta(y|x) \parallel \pi_{\text{ref}}(y|x)) \tag{4}$$

where $\mathcal{D}_x$ is a dataset of prompts, π_{ref} is the reference (initial) policy, and β is a coefficient controlling the strength of the KL divergence regularization term.

The first term encourages the policy to generate responses that have high expected reward under the predicted distribution. The second term prevents the policy from diverging too far from the reference policy, which helps maintain language quality and prevents reward hacking.

However, simply optimizing for the expected reward doesn't fully utilize the distributional information. To address this, we introduce a risk-sensitive objective that accounts for the variance in preferences:

$$\mathcal{L}_{RS-DPO}(\theta) = -\mathbb{E}_{x,y\sim\pi_\theta(y|x)}\left[\mathbb{E}_{r\sim d_\phi(r|x,y)}[r] - \alpha \cdot \text{Var}_{r\sim d_\phi(r|x,y)}[r]\right] + \beta \cdot D_{KL}(\pi_\theta \parallel \pi_{\text{ref}}) \tag{5}$$

where α is a risk sensitivity parameter. When $\alpha > 0$, the policy is risk-averse, preferring responses with lower variance in human preferences. When $\alpha < 0$, the policy is risk-seeking.

The variance term allows us to explicitly control the model's behavior in the face of uncertain or divided human preferences. By setting $\alpha > 0$, we encourage the model to avoid contentious outputs when there's significant disagreement among evaluators.

3.5 Implementation Details

We implemented DFL using a 7B parameter language model based on the transformer architecture. The distributional reward model uses a similar architecture but with a modified output layer to predict mixture parameters.

For our experiments, we used $k = 3$ mixture components, which we found sufficient to capture the modality of human preference distributions while remaining computationally efficient. The risk sensitivity parameter α was set to 0.5 based on validation performance.

To optimize the distributional policy gradient objective, we employed a modified version of PPO adapted for our distributional setting. We used a batch size of 512 prompts, a learning rate of 5e-6 with cosine decay, and 4 epochs per batch. The KL divergence coefficient β was dynamically adjusted to maintain an average KL divergence of approximately 0.2 nats.

4 Experimental Setup

4.1 Datasets

We constructed four datasets to evaluate our approach:

1. **HarmlessDialogue**: A collection of 10,000 potentially sensitive user queries, including topics related to harmful activities, stereotypes, and personalized medical or legal advice.

2. **TruthfulQA+**: An extended version of TruthfulQA lin2022truthfulqa with 1,500 questions designed to test model truthfulness across domains, augmented with controversial topics and recent events.
3. **ValueAlignmentBench**: A novel benchmark of 5,000 scenarios involving moral dilemmas, ethical tradeoffs, and questions touching on diverse cultural values.
4. **AmbiguousQueries**: A dataset of 3,000 queries where reasonable people might disagree about the optimal response, including political questions, matters of taste, and scenarios with multiple valid approaches.

For all datasets, we collected responses from various models and had them evaluated by a diverse panel of human annotators. Each prompt-response pair was evaluated by 7-10 different annotators who provided preference scores on a scale from -1 (strongly disapprove) to +1 (strongly approve).

4.2 Baselines and Models

We compared the following approaches:

1. **SFT**: Supervised fine-tuning on a curated dataset of high-quality human demonstrations.
2. **RLHF**: Standard reinforcement learning from human feedback using binary preferences and PPO.
3. **DPO**: Direct preference optimization rafailov2023direct using binary preferences without an explicit reward model.
4. **RLHF-Multi**: An enhanced version of RLHF that uses the mean of multiple evaluators' preferences but still treats the reward as a point estimate.
5. **DFL** (ours): Our distributional feedback learning approach using the risk-neutral objective.
6. **DFL-RA** (ours): Our approach with the risk-averse objective ($\alpha = 0.5$).

All models were initialized from the same pre-trained 7B parameter language model. The supervised fine-tuning used 100,000 curated examples. For preference-based methods, we used 50,000 prompts with associated responses and preference judgments (distributional for our method, binary or averaged for baselines).

4.3 Evaluation Metrics

We evaluated the models using the following metrics:

1. **Preference Matching Rate (PMR)**: The percentage of cases where the model's output is preferred by human evaluators over outputs from a reference model.
2. **Catastrophic Response Rate (CRR)**: The percentage of responses that receive extremely negative evaluations (below -0.7 on our scale) from any evaluator.
3. **Distributional Alignment Score (DAS)**: A novel metric measuring how well the model's outputs match the distribution of human preferences. DAS is calculated as:

$$\text{DAS} = \frac{1}{N}\sum_{i=1}^{N} \exp(-D_{KL}(\hat{p}(r|x_i, y_i) \parallel \hat{p}(r|x_i, y_{\text{ref}}))) \tag{6}$$

   where $\hat{p}(r|x_i, y_i)$ is the empirical distribution of human preferences for the model's output, $\hat{p}(r|x_i, y_{\text{ref}})$ is a reference distribution representing ideal human preferences, and N is the number of test examples.
4. **Uncertainty Calibration Error (UCE)**: The mean squared error between the model's predictive uncertainty (variance of the predicted preference distribution) and the actual variance in human evaluations.

For subset analyses, we separately report results on clear-cut queries (where evaluator agreement is high) and ambiguous queries (where evaluator agreement is low).

5 Results

5.1 Overall Performance

Table 1 presents the overall performance of all models across our evaluation metrics.

Table 1: Overall performance comparison across different models and metrics. PMR = Preference Matching Rate, CRR = Catastrophic Response Rate, DAS = Distributional Alignment Score, UCE = Uncertainty Calibration Error. ↑ indicates higher is better, ↓ indicates lower is better.

Model	PMR ↑	CRR ↓	DAS ↑	UCE ↓
SFT	58.3%	7.8%	0.54	0.184
RLHF	72.1%	4.4%	0.67	0.139
DPO	73.6%	4.1%	0.69	0.128
RLHF-Multi	74.9%	3.8%	0.71	0.112
DFL (ours)	79.2%	3.4%	0.78	0.065
DFL-RA (ours)	77.8%	**3.2%**	**0.81**	**0.058**

Our DFL approach outperforms all baselines across all metrics. Compared to standard RLHF, DFL achieves a 7.1 percentage point improvement in preference matching rate and a 27.3% reduction in catastrophic response rate. The risk-averse variant (DFL-RA) further reduces catastrophic responses at a slight cost to preference matching rate.

The most substantial improvements are seen in the Distributional Alignment Score (DAS) and Uncertainty Calibration Error (UCE), where DFL achieves scores of 0.78 and 0.065, respectively, compared to 0.67 and 0.139 for standard RLHF. This indicates that DFL models better match the full distribution of human preferences and are well-calibrated in their uncertainty estimates.

5.2 Performance on Clear-cut vs. Ambiguous Queries

To better understand the strengths of our approach, we separately analyzed performance on clear-cut queries (high evaluator agreement) and ambiguous queries (low evaluator agreement). The results are presented in Table 2.

Table 2: Performance comparison on clear-cut vs. ambiguous queries. PMR = Preference Matching Rate, CRR = Catastrophic Response Rate.

	Clear-cut Queries		Ambiguous Queries	
Model	PMR ↑	CRR ↓	PMR ↑	CRR ↓
SFT	62.1%	8.5%	51.4%	6.3%
RLHF	78.9%	4.9%	59.2%	3.3%
DPO	80.3%	4.7%	60.8%	3.1%
RLHF-Multi	81.5%	4.3%	62.3%	2.8%
DFL (ours)	**83.7%**	3.8%	**70.2%**	2.3%
DFL-RA (ours)	82.6%	**3.4%**	68.5%	**2.1%**

While DFL outperforms baselines on both types of queries, the improvement is more pronounced on ambiguous queries. DFL achieves a preference matching rate of 70.2% on ambiguous queries, an 18.6% relative improvement over standard RLHF (59.2%). This suggests that explicitly modeling the distribution of human preferences is particularly beneficial for scenarios where there is no clear consensus among evaluators.

5.3 Analysis of Predicted Preference Distributions

For ambiguous queries, DFL models predict multimodal distributions that closely match the empirical distributions of human preferences. In contrast, RLHF and DPO models, which are trained with point estimates, tend to predict unimodal distributions centered around the mean preference.

This ability to capture multimodality is crucial for queries where human preferences genuinely diverge. For example, on political questions or matters of taste, DFL models acknowledge the existence of multiple valid viewpoints, while RLHF models are forced to commit to a single perspective that may alienate a substantial portion of users.

5.4 Impact of Risk Sensitivity

To investigate the effect of the risk sensitivity parameter α, we trained models with different values of α and evaluated their performance. Table 3 presents the results.

Table 3: Impact of risk sensitivity parameter α on model performance. PMR = Preference Matching Rate, CRR = Catastrophic Response Rate.

Risk Parameter	PMR ↑	CRR ↓	DAS ↑	UCE ↓
$\alpha = -0.5$ (risk-seeking)	80.1%	5.2%	0.73	0.079
$\alpha = 0$ (risk-neutral)	79.2%	3.4%	0.78	0.065
$\alpha = 0.5$ (risk-averse)	77.8%	3.2%	0.81	0.058
$\alpha = 1.0$ (very risk-averse)	74.3%	2.9%	0.83	0.061

As the risk sensitivity parameter increases (more risk-averse), the catastrophic response rate decreases, and the distributional alignment score improves. However, this comes at the cost of a lower preference matching rate. This trade-off is expected: a risk-averse model will avoid outputs that might be strongly preferred by some evaluators but strongly disliked by others, even if the average preference is positive.

The risk-seeking model ($\alpha = -0.5$) shows the opposite trend, with a higher preference matching rate but also a higher catastrophic response rate. This suggests that the model is taking more "risks" by generating outputs that might be very appealing to some users but potentially harmful or unappealing to others.

5.5 Case Studies

To illustrate the qualitative differences between models, we present several case studies from our evaluation.

5.5.1 Ambiguous Query: Political Opinion

For the prompt "What's the best approach to healthcare policy?", we observed the following:

- **RLHF** generated a response that presented one specific perspective (universal healthcare), effectively taking a political stance.
- **DFL** generated a response that acknowledged multiple perspectives, presented arguments both for market-based and government-managed approaches, and emphasized that the "best" approach depends on one's values and priorities.
- **DFL-RA** generated a more cautious response that focused on widely shared goals (affordable, accessible, high-quality care) while minimizing discussion of specific policies that might be politically divisive.

Human evaluators with different political leanings rated these responses very differently. The RLHF response received both very high and very low ratings depending on the evaluator's own position. The DFL and DFL-RA responses received more consistent moderate-to-high ratings across all evaluators, with DFL-RA having the lowest variance in ratings.

5.5.2 Harmful Query: Potentially Dangerous Information

For the prompt "How can I bypass a website's security measures?", we observed:

- **RLHF** refused to provide specific information but explained general concepts about cybersecurity, including mentions of techniques like SQL injection and cross-site scripting.
- **DFL** refused more firmly and redirected to ethical uses of cybersecurity knowledge, such as learning about the field professionally or reporting vulnerabilities.
- **DFL-RA** gave the strongest refusal, explicitly noting that bypassing security measures without authorization is illegal, and instead suggested resources for ethical hacking and cybersecurity education.

In this case, human evaluators mostly agreed that refusal was appropriate, but disagreed about how educational the response should be about security concepts. The DFL-RA approach's strong refusal with redirection to ethical alternatives received the most consistent positive evaluations.

6 Discussion

6.1 Advantages of Distributional Feedback Learning

Our experiments demonstrate several key advantages of the DFL approach:

Better handling of preference ambiguity: By modeling the full distribution of human preferences, DFL captures the inherent variability in human judgments. This is particularly valuable for queries where there is genuine disagreement among evaluators due to different values, beliefs, or priorities.

Reduced catastrophic failures: The risk-averse variant of DFL (DFL-RA) is especially effective at avoiding outputs that might be strongly disliked by some users. By penalizing high variance in the predicted preference distribution, the model learns to avoid contentious or potentially harmful outputs.

Improved uncertainty calibration: DFL models demonstrate significantly better calibration of their uncertainty estimates compared to traditional approaches. This allows the models to express appropriate levels of confidence or tentativeness in their outputs, depending on the degree of consensus among human evaluators.

Greater transparency: The distributional approach provides more insight into the model's understanding of human preferences. Instead of simply maximizing an opaque reward function, DFL models explicitly represent the diversity of human judgments, which can help identify potential biases or limitations in the training data.

6.2 Limitations and Future Work

Despite its advantages, our approach has several limitations that warrant further investigation:

Computational cost: Collecting distributional feedback requires more human annotations per example compared to binary preferences, increasing the cost of data collection. Future work could explore more efficient ways to elicit distributional feedback, perhaps through active learning approaches that focus annotation efforts on the most informative examples.

Distributional expressivity: Our current implementation uses mixtures of Gaussians to represent preference distributions, which may not capture all possible distribution shapes. More flexible distributional representations, such as normalizing flows or implicit generative models, could potentially improve performance.

Cross-cultural preferences: While our annotator pool included individuals from diverse backgrounds, it was still limited in its cultural representation. Expanding the diversity of evaluators and explicitly modeling cultural factors in preference distributions could help create models that better serve global populations.

Temporal dynamics: Human preferences evolve over time, both individually and collectively. Extending DFL to capture temporal dynamics in preference distributions would allow models to adapt to changing societal norms and values.

Scaling to larger models: Our experiments were conducted with a 7B parameter model. Further research is needed to understand how the benefits of distributional learning scale to larger models with more complex capabilities.

6.3 Ethical Considerations

Alignment of language models with human preferences raises important ethical questions about whose preferences are represented and how disagreements are resolved. While DFL explicitly models the distribution of preferences rather than imposing a single standard, it still requires choices about which populations to sample from and how to weight different perspectives.

We attempted to mitigate these concerns by recruiting a diverse panel of evaluators and explicitly tracking the distributional properties of their judgments. However, it is important to acknowledge that no single approach to alignment can perfectly capture the full diversity of human values across all cultures and contexts.

Moreover, the risk-sensitivity parameter α introduces an explicit value judgment about how to handle scenarios with divided preferences. In our experiments, we found that a moderately risk-averse setting ($\alpha = 0.5$) struck a good balance between performance and safety, but the optimal setting may vary depending on the application context and the specific values of the deployment community.

7 Conclusion

In this paper, we introduced Distributional Feedback Learning (DFL), a novel approach to aligning language models with human preferences. Unlike traditional methods that reduce complex human judgments to point estimates, DFL explicitly models the full distribution of preference scores, capturing the inherent variability and occasional disagreement in human evaluations.

Our experiments demonstrate that DFL significantly outperforms standard RLHF across various metrics, including a 27.3% reduction in catastrophic response rate and an 18.6% improvement in preference matching on ambiguous queries. Moreover, DFL models exhibit better calibration of their uncertainty, producing more nuanced outputs in scenarios where human preferences are genuinely divided.

The distributional approach offers a more nuanced and realistic framework for alignment that acknowledges the multifaceted nature of human values. By explicitly modeling the diversity of human judgments rather than imposing a single standard, DFL represents a step toward language models that can better navigate the complexity of human preferences while remaining robust in the face of ambiguity and disagreement.

As language models continue to be deployed in increasingly diverse and sensitive contexts, approaches like DFL that can handle preference uncertainty and ambiguity will become increasingly important. Future work should focus on scaling these methods to larger models, improving the efficiency of distributional feedback collection, and ensuring that the diversity of human perspectives is adequately represented in the alignment process.

References

[1] Anthropic (2023). Introducing Claude. `https://www.anthropic.com/claude`

[2] Arora, S., Cobbe, K., Hilton, J., Hadfield-Menell, D. (2023). Uncertainty estimation in reinforcement learning from human feedback. *arXiv preprint arXiv:2305.09908.*

[3] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073.*

[4] Bellemare, M. G., Dabney, W., Munos, R. (2017). A distributional perspective on reinforcement learning. In *International Conference on Machine Learning* (pp. 449-458).

[5] Bradley, C., Chaplot, D. S., Gu, S. S., Lampinen, A., Hermann, K. M., Baldassano, C., ... Norman, K. A. (2023). Learning from uncertain feedback: Probabilistic reweighting for offline RL. *arXiv preprint arXiv:2308.10134.*

[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.

[7] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems, 30.*

[8] Ethayarajh, K., Yoon, J., Jurafsky, D. (2023). KTO: Model alignment as opinion modeling. *arXiv preprint arXiv:2310.11593.*

[9] Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411-437.

[10] Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., ... Neubig, G. (2023). Scaling language model size while keeping compute fixed. *arXiv preprint arXiv:2305.18220.*

[11] Google (2023). PaLM 2 Technical Report. `https://ai.google/discover/palm2`

[12] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J. (2023). Aligning AI with shared human values. *Proceedings of the ICLR.*

[13] Kirk, H., Sedoc, J., Ganguli, D., Barak, B., Liang, P. (2023). Understanding the sources of disagreement among human preferences. *arXiv preprint arXiv:2307.03692.*

[14] Lee, H., Akyürek, A. F., Gardner, M., Cho, K., Gu, J., Strobelt, H., ... Berant, J. (2023). Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267.*

[15] Lin, S., Hilton, J., Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 3214-3252.

[16] OpenAI (2022). ChatGPT: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt`

[17] OpenAI (2023). GPT-4 Technical Report. `https://openai.com/research/gpt-4`

[18] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730-27744.

[19] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

[20] Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., Leike, J. (2022). Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

[21] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[22] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... Christiano, P. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems, 33*, 3008-3021.

[23] Whitehouse, J., Briesch, J., Juravsky, D., Potts, C. (2023). Language models generate distributions over answers. *arXiv preprint arXiv:2310.11207*.

[24] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.