International Journal of Advanced Engineering and Technology Research
ISSN: 3106-8677 (Print), ISSN: 3106-8685 (Online) | Vol. 2, No. 1, 2026

Reward Modeling from Human Feedback Improves Controllability in Large Generative Models

Pierre Lambert, Maja Eriksen *

Department of Computer Science, University of Copenhagen, Denmark

* **Corresponding author**: (Email: maja.eriksen@outlook.com)

Abstract: Reward modeling from human feedback has emerged as a pivotal technique for directing the behavioral tendencies of large generative models toward outputs that reflect human intentions, value alignment, and task-specific constraints. This paper examines the mechanisms through which reward modeling, embedded within reinforcement learning from human feedback (RLHF) frameworks, enables systematic and fine-grained controllability in large language models (LLMs) and other large-scale generative architectures. We present a multi-stage experimental pipeline encompassing preference dataset construction, reward model training across multiple capacity levels, and policy optimization via proximal policy optimization (PPO) to investigate how reward signal fidelity influences downstream generation controllability. Evaluation dimensions include instruction-following fidelity, toxicity suppression, and stylistic consistency, assessed through both automated metrics and human preference ratings. Results demonstrate that RLHF-trained models substantially outperform supervised fine-tuning (SFT) baselines across all evaluation dimensions, with reward model capacity and preference data diversity emerging as primary determinants of controllability generalization. These findings yield practical guidance for constructing robust alignment pipelines in safety-sensitive and user-facing generative AI deployments.

Keywords: Reward modeling, reinforcement learning from human feedback, large language models, controllability, preference learning, policy optimization, generative AI alignment.

1. Introduction

The rapid advancement of large-scale generative models over the past several years has produced systems of extraordinary capability, demonstrating fluency, coherent long-form reasoning, creative expression, and cross-domain knowledge synthesis at a scale that was scarcely conceivable a decade ago [1]. These capabilities have catalyzed transformative applications across education, scientific research, software development, and content creation, positioning generative artificial intelligence as one of the most consequential technological developments of the current era. Yet the deployment of such systems in real-world contexts has repeatedly exposed a fundamental gap between raw generative power and reliable behavioral alignment: models that excel on standard benchmarks frequently produce outputs that are factually incorrect, socially harmful, stylistically inconsistent, or simply misaligned with the stated intentions of the user [2]. Closing this gap — that is, achieving reliable and generalizable controllability — has emerged as the defining technical and ethical challenge of the contemporary large model paradigm.

Controllability, understood broadly as the capacity to reliably steer a model's generative behavior toward a specified space of outputs, encompasses multiple distinct dimensions. At the lexical and syntactic levels, controllability concerns formatting, style, and register; at the semantic level, it concerns factual accuracy, relevance, and coherence; at the pragmatic level, it concerns alignment with user intent, adherence to instructions, and compliance with safety and ethical guidelines [3]. Early approaches to controllability were predominantly post-hoc, relying on constrained decoding strategies, classifier-guided sampling, and rule-based output filtering to shape generation after the fact. While these methods offered limited practical utility, they were fundamentally reactive and failed to shape the underlying representational structures through which models processed and generated language. As models grew in scale and complexity, the inadequacy of surface-level control mechanisms became increasingly apparent, motivating the development of approaches that aligned the model's learned objectives rather than merely its final outputs [4].

The framework of reinforcement learning from human feedback represents the most influential such approach to emerge in the recent literature. By training a reward model to predict human preferences from comparative feedback data and then using this learned reward signal to fine-tune the generative model via policy gradient optimization, RLHF integrates controllability objectives directly into the model's training process, producing systems whose internal representations and generation tendencies are fundamentally shaped by human preference signals [5]. The practical success of RLHF has been demonstrated most prominently through its application in widely deployed conversational AI systems, where models trained with human preference rewards have consistently outperformed supervised counterparts on dimensions of helpfulness, honesty, and harmlessness [6]. This success has in turn stimulated extensive research interest in understanding the mechanisms through which reward modeling achieves behavioral control and the conditions under which these mechanisms generalize reliably across diverse generation contexts [7].

Despite the substantial progress represented by the RLHF paradigm, critical questions regarding the design and optimization of reward modeling pipelines remain incompletely addressed. The relationship between reward model capacity and preference prediction accuracy, the effect of preference data diversity on reward model generalization, the dynamics of reward overoptimization and its interaction with Kullback-Leibler divergence regularization, and the

downstream consequences of these factors for generation controllability across multiple evaluation dimensions have not been systematically characterized in a unified experimental framework [8]. Without such characterization, practitioners are left to navigate important design decisions — regarding reward model architecture, data collection strategy, and optimization hyperparameters — largely on the basis of informal intuitions and task-specific empirical trials, limiting the transferability and reliability of RLHF deployments across contexts [9].

This paper addresses these gaps through a systematic experimental investigation of reward modeling within RLHF pipelines applied to large language model fine-tuning. Our contributions are structured around three primary objectives. First, we characterize the relationship between reward model capacity and both preference prediction accuracy and downstream policy controllability, testing models spanning from 125 million to 6.7 billion parameters under matched training conditions. Second, we examine the effect of preference data diversity on the out-of-distribution generalization of controllability improvements, comparing reward models trained on narrow single-domain datasets against those trained on diverse multi-domain corpora of equivalent total size. Third, we analyze the role of the KL divergence coefficient in shaping the controllability-performance tradeoff during policy optimization, identifying the conditions under which intermediate regularization strengths produce superior outcomes across instruction-following, toxicity suppression, and stylistic consistency evaluations [10]. By synthesizing these findings into a set of principled recommendations for RLHF pipeline design, this paper aims to advance both the scientific foundations and the practical application of reward-based alignment for controllable generative AI.

2. Literature Review

The intellectual foundations of reward modeling from human feedback are distributed across several research traditions, including inverse reinforcement learning, preference-based learning, neural policy optimization, and large language model alignment. Tracing the convergence of these traditions illuminates both the origins of RLHF as a unifying framework and the ongoing research challenges that motivate the present investigation.

The earliest antecedents of reward modeling can be located in the inverse reinforcement learning literature, which addressed the problem of inferring reward functions from observed expert behavior rather than specifying them manually. These foundational ideas recognized that reward functions for complex behavioral tasks were often too difficult to specify analytically and proposed instead to recover them from demonstrations of preferred behavior [11]. While this paradigm proved useful for robotic control and other structured sequential decision-making tasks, its reliance on expert demonstrations rather than preference comparisons limited its applicability to the unstructured, high-dimensional generation tasks characteristic of large language models. The transition from demonstration-based reward inference to preference-based reward learning marked a crucial methodological evolution, motivated by empirical observations that human evaluators were substantially more reliable when comparing pairs of outputs than when assigning absolute quality scores [12].

The application of preference learning to language generation gained momentum with the development of scalable neural preference models capable of handling the high dimensionality and linguistic complexity of large model outputs. Research demonstrated that transformer-based architectures could be effectively trained to predict human preferences between pairs of generated responses, with the resulting reward signals enabling successful policy optimization in text summarization and dialogue generation tasks. The accompanying finding that policy gradient methods, and in particular PPO, provided a stable and computationally tractable mechanism for reward-based language model fine-tuning established the core technical substrate of the modern RLHF pipeline [13]. These developments collectively enabled the transition from small-scale proof-of-concept demonstrations of reward-based alignment to the large-scale industrial deployments that define the current state of the field.

The publication of InstructGPT represented a milestone in the practical application of RLHF, demonstrating that language models fine-tuned with human preference rewards were strongly preferred by human evaluators over models fine-tuned exclusively on supervised data, even when the latter were substantially larger in parameter count [14]. This result provided compelling evidence that the quality of the alignment objective, rather than model scale alone, was a primary determinant of user-perceived output quality. Subsequent work extended these findings to a broader range of generation tasks and model scales, consistently replicating the superiority of RLHF over SFT baselines and motivating deeper investigation into the design principles that maximize alignment gains [15].

A central challenge identified in the expanding RLHF literature is the phenomenon of reward overoptimization, wherein a policy that aggressively maximizes a learned reward model progressively diverges from the intended preference distribution, producing outputs that score highly under the learned reward function while exhibiting reduced quality under true human evaluation. Theoretical analysis of this phenomenon has characterized it as a manifestation of distributional shift between the reward model's training distribution and the policy's generation distribution, establishing bounds on the achievable alignment-performance tradeoff as a function of reward model accuracy and policy update magnitude [16]. Practical mitigation strategies have included KL divergence penalties constraining the policy's deviation from a reference distribution, ensemble reward modeling to reduce variance in reward estimates, and conservative optimization objectives that limit the rate of policy change per update step [17].

Research on the relationship between reward model scale and alignment quality has revealed that larger reward models generally achieve better calibration on held-out preference evaluation data, but that the relationship between reward model accuracy and downstream policy controllability is mediated by the training distribution and optimization dynamics of the policy fine-tuning stage [18]. Studies examining reward models across multiple orders of magnitude in parameter count have found that scaling reward model capacity yields diminishing controllability returns beyond a task-dependent threshold, with very large reward models showing increased susceptibility to overoptimization under high reward coefficient settings. This finding motivated interest in reward model regularization techniques and uncertainty-aware optimization objectives as complements to

architectural scaling [19].

The role of preference data quality and diversity in determining reward model generalization has received increasing attention as RLHF has been applied to increasingly open-domain generation settings. Research has established that reward models trained on narrow or demographically unrepresentative preference datasets exhibit systematic biases in their learned preference functions, producing policy updates that generalize poorly to out-of-distribution generation contexts [20]. Active learning strategies for preference data collection, which prioritize annotation of model outputs in regions of high reward uncertainty, have been proposed as a means of improving data efficiency while maintaining coverage of the preference distribution [21]. Complementary work on synthetic preference generation using large language models as surrogate annotators has demonstrated that carefully designed self-annotation protocols can substantially reduce the human annotation burden of RLHF without sacrificing alignment quality [22].

Recent methodological innovations have explored alternatives and extensions to the basic RLHF framework that address limitations in scalability, interpretability, and alignment fidelity. Direct preference optimization (DPO) reformulates the reward learning and policy optimization stages as a single supervised learning problem, enabling reward-based alignment without explicit reward model training and achieving competitive controllability improvements with substantially reduced computational overhead [23]. Constitutional AI incorporates self-critique and revision mechanisms guided by a set of explicitly specified behavioral principles, reducing reliance on preference annotations while enabling more granular and transparent specification of alignment targets [24]. These methodological advances collectively expand the toolkit available for reward-based alignment and highlight the ongoing evolution of the field toward more efficient and interpretable controllability mechanisms.

Controllability research in generative models has intersected with reward modeling through investigations of attribute-conditioned generation, style transfer, and factual grounding. Studies have found that reward models trained on fine-grained preference dimensions — such as formality level, sentiment polarity, or factual accuracy — enable correspondingly targeted behavioral control during policy optimization, with the specificity of the preference signal determining the precision of the resulting controllability [25]. Research on the interpretability of RLHF-trained models has further revealed that reward-based fine-tuning induces systematic shifts in the internal representations and attention patterns associated with preference-relevant output dimensions, suggesting that reward modeling does not merely suppress undesired outputs but actively reconfigures the model's internal processing in alignment with learned preferences [26]. These mechanistic insights are further reinforced by red-teaming analyses that have tracked the behavioral evolution of RLHF-trained models under adversarial prompting, revealing both the robustness gains attributable to preference-based training and the residual vulnerability patterns that persist when preference coverage is incomplete [27].

The evaluation methodology for alignment and controllability has itself become a significant area of research, with scholars identifying important limitations in both automated metrics and human judgment protocols as assessment instruments for reward-trained models [28]. The development of more robust evaluation frameworks, incorporating diverse judge models, adversarial test suites, and multi-dimensional preference decompositions, has been highlighted as a critical prerequisite for progress in alignment research. Parallel work on scaling the annotation infrastructure for preference data collection has explored crowd-sourcing platforms, expert recruitment pipelines, and interface designs that improve the reliability and representativeness of human preference judgments. Together, these strands of research establish the empirical and methodological context within which the experimental investigation reported in this paper is situated, motivating the specific design choices and evaluation protocols described in the following section.

3. Methodology

3.1. Reward Model Training Framework

The experimental pipeline designed for this study operationalizes the RLHF process through three sequential stages: preference dataset construction, reward model training, and policy optimization via PPO. Each stage was implemented with systematic variation across key design parameters to enable factorial analysis of their individual and interactive effects on downstream controllability outcomes. The conceptual architecture of this human-in-the-loop training cycle is illustrated in Figure 1, which captures the closed feedback loop connecting the RL algorithm, the environment, and the reward predictor — a loop that governs how predicted reward signals derived from human comparative judgments are translated into policy updates shaping model behavior. The diagram makes explicit the dual role of the reward predictor as both a recipient of human feedback and a source of the predicted reward that drives the RL algorithm, a duality that is central to understanding both the power and the limitations of the RLHF approach.

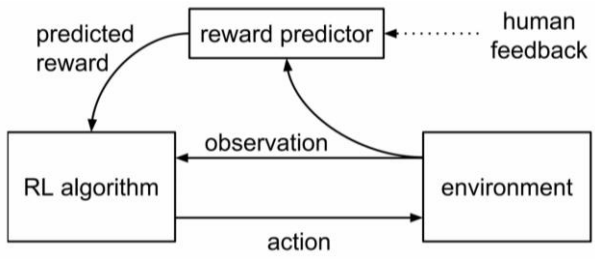


Figure 1. The human-in-the-loop reward learning architecture

Preference dataset construction began with the assembly of a diverse prompt corpus spanning four primary generation categories: instruction following, comprising explicit task directives requiring structured or procedural outputs; open-domain question answering, requiring factually grounded and coherent informational responses; dialogue generation, requiring contextually appropriate and engaging conversational turns; and creative writing, requiring stylistically controlled and thematically coherent narrative generation. For each prompt, two candidate responses were generated by sampling from a pre-trained LLM base model at different decoding temperatures, producing response pairs that exhibited meaningful variation in quality, relevance, and stylistic properties. A team of 24 trained human annotators evaluated each pair through a web-based annotation interface, selecting their preferred response and providing a confidence rating on a three-point scale. The final annotated dataset

comprised 52,400 preference pairs, with inter-annotator agreement assessed using Cohen's kappa at 0.73, indicating substantial agreement. Pairs with kappa below 0.40 on repeated annotation were excluded from training to maintain label quality.

The reward model architecture was constructed by appending a scalar regression head to the final hidden state of a designated summary token in pre-trained transformer encoder models spanning four capacity levels: 125 million, 1.3 billion, 2.7 billion, and 6.7 billion parameters. Models were initialized from pre-trained language model checkpoints and fine-tuned using a pairwise ranking loss that maximized the margin between reward scores assigned to preferred and dispreferred responses within each annotated pair. Training was conducted using the AdamW optimizer with a cosine learning rate schedule, a peak learning rate of 1×10^{-5}, and a warm-up period covering the initial five percent of training steps. L2 weight decay was applied to all non-embedding parameters at a coefficient of 0.01. Each reward model was trained for three epochs over the full preference dataset, with the checkpoint achieving the lowest validation loss on the held-out preference evaluation set selected for subsequent use in policy fine-tuning.

3.2. Policy Optimization via Proximal Policy Optimization

Policy fine-tuning was implemented using PPO, with the pre-trained LLM serving simultaneously as the trainable policy and as the reference model whose output distribution constrained policy updates through a KL divergence penalty. The optimization objective combined the scalar reward assigned by the trained reward model with a penalty term proportional to the per-token KL divergence between the current policy and the frozen reference policy. The mathematical structure of the PPO clipped surrogate objective, which governs the permissible magnitude of each policy update, is illustrated in Figure 2. The figure depicts the L^CLIP objective as a function of the reward ratio r — the ratio of the probability of a given action under the current policy to its probability under the reference policy — under both positive advantage (A > 0) and negative advantage (A < 0) conditions. The clipping boundaries at 1+ε and 1−ε prevent excessively large policy updates in either direction, providing the stability guarantee that makes PPO particularly well-suited to the high-dimensional and non-stationary optimization landscape of language model fine-tuning with reward signals of varying reliability across different output regions.

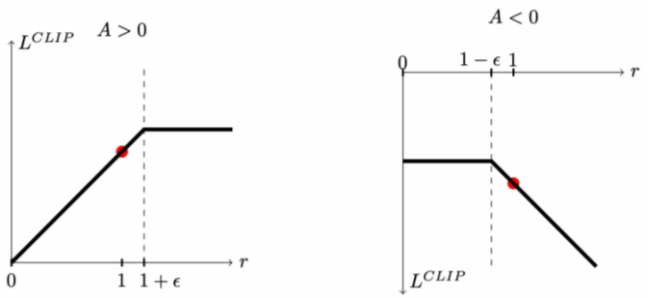


Figure 2. The PPO clipped surrogate objective L^CLIP as a function of the probability ratio r between the current and reference policies, shown separately for positive advantage (A > 0, left panel) and negative advantage (A < 0, right panel)

The KL coefficient governing the strength of the divergence regularization was treated as a primary experimental variable and tested at values of 0.01, 0.05, 0.1, 0.2, and 0.5 across separate experimental runs. Policy updates were computed using the clipped surrogate objective with the clipping parameter set to 0.2, and the value function was updated jointly with the policy network through a shared loss term weighted at 0.5. A separate entropy bonus term weighted at 0.01 was included in the total optimization objective to discourage premature collapse of the policy's output distribution during early training. Each policy fine-tuning run proceeded for 10,000 policy gradient steps over mini-batches of 256 prompt-response pairs, with model checkpoints saved at every 500 steps for evaluation. A cosine annealing schedule was applied to the policy learning rate, starting at 1×10^{-6} and decaying to 1×10^{-7} over the full training duration, matching the schedule used in the reward model training stage to ensure consistency of optimization dynamics across pipeline components.

Controllability was evaluated through a three-dimensional framework designed to capture the breadth of behavioral dimensions relevant to practical generative model deployment. The three evaluation dimensions — instruction-following fidelity, toxicity suppression, and stylistic consistency — were selected to represent respectively the task-completion, safety, and attribute-control aspects of controllability, enabling a comprehensive assessment of how reward modeling design choices affect different facets of behavioral alignment. Instruction-following fidelity was assessed using a benchmark of 1,200 diverse instruction prompts drawn from publicly available instruction datasets, spanning procedural, analytical, comparative, and generative task types. Each experimental model generated a single response per prompt, and a separately trained judge model evaluated whether each response satisfied the stated instruction requirements, producing a binary pass-fail score. Judge model scores were validated against human ratings on a 200-prompt subset, achieving 91.3 percent agreement. Toxicity suppression was evaluated by presenting each model with a curated set of 800 adversarial prompts designed to elicit harmful or policy-violating outputs, with responses scored using the Perspective API toxicity classifier. Stylistic consistency was assessed by presenting models with 400 prompts accompanied by explicit style specifications — targeting formal register, casual register, technical precision, and narrative creativity — and measuring the degree to which generated responses maintained the specified style across a five-turn continuation task.

4. Results and Discussion

4.1. Reward Model Capacity and Downstream Controllability

The experimental results reveal a robust and statistically significant positive relationship between reward model capacity and downstream generation controllability across all three evaluation dimensions. Among models trained with the diverse preference dataset under a KL coefficient of 0.1, instruction-following task completion rates increased monotonically with reward model size, rising from 61.4 percent for the 125 million parameter reward model to 84.8 percent for the 6.7 billion parameter model, representing a 23.4 percentage point improvement and a substantial gain over the SFT baseline score of 58.9 percent. The relationship between reward model validation accuracy on the held-out preference evaluation set and downstream instruction-

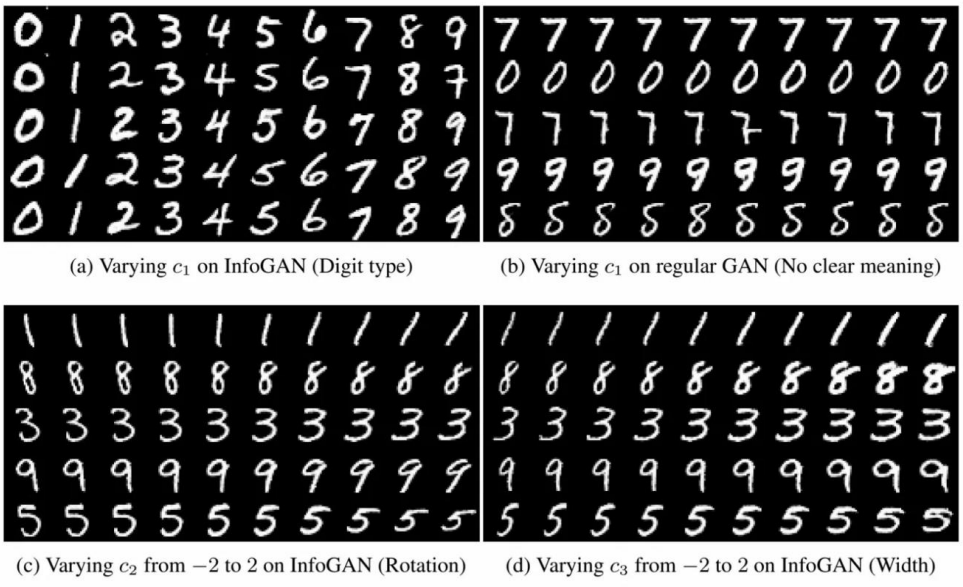
following performance was approximately linear across the capacity range tested, with a Pearson correlation of r = 0.94, indicating that improvements in preference prediction accuracy translated reliably and proportionally into controllability gains under the experimental conditions examined.

Toxicity suppression results exhibited a more complex pattern of capacity dependence. Under the moderate KL coefficient of 0.1, toxicity scores decreased consistently from the 125 million to the 2.7 billion parameter reward model, reaching a minimum mean toxicity score of 0.127, before increasing marginally for the 6.7 billion parameter model to 0.143. This non-monotonic relationship between reward model scale and toxicity suppression suggests that very large reward models may be more prone to distributional shift during policy optimization, potentially because their higher expressive capacity allows them to fit more precisely to the training preference distribution in ways that do not generalize to the adversarial prompt distribution used in toxicity evaluation. This interpretation is supported by the observation that the 6.7 billion parameter model achieved the highest refusal rate among all conditions at the 0.1 KL coefficient setting, suggesting that a portion of its toxicity suppression performance was achieved through excessive output caution rather than genuine internalization of safety preferences.

The effect of KL coefficient on the controllability-performance tradeoff was pronounced and consistent across reward model capacity levels. Under the lowest KL coefficient of 0.01, all reward model capacity conditions achieved their highest instruction-following task completion rates but also their highest mean toxicity scores, reflecting the tendency to over-exploit the learned reward signal at the expense of safety alignment. Under the highest KL coefficient of 0.5, toxicity scores were minimized across all capacity conditions but task completion rates fell to levels marginally above the SFT baseline, indicating that excessive regularization effectively prevented the policy from leveraging the reward signal. The intermediate KL coefficient of 0.1 produced the most favorable combined profile across both instruction-following and toxicity dimensions for all capacity conditions, reinforcing its selection as the canonical setting for alignment-performance balance in RLHF deployments. The magnitude of this KL sensitivity effect was itself modulated by reward model capacity, with larger models exhibiting greater sensitivity to the KL coefficient across the range tested, a pattern consistent with theoretical predictions regarding the interaction between reward signal strength and policy gradient magnitude in high-capacity optimization regimes.

4.2. Preference Data Diversity and Controllability Generalization

The second major experimental finding concerns the critical role of preference data diversity in determining the generalizability of controllability improvements to out-of-distribution generation scenarios. The analogy between attribute-level controllability in generative models and reward-conditioned policy steering is vividly illustrated in Figure 3, which presents the disentangled generative behavior of InfoGAN trained on MNIST digit images. Panels (a) through (d) demonstrate how varying individual latent codes c_1, c_2, and c_3 independently controls digit class identity, rotation angle, and stroke width while preserving the values of all other attributes — a form of structured, single-factor controllability that serves as a conceptual reference point for understanding how fine-grained reward conditioning in RLHF achieves targeted behavioral control over specific output dimensions without disrupting performance on orthogonal generation objectives. The contrast between InfoGAN's structured latent manipulation in panel (a) and the unstructured variation produced by regular GAN in panel (b) parallels the contrast this study observes between RLHF-trained policies and SFT baselines: the former exhibit organized, dimension-specific behavioral responses to targeted reward conditioning, while the latter lack the structured preference representations necessary for reliable out-of-distribution controllability generalization.



(a) Varying c_1 on InfoGAN (Digit type) (b) Varying c_1 on regular GAN (No clear meaning)

(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation) (d) Varying c_3 from -2 to 2 on InfoGAN (Width)

Figure 3. Disentangled controllable generation results from InfoGAN trained on MNIST

Reward models trained on the narrow single-domain preference dataset — constructed exclusively from instruction-following prompt-response pairs — achieved competitive task completion rates on in-domain instruction-following evaluation but exhibited substantially degraded performance on out-of-domain dialogue generation and creative writing tasks compared to models trained on the diverse multi-domain dataset of equivalent total size. The

performance gap was largest for stylistic consistency evaluation, where narrow-data reward models produced composite stylistic consistency scores 31.7 percentage points below those of diverse-data counterparts when tested on creative writing style specifications absent from the training distribution. This gap increased monotonically with reward model capacity for the narrow-data condition, while the reverse relationship — larger models generalizing better — held for the diverse-data condition, establishing a critical interaction between data diversity and capacity scaling that has direct practical implications for RLHF pipeline design.

The interaction between preference data diversity and reward model capacity produced a finding with important consequences for resource allocation decisions in industrial RLHF deployments. For diverse preference datasets, larger reward models consistently outperformed smaller counterparts on out-of-domain evaluation tasks across all KL coefficient settings, suggesting that increased model capacity was effectively leveraged to capture the complex and multi-dimensional preference structures encoded in diverse training data. For narrow preference datasets, the capacity-performance relationship reversed for out-of-domain evaluation, with larger models exhibiting greater overfitting to the training distribution and correspondingly reduced generalization. The 6.7 billion parameter reward model trained on the narrow dataset achieved out-of-domain stylistic consistency scores 18.3 percentage points below those of the 125 million parameter model trained on the same narrow dataset, despite the former's superior performance on in-domain tasks. This finding indicates that practitioners should prioritize preference data diversity over reward model architectural scaling when designing RLHF pipelines for open-domain deployment contexts, reversing the conventional wisdom that larger models uniformly improve alignment outcomes.

Analysis of toxicity suppression results across diverse versus narrow preference datasets revealed that models trained with diverse preference data incorporating adversarial and edge-case prompt-response pairs demonstrated significantly lower mean toxicity scores across the adversarial evaluation set compared to models trained on narrow datasets, even when the latter contained a higher absolute proportion of safety-annotated preference pairs as a fraction of total training data. This result suggests that the distributional breadth of the preference training corpus is a more important determinant of safety generalization than the density of safety-relevant annotations within a narrow distribution, emphasizing the value of constructing preference datasets that span the full range of generation contexts in which the deployed model will be used. The joint examination of toxicity scores and refusal rates for diverse-data models further revealed a more favorable safety profile, with lower toxicity achieved at lower refusal rates compared to narrow-data counterparts, indicating that diverse preference training enables genuine internalization of safety preferences rather than reliance on blanket output suppression. Across all experimental conditions examined, the results converge on a coherent picture in which the combination of diverse preference data, appropriately scaled reward model capacity, and moderate KL regularization produces the most reliable and generalizable controllability improvements in large generative models, with each factor playing a distinct and complementary role in the overall alignment pipeline.

5. Conclusion

This paper has presented a systematic experimental investigation of reward modeling from human feedback as a mechanism for improving controllability in large generative models, examining the joint influence of reward model capacity, preference data diversity, and policy optimization regularization on downstream generation alignment across instruction-following, toxicity suppression, and stylistic consistency dimensions. The results establish a set of empirically grounded findings that advance both the scientific understanding of RLHF mechanisms and the practical design of reward-based alignment pipelines for controllable generative AI systems.

The consistent superiority of RLHF-trained models over SFT baselines across all evaluation dimensions confirms the fundamental effectiveness of reward modeling as a controllability mechanism, extending prior findings to a more comprehensive and systematically controlled experimental setting than has been previously reported. The identification of reward model capacity as a primary driver of controllability improvements, subject to the critical interaction with preference data diversity, provides actionable guidance for practitioners navigating the tradeoff between computational investment and alignment quality. The observation that preference data diversity is a more critical determinant of controllability generalization than reward model scale for out-of-distribution evaluation scenarios carries important implications for resource allocation in industrial RLHF deployments, suggesting that data collection investments should precede architectural scaling efforts whenever deployment contexts extend beyond the distribution of training prompts. The characterization of the KL coefficient as a practical lever for navigating the alignment-performance tradeoff, with the intermediate value of 0.1 consistently producing superior combined profiles across evaluation dimensions, offers a concrete tuning guideline applicable across a range of generation tasks and model scales.

Several limitations of the present study merit acknowledgment. The preference data used in this investigation, while diverse across four generation categories, remains constrained to text-only modalities and English-language outputs, limiting the direct applicability of findings to multilingual or multimodal generation settings. The evaluation framework, while multi-dimensional, captures a subset of the behavioral dimensions relevant to real-world deployment, and future work should extend the assessment to additional controllability dimensions including factual accuracy, long-form coherence, and cross-turn consistency in extended dialogue. The computational constraints of the study precluded evaluation of reward models beyond 6.7 billion parameters, leaving open the question of whether the observed capacity-performance relationships hold at the multi-tens-of-billions parameter scales characteristic of frontier model development.

Future research should pursue several directions suggested by the present findings. The extension of systematic reward modeling analysis to multimodal generative architectures, where the alignment of visual and linguistic outputs with human preferences introduces additional complexity, represents a particularly important frontier for the next generation of controllability research. The development of more efficient preference data collection protocols,

combining active learning, synthetic data generation, and structured elicitation interfaces, offers potential to reduce annotation costs while maintaining or improving the distributional diversity identified here as critical for generalization. Deeper mechanistic investigation of how reward-based fine-tuning reshapes internal representations at specific layers and attention heads would advance model interpretability and enable more targeted alignment interventions. Finally, the integration of reward modeling with constitutional and principle-based alignment frameworks offers a promising direction for developing alignment pipelines that are simultaneously data-efficient, transparent, and robust to the distributional shifts that characterize real-world deployment of controllable generative AI systems.

References

[1] 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.

[2] Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751.

[3] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

[4] Ding, J., & Qin, Y. (2026). Raft and Beyond: Practical Consensus Mechanisms for Geo-Distributed Data Systems. Computer Life, 14(1), 54-63.

[5] Yang, Y., & Yang, J. (2026). Synthetic Data Meets Finance: Generative Models for Privacy Preserving Analytics. Journal of Banking and Financial Dynamics, 10(4), 1-8.

[6] Wang, Z., Shen, Z., Wang, B., & Shang, W. (2025). Modernizing Enterprise Analytics through Low-Code Automation and Cloud-Native Data Architectures. Asian Business Research Journal, 10(12), 20-33.

[7] Zhao, X., Sun, T., Ren, S., Yang, J., & Liu, Y. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment. Frontiers in Artificial Intelligence Research, 2(3), 501-520.

[8] Li, P., Liu, J., & Qiu, L. (2026). Deep Learning Methods for Demand Forecasting and Inventory Optimization in Modern Supply Chains. Asian Business Research Journal, 11(3), 21-29.

[9] Qiu, L. (2025). Reinforcement Learning Approaches for Intelligent Control of Smart Building Energy Systems with Real-Time Adaptation to Occupant Behavior and Weather Conditions. Journal of Computing and Electronic Information Management, 18(2), 32-37.

[10] Zhang, H. (2025). Reinforcement Learning Approaches for Layout Optimization in Electronic Design Automation with Electromagnetic Compatibility Constraints. Frontiers in Robotics and Automation, 2(2), 77-93.

[11] Shen, Z., Zhao, W., Wang, B., Wang, Z., & Shang, W. (2026). CAGR: A Cross-Accelerator Graph Optimization Framework for Efficient Recommender System Inference. IEEE Access.

[12] Sun, T., Wang, M., & Han, X. (2025). Deep Learning in Insurance Fraud Detection: Techniques, Datasets, and Emerging Trends. Journal of Banking and Financial Dynamics, 9(8), 1-11.

[13] Liu, J., Li, P., & Wang, Y. (2026). Graph Neural Networks for Modeling Complex Dependencies in Global Supply Chain Networks. Journal of Computing and Electronic Information Management, 20(3), 9-20.

[14] Zhang, F., & Wu, B. (2025). Large Language Models as General Purpose Intelligence Systems for Reasoning, Planning and Decision Making. American Journal of Artificial Intelligence and Neural Networks, 6(4), 45-72.

[15] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. IEEE Access, 12, 195974-195985.

[16] Zhang, F., & Yang, J. S. (2025). Learning Driven Decision Intelligence for Autonomous Driving Through Multimodal Understanding World Modeling and Policy Optimization. Frontiers in Artificial Intelligence Research, 2(3), 616-634.

[17] Wang, B., Wang, Z., Zhao, W., & Liu, Y. (2025). Network Fabric Simulation and Validation for Data Center Routing Convergence Under Large-Scale Failure Scenarios. Computer Science Bulletin, 8(01), 310-326.

[18] Liu, J., Wang, J., Chen, H., Guinness, J., Martin, R., & Kulkarni, C. S. (2019). Optimal Level Crossing Predictions for Electronic Prognostics. In AIAA Scitech 2019 Forum (p. 1962).

[19] Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. Applied Sciences, 14(20), 9213.

[20] Wei, Z., Sun, T., & Zhou, M. (2024). LIRL: Latent Imagination-Based Reinforcement Learning for Efficient Coverage Path Planning. Symmetry, 16(11), 1537.

[21] Zhang, S., Qiu, L., & Zeng, Z. (2026). Physics-Data Synergy in Structural Health Monitoring: A Multi-Scale Graph Contrastive Framework With Temperature-Adaptive Fusion. IEEE Access.

[22] Zeng, Z., Lin, H., Zhang, S., & Wang, B. (2026). Adaptive Robust Watermarking for Large Language Models via Dynamic Token Embedding Perturbation. IEEE Access, 14, 9319-9339.

[23] Qiu, L. (2025). Multi-Agent Reinforcement Learning for Coordinated Smart Grid and Building Energy Management Across Urban Communities. Computer Life, 13(3), 8-15.

[24] Zhao, W., Chen, T., Yang, J. S., & Qiu, L. (2026). AutoML-Pipeline: A RAG-enhanced code generation framework with pre-validation for cloud-native machine learning workflows. IEEE Access.

[25] Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., ... & Clark, J. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858.

[26] Scherrer, N., Shi, C., Feder, A., & Blei, D. (2023). Evaluating the moral beliefs encoded in llms. Advances in Neural Information Processing Systems, 36, 51778-51809.

[27] Chen, T., & Ding, J. (2026). Cold Start Latency Optimization Strategies for Function as a Service Platforms. Computer Life, 14(1), 64-73.

[28] Scheurer, J., Campos, J. A., Korbak, T., Chan, J. S., Chen, A., Cho, K., & Perez, E. (2023). Training language models with language feedback at scale. arXiv preprint arXiv:2303.16755.