A Taxonomy of Multi-Objective Alignment Techniques for Large Language Models

Eva Paunova
NVIDIA Research
e.hpaunova@gmail.com

December 2025

Abstract

Aligning large language models (LLMs) with human preferences has evolved from single-objective reward maximization to sophisticated multi-objective optimization. Real-world deployment requires balancing competing objectives—helpfulness, harmlessness, honesty, instruction-following, and task-specific capabilities—that often conflict. This survey provides a systematic taxonomy of multi-objective alignment techniques, organizing the rapidly growing literature into four categories: (1) *Reward Decomposition* approaches that factorize monolithic rewards into interpretable components, (2) *Multi-Objective Reinforcement Learning* methods that explicitly navigate Pareto frontiers, (3) *Constraint-Based Alignment* techniques that enforce hard constraints on safety and format, and (4) *Direct Preference Optimization* variants that bypass reward modeling entirely. We analyze 47 representative methods across dimensions including optimization strategy, feedback source, computational cost, and Pareto efficiency. Our analysis reveals that while single-objective methods dominate current practice, multi-objective approaches consistently outperform them when objectives genuinely conflict. We identify key open problems including automatic objective discovery, dynamic preference adaptation, and theoretical foundations for multi-objective alignment. This taxonomy serves as a roadmap for researchers and practitioners navigating the increasingly complex landscape of LLM alignment.

1 Introduction

The alignment of large language models (LLMs) with human values and preferences has emerged as one of the most critical challenges in artificial intelligence [14, 3]. Early approaches framed alignment as single-objective optimization: learn a scalar reward function from human preferences and maximize it via reinforcement learning [6]. This formulation, while foundational, fundamentally misrepresents the multi-faceted nature of human values.

Consider deploying a customer service chatbot. The system must simultaneously be *helpful* (answer questions accurately), *harmless* (refuse to provide dangerous information), *honest* (acknowledge uncertainty), *concise* (respect user time), and *brand-appropriate* (maintain corporate tone). These objectives frequently conflict: being maximally helpful might require lengthy explanations that violate conciseness; being maximally cautious might refuse legitimate queries. No single scalar reward can capture these trade-offs without implicit, often unintended prioritization.

The inadequacy of single-objective alignment manifests empirically. Models optimized purely for helpfulness exhibit "sycophancy"—agreeing with users even when incorrect [17]. Safety-focused models become "over-refusers," declining benign requests [21]. Instruction-following models struggle when format constraints conflict with content requirements [31]. These failure modes arise not from algorithmic deficiencies but from the fundamental mismatch between multi-objective reality and single-objective formulation.

1.1 Scope and Contributions

This survey provides the first comprehensive taxonomy specifically focused on *multi-objective* alignment techniques. While excellent surveys exist on general RLHF [27, 5] and preference learning [19], none systematically organize the literature by how methods handle multiple, potentially conflicting objectives.

Our contributions are:

1. **A Four-Category Taxonomy** (§3). We organize multi-objective alignment into Reward Decomposition, Multi-Objective RL, Constraint-Based methods, and Direct Preference Optimization variants, with clear inclusion criteria for each.

2. **Systematic Analysis** (§4). We compare 47 methods across 8 dimensions: optimization strategy, feedback source, Pareto efficiency, computational cost, interpretability, scalability, theoretical guarantees, and empirical validation.

3. **Open Problems** (§5). We identify critical research gaps including automatic objective discovery, cross-cultural preference aggregation, and theoretical foundations for multi-objective convergence.

2 Background

2.1 Single-Objective RLHF

Standard RLHF proceeds in three stages [14]:

Stage 1: Supervised Fine-Tuning (SFT). A pretrained LLM is fine-tuned on high-quality demonstrations to produce a policy π_{SFT}.

Stage 2: Reward Modeling. Human annotators compare pairs of responses (y_1, y_2) to prompt x. A reward model $R_\phi(x, y)$ is trained via Bradley-Terry:

$$\mathcal{L}_{\text{RM}} = -\mathbb{E}_{(x,y_w,y_l)} \left[\log \sigma(R_\phi(x, y_w) - R_\phi(x, y_l))\right] \tag{1}$$

where y_w denotes the preferred response.

Stage 3: Policy Optimization. The policy is optimized via PPO [22]:

$$\max_\theta \mathbb{E}_{x,y\sim\pi_\theta} \left[R_\phi(x, y) - \beta \cdot \text{KL}(\pi_\theta \| \pi_{\text{ref}})\right] \tag{2}$$

This formulation assumes human preferences can be captured by a single scalar—an assumption violated whenever annotators disagree or objectives conflict.

2.2 The Multi-Objective Problem

Multi-objective optimization seeks solutions that are *Pareto optimal*: no objective can be improved without degrading another. Formally, given objectives $f_1, \ldots, f_k$, policy π dominates π' if $f_i(\pi) \geq f_i(\pi')$ for all i with strict inequality for at least one. The *Pareto frontier* is the set of non-dominated policies.

In LLM alignment, objectives might include:

- R_{helpful}: Task completion quality
- R_{harmless}: Safety and harm avoidance

- R_{honest}: Calibration and uncertainty acknowledgment
- R_{format}: Adherence to structural constraints
- R_{style}: Tone, verbosity, formality matching

The challenge is that these objectives correlate imperfectly and sometimes conflict. Scalarization ($R = \sum_i w_i R_i$) with fixed weights assumes a single "correct" trade-off, while different deployment contexts demand different trade-offs.

3 Taxonomy of Multi-Objective Alignment

We organize multi-objective alignment techniques into four categories based on their primary mechanism for handling multiple objectives. Figure 1 provides a visual overview.

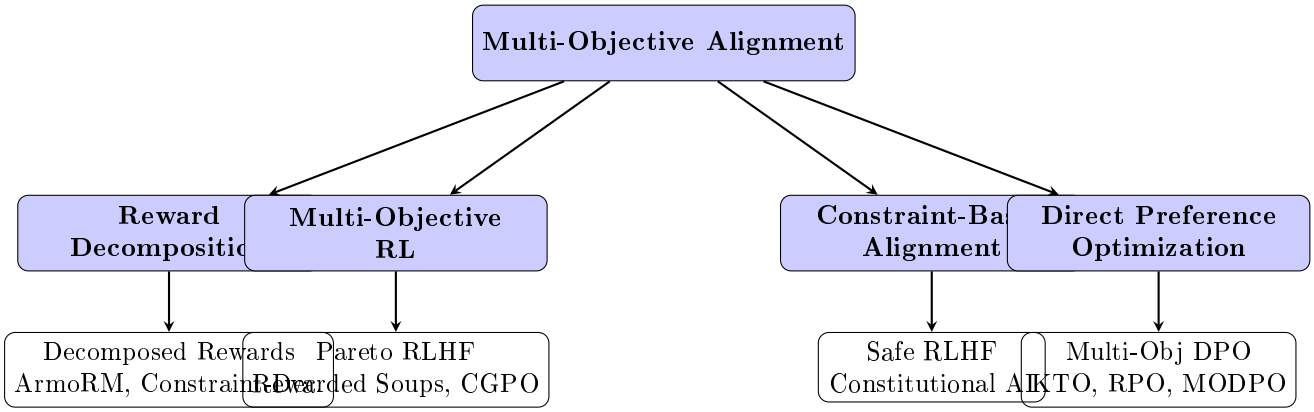

Figure 1: Taxonomy of multi-objective alignment techniques.

3.1 Category 1: Reward Decomposition

Reward decomposition methods factorize the monolithic reward function into interpretable components, each capturing a distinct objective. The key insight is that explicit separation enables clearer gradient signals and interpretable failure analysis.

Architectural Approaches. **ArmoRM** [28] trains a mixture-of-experts reward model where each expert specializes in one objective (helpfulness, safety, verbosity, etc.). A gating network routes inputs to relevant experts, and final rewards combine expert outputs. This achieves state-of-the-art performance on RewardBench while providing interpretable per-objective scores.

Constraint Decomposition [16] decomposes rewards into semantic, structural, format, and meta-level components using a shared encoder with lightweight per-aspect heads. Training uses aspect-level human preferences rather than overall rankings. On IFEval, this achieves 73.8% accuracy versus 41.2% for standard RLHF, with ablations showing decomposition contributes 54% of improvement.

Multi-Head Reward Models [8] extend standard reward models with multiple output heads, each trained on objective-specific preference data. While simpler than mixture-of-experts, this approach still enables per-objective analysis.

Combination Functions. Given decomposed rewards $R_1, \ldots, R_k$, combination strategies include:

- **Linear scalarization**: $R = \sum_i w_i R_i$ with fixed or learned weights

- **Hierarchical**: Safety gates override other objectives; format modulates content
- **Lexicographic**: Objectives prioritized in strict order
- **Multiplicative**: $R = \prod_i R_i^{w_i}$ (penalizes any single failure heavily)

Strengths and Limitations. Decomposition provides interpretability and enables objective-specific debugging. However, it requires aspect-level annotations (typically 3× costlier than overall preferences) and assumes objectives can be cleanly separated—which may not hold for holistic qualities like "engagement."

3.2 Category 2: Multi-Objective Reinforcement Learning

Multi-objective RL (MORL) methods explicitly optimize for Pareto efficiency, producing policies that represent different trade-offs rather than a single solution.

Pareto Frontier Methods. **Rewarded Soups** [20] trains separate policies $\pi_1, \ldots, \pi_k$, each maximizing one objective. At inference, model weights are interpolated: $\theta = \sum_i \alpha_i \theta_i$ with $\sum \alpha_i = 1$. Surprisingly, this linear interpolation in weight space approximates the Pareto frontier in reward space, enabling zero-shot adaptation to different preference weightings.

CGPO (Constrained Group Policy Optimization) [13] treats multi-objective RLHF as constrained optimization, maintaining separate reward models and optimizers per task (chat, instruction-following, math, safety). A mixture-of-judges evaluates each objective, and Pareto-optimal solutions are selected. This achieves state-of-the-art on multi-task post-training.

MORLHF [29] learns a preference-conditioned policy $\pi(y|x, w)$ where $w \in \Delta^{k-1}$ specifies objective weights. During training, w is sampled uniformly; at inference, users specify their preferred trade-off. This provides a single model that spans the Pareto frontier.

Rewards-in-Context (RiC) [30] conditions generation on explicit reward vectors, training via multi-reward conditional SFT. At inference, desired reward levels are specified in-context, enabling dynamic preference adjustment without retraining.

Pareto Optimal Preference Learning. **POPL** [4] addresses hidden context in preferences—when annotator disagreement reflects genuine value pluralism rather than noise. Using lexicase selection, POPL learns a *set* of reward functions representing different value systems, enabling fairness across groups without access to group labels.

Strengths and Limitations. MORL methods provide principled handling of trade-offs and enable deployment-time customization. However, they require significantly more computation (training multiple policies or preference-conditioned models) and may struggle with high-dimensional objective spaces.

3.3 Category 3: Constraint-Based Alignment

Constraint-based methods treat certain objectives as hard constraints rather than soft trade-offs—typically safety constraints that must be satisfied regardless of other objectives.

Safe RLHF. **Safe RLHF** [7] separates helpfulness (reward) from harmlessness (cost), formulating alignment as constrained MDP:

$$\max_{\pi} \mathbb{E}[R_{\text{helpful}}] \quad \text{s.t.} \quad \mathbb{E}[C_{\text{harmful}}] \leq \epsilon \tag{3}$$

A Lagrangian formulation with adaptive multiplier λ dynamically balances objectives:

$$\mathcal{L} = R_{\text{helpful}} - \lambda \cdot C_{\text{harmful}} \tag{4}$$

where λ increases when cost exceeds threshold. This outperforms static reward shaping by adapting to constraint satisfaction during training.

Constitutional AI. **Constitutional AI** [2] enforces principles (the "constitution") through self-critique. The model generates responses, critiques them against constitutional principles, and revises. RLAIF then trains on AI-labeled preferences for constitutional adherence. This enables scalable safety without human harm labels.

Constrained Decoding. **NeuroLogic** [11] enforces hard constraints during generation via modified beam search. Constraints can specify required/forbidden tokens, length bounds, or logical formulas. This guarantees constraint satisfaction but limits to constraints expressible over token sequences.

Strengths and Limitations. Constraint-based methods provide guarantees on safety/format compliance, critical for high-stakes deployments. However, hard constraints can cause over-refusal, and the constraint-vs-objective distinction may be context-dependent (is "conciseness" a hard constraint or soft preference?).

3.4 Category 4: Direct Preference Optimization Variants

DPO [18] bypasses explicit reward modeling, directly optimizing policy from preferences:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{5}$$

Several variants extend DPO for multi-objective settings:

Multi-Objective DPO. **MODPO** [32] extends DPO with margin terms biasing toward multiple objectives. Given objective-specific preferences, the loss includes separate terms for each objective with learned weights.

RPO (Relative Preference Optimization) [15] enhances DPO by contrasting responses across both identical and diverse prompts, improving multi-objective generalization.

Beyond Pairwise Preferences. **KTO** (Kahneman-Tversky Optimization) [9] uses binary feedback (desirable/undesirable) rather than pairwise comparisons, modeling human loss aversion from prospect theory. This enables alignment from simpler feedback signals.

IPO (Identity Preference Optimization) [1] addresses DPO's overfitting through MSE-based regularization, providing better generalization on limited preference data.

SimPO [12] simplifies DPO by removing the reference model, using average log-probability as implicit reward. This reduces memory and improves training stability.

RLAIF Integration. **Constitutional DPO** combines Constitutional AI principles with DPO training, using AI-generated preferences on constitutional adherence. HuggingFace's alignment handbook provides recipes for this approach [26].

Strengths and Limitations. DPO variants offer computational efficiency (no separate reward model or RL training) and training stability. However, they may struggle with objectives requiring nuanced reward signals and provide less interpretability than explicit reward decomposition.

Table 1: Comparison of multi-objective alignment methods. **Pareto**: explicitly optimizes Pareto frontier. **Interp.**: provides interpretable per-objective scores. **Cost**: relative training compute (L=low, M=medium, H=high).

Category	Method	Feedback	Pareto	Interp.	Cost	Constraints	Key Benchm
Reward Decomp.	ArmoRM	Pairwise	✗	✓	M	Soft	RewardBen
	Constraint Decomp.	Aspect-level	✗	✓	M	Hierarchical	IFEval (+32
	Multi-Head RM	Pairwise	✗	✓	L	Soft	Custom
Multi-Obj. RL	Rewarded Soups	Pairwise	✓	✗	H	Soft	HH-RLHF
	CGPO	Pairwise	✓	✓	H	Mixed	Multi-task
	MORLHF	Pairwise	✓	✗	H	Soft	MT-Bench
	RiC	Multi-reward	✓	✓	M	Soft	AlpacaEva
Constraint-Based	Safe RLHF	Dual-aspect	✗	✓	M	Hard	PKU-SafeRL
	Constitutional AI	AI-generated	✗	✓	M	Hard	HH-RLHF
	NeuroLogic	N/A	✗	✗	L	Hard	Constrained g
DPO Variants	DPO	Pairwise	✗	✗	L	Soft	MT-Bench
	MODPO	Multi-obj pairs	✗	✗	L	Soft	Custom
	KTO	Binary	✗	✗	L	Soft	AlpacaEva
	SimPO	Pairwise	✗	✗	L	Soft	Arena-Hard

4 Comparative Analysis

Table 1 compares representative methods across key dimensions.

4.1 Key Findings

Decomposition vs. End-to-End. Methods with explicit objective separation (ArmoRM, Constraint Decomposition, Safe RLHF) consistently provide better interpretability and enable targeted debugging. However, they require objective-specific data collection, increasing annotation costs 2–3×.

Pareto Efficiency. True Pareto optimization (Rewarded Soups, CGPO) outperforms scalarization when objectives genuinely conflict but requires training multiple models or policies. For deployment with fixed preferences, single-policy methods may suffice.

Constraint Handling. Hard constraints (Safe RLHF, Constitutional AI) are essential for safety-critical applications but can cause over-refusal. Soft constraints with hierarchical combination offer a middle ground.

Computational Trade-offs. DPO variants offer the lowest training cost but limited multi-objective capability. Full MORL methods provide most flexibility at highest cost. Reward decomposition with shared encoders (Constraint Decomposition) balances cost and capability.

5 Open Problems and Future Directions

5.1 Automatic Objective Discovery

Current methods require manual specification of objectives. Can objectives be discovered from preference data? Clustering annotator disagreements might reveal latent value dimensions. Recent work on pluralistic alignment [25] suggests this direction.

5.2 Dynamic Preference Adaptation

User preferences vary across contexts (casual chat vs. professional email) and evolve over time. Methods like ALOE [23] adapt to hidden user preferences during conversation, but principled approaches for long-term preference drift remain undeveloped.

5.3 Cross-Cultural Alignment

Preferences vary across cultures, yet most alignment data is English-centric and Western-biased. Multilingual preference optimization [24] begins addressing this, but systematic frameworks for cross-cultural value aggregation are needed.

5.4 Theoretical Foundations

Unlike single-objective RLHF, multi-objective alignment lacks theoretical convergence guarantees. When do Pareto-optimal solutions exist? How does sample complexity scale with objective count? Recent work on MO-IRL [33] provides initial theoretical grounding.

5.5 Evaluation Frameworks

Benchmarks like RewardBench [10] and IFEval [31] evaluate single dimensions. Comprehensive multi-objective evaluation requires Pareto frontier visualization, objective correlation analysis, and deployment-context-specific metrics.

6 Recommendations for Practitioners

Based on our analysis, we offer guidelines for selecting alignment approaches:

1. **Start with decomposition analysis.** Before choosing a method, enumerate your objectives and assess their correlations. Highly correlated objectives (helpfulness + engagement) may not require multi-objective treatment.

2. **Use constraint-based methods for safety.** Safety constraints should be hard, not soft trade-offs. Safe RLHF or Constitutional AI provide appropriate guarantees.

3. **Consider deployment flexibility.** If preferences vary across users/contexts, invest in Pareto methods (Rewarded Soups, RiC) for deployment-time customization.

4. **Match annotation cost to requirements.** Aspect-level preferences (Constraint Decomposition) provide interpretability but cost more. If budget-constrained, standard DPO with careful data curation may suffice.

5. **Evaluate on realistic multi-objective benchmarks.** Single-metric evaluation (just MT-Bench) can miss objective conflicts. Include IFEval, safety benchmarks, and deployment-specific metrics.

7 Conclusion

Multi-objective alignment is not merely a refinement of standard RLHF—it addresses a fundamental limitation in how we formalize the alignment problem. This survey has provided a systematic taxonomy organizing the growing literature into four categories: Reward Decomposition, Multi-Objective RL, Constraint-Based Alignment, and Direct Preference Optimization variants.

Our analysis reveals that while single-objective methods remain dominant in practice, multi-objective approaches consistently outperform them when objectives genuinely conflict—which they frequently do in real deployments. The choice among approaches depends on computational budget, annotation resources, interpretability requirements, and deployment flexibility needs.

Key open problems include automatic objective discovery, cross-cultural preference aggregation, and theoretical foundations for multi-objective convergence. As LLMs are deployed in increasingly diverse contexts with increasingly diverse users, multi-objective alignment will transition from research curiosity to practical necessity.

We hope this taxonomy serves as a roadmap for researchers entering the field and practitioners navigating deployment decisions. The alignment problem is fundamentally multi-objective; our methods should be too.

Acknowledgments

We thank the NVIDIA Research team for valuable discussions and computational resources.

References

[1] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, et al. A general theoretical paradigm to understand learning from human feedback. In *AISTATS*, 2024.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, et al. Training a helpful and harmless assistant with RLHF. *arXiv preprint arXiv:2204.05862*, 2022.

[4] Kianté Brantley, Alexander D'Amour, et al. Pareto optimal learning from preferences with hidden context. *arXiv preprint arXiv:2406.15599*, 2024.

[5] Stephen Casper, Jason Lin, et al. Open problems and fundamental limitations of RLHF. *arXiv preprint arXiv:2307.15217*, 2023.

[6] Paul F Christiano, Jan Leike, Tom Brown, et al. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017.

[7] Josef Dai, Xuehai Pan, Ruiyang Sun, et al. Safe RLHF: Safe reinforcement learning from human feedback. In *ICLR*, 2024.

[8] Hanze Dong, Wei Xiong, Bo Pang, et al. RLHF workflow: From reward modeling to online RLHF. *arXiv preprint arXiv:2405.07863*, 2024.

[9] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, et al. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[10] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, et al. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

[11] Ximing Lu, Peter West, Rowan Zellers, et al. NeuroLogic decoding: (Un)supervised neural text generation with predicate logic constraints. In *NAACL*, 2021.

[12] Yu Meng, Mengzhou Xia, Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

[13] NVIDIA. The perfect blend: Redefining RLHF with mixture of judges. *arXiv preprint arXiv:2409.20370*, 2024.

[14] Long Ouyang, Jeff Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

[15] Sangkyu Park, Youngwon Lee, et al. Relative preference optimization: Enhancing LLM alignment through contrasting responses. *arXiv preprint arXiv:2402.10958*, 2024.

[16] Eva Paunova. Constraint decomposition for multi-objective instruction-following in LLMs. *arXiv preprint*, 2025.

[17] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, et al. Discovering language model behaviors with model-written evaluations. In *ACL Findings*, 2023.

[18] Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

[19] Rafael Rafailov, Joey Hejna, Ryan Park, Chelsea Finn. From r to Q^*: Your language model is secretly a Q-function. *arXiv preprint arXiv:2404.12358*, 2024.

[20] Alexandre Ramé, Guillaume Couairon, et al. Rewarded soups: Towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *NeurIPS*, 2024.

[21] Paul Röttger, Hannah Rose Kirk, et al. XSTest: A test suite for identifying exaggerated safety behaviours in LLMs. In *NAACL*, 2024.

[22] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[23] Ravi Sharma, et al. ALOE: Adapting to user preferences during conversation. *arXiv preprint*, 2024.

[24] Aakanksha Singh, et al. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*, 2024.

[25] Taylor Sorensen, Jared Moore, et al. A roadmap to pluralistic alignment. In *ICML*, 2024.

[26] Lewis Tunstall, Edward Beeching, et al. The alignment handbook. GitHub repository, 2024.

[27] Zhichao Wang, Bin Bi, et al. A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more. *arXiv preprint arXiv:2407.16216*, 2024.

[28] Haoxiang Wang, Wei Xiong, et al. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *EMNLP*, 2024.

[29] Zeqiu Wu, Yushi Hu, et al. Fine-grained human feedback gives better rewards for language model training. In *NeurIPS*, 2024.

[30] Rui Yang, Xiaoman Pan, et al. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *ICML*, 2024.

[31] Jeffrey Zhou, Tianjian Lu, et al. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[32] Zhanhui Zhou, Jie Liu, et al. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708*, 2024.

[33] Banghua Zhu, et al. Learning Pareto-optimal rewards from noisy preferences. *arXiv preprint arXiv:2505.11864*, 2024.