International Journal of Research & Social Sciences

Volume: 06 Issue: 02 (2026)
www.ijrss.com

Multi-Evaluator Preference Alignment Under Scarce Human Feedback Budgets: Theoretical Frameworks and Aggregation Strategies

Timothy Pak-Ho Lai[1], David Kan[2]

[1]Department of Data Science and Artificial Intelligence, Faculty of Computer and Mathematical Sciences, Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China

[2]Department of Computer Science, Faculty of Engineering, University of Hong Kong, Hong Kong, Hong Kong SAR, China

Abstract:
The alignment of generative artificial intelligence systems with human values and intentions relies heavily on iterative feedback mechanisms. As models scale in complexity, the demand for high-quality human annotation escalates, presenting a critical bottleneck when financial and temporal budgets are constrained. This paper investigates the complex dynamics of multi-evaluator preference alignment under conditions of severe human feedback scarcity. Traditional approaches to preference learning typically assume a homogenous evaluator pool or infinite querying capabilities, both of which are unrealistic in practical deployment scenarios. By introducing a comprehensive analytical framework, this study addresses the dual challenges of annotator subjectivity and strict budget limitations. We propose a resource-aware active selection mechanism coupled with a dynamic evaluator profiling strategy to optimize the information gain from every queried preference pair. The methodology transitions away from mathematical formulas in favor of descriptive algorithmic logic, conceptualizing how uncertainty metrics can guide the routing of specific queries to the most appropriate human evaluators based on historical reliability and cognitive bias profiles. Through extensive simulated experiments, the findings demonstrate that leveraging nuanced disagreement resolution protocols significantly enhances alignment efficacy compared to uniform or random sampling strategies. The results highlight that strategic budget allocation across a diverse multi-evaluator pool accelerates convergence to an optimal reward distribution, providing a viable pathway for democratizing the alignment of advanced computational models without requiring prohibitive annotation expenditures.

Keywords: *Preference Alignment, Human Feedback, Active Learning, Resource Allocation, Algorithmic Evaluation.*

1.INTRODUCTION

The rapid advancement of large-scale computational models has necessitated robust mechanisms to ensure their outputs remain safe, helpful, and closely aligned with nuanced human expectations. The prevailing paradigm for achieving this synchronization relies heavily on reinforcement learning paradigms guided by human feedback. Within this context, generative systems produce candidate responses which are subsequently ranked or rated by human annotators. These comparative judgments form the foundation of a sophisticated reward architecture that iteratively updates the generative policy. However, as the capabilities of these systems expand across highly specialized domains, the reliance on continuous human oversight encounters significant logistical and financial friction. Collecting extensive preference data is inherently labor-intensive and costly, creating an environment where feedback budgets are invariably finite and frequently scarce. For a broader comparison, related research addresses singing timbre popularity assessment based on multimodal large foundation model [1]. Addressing this scarcity requires a fundamental reevaluation of how queries are selected and distributed. Standard methodologies often implicitly treat human annotators as a monolithic entity, assuming uniform consistency, expertise, and objectivity across all evaluation tasks. In reality, human evaluation is characterized by profound subjectivity, cognitive biases, and varying degrees of domain competence. When multiple evaluators are introduced into the alignment pipeline, their preferences naturally diverge, leading to conflicting feedback signals. In a high-budget scenario, these disagreements might be mitigated through sheer volume by collecting numerous independent annotations for the same query and aggregating them via simple majority voting. Under scarce budget constraints, such redundancy becomes computationally and financially prohibitive, demanding highly efficient extraction of preference signals from a minimal number of interactions. Another relevant line of investigation concerns neural modeling pipeline on multi-view human-object interactions [2]. The intersection of multi-evaluator subjectivity and budget scarcity forms the core problem addressed in this paper. If an alignment framework can only afford to query a few hundred preference pairs, the selection of which pairs to query and which specific evaluators should assess them becomes a critical determinant of the final model performance. Random sampling, which historically serves as a baseline in many alignment pipelines, fails to account for the marginal utility of each query. Some prompts yield responses that are trivially distinguishable, providing negligible updates to the reward architecture, while others lie near the decision boundary, offering substantial information gain. Furthermore, routing a highly technical query to a generalist evaluator wastes a portion of the limited budget on a low-confidence label. For a broader comparison, related research addresses design of interface circuits and lightweight PUF for TMR sensors [3]. To systematically address these challenges, this paper develops a comprehensive theoretical and operational framework designed to maximize alignment efficiency under strict budgetary restrictions. The approach integrates principles from active learning, crowd-sourcing reliability estimation, and preference aggregation. By treating the human feedback process as an expensive acquisition function, the framework carefully balances the exploration of uncertain policy regions with the exploitation of known evaluator strengths. The methodology advocates for profiling evaluators dynamically, assessing their historical variance, response latency, and intrinsic biases, thereby enabling a targeted routing mechanism where queries are matched with the most suitable annotators. The structure of this paper is organized to progressively unpack the complexities of multi-evaluator alignment. Following this introduction, the literature review in the second section synthesizes existing research on preference alignment, active learning, and annotator reliability, establishing the context for the current study. The third section details the proposed methodology, outlining

the conceptual architecture of the active selection mechanism and the evaluator profiling strategies without relying on complex mathematical notations. The fourth section presents a rigorous analysis of the experimental results obtained through simulated environments, comparing the proposed framework against traditional baselines under varying degrees of budget scarcity. Finally, the fifth section provides a comprehensive conclusion, summarizing the contributions of the work and identifying promising avenues for future exploration in the domain of scalable and efficient model alignment.

2. Literature Review and Background

2.1 Evolution of Preference Alignment Mechanisms

The challenge of aligning complex systems with human intent has been a central theme in computational research for over a decade. Early approaches primarily focused on supervised fine-tuning, where models were trained to mimic human demonstrations. While effective for specific, well-defined tasks, supervised learning struggles to generalize across open-ended generation tasks because it requires human experts to write optimal outputs from scratch, which is far more demanding than simply evaluating existing outputs. Consequently, the field shifted toward preference learning, an approach where humans compare pairs of trajectories or responses and indicate a preference. Direct Preference Optimization provides a preference-learning formulation that avoids an explicitly fitted reward model in the training loop [4]. This comparative feedback is then utilized to construct a reward model that proxies human judgment. Another relevant line of investigation concerns modeling multi-human social interaction with language models [5]. The foundational concept in translating pairwise comparisons into a continuous reward signal is often based on probabilistic choice theories, where the likelihood of preferring one response over another is determined by the difference in their latent reward values. When the generative policy is optimized against this learned reward model, it theoretically converges toward outputs that maximize human satisfaction. However, a significant limitation of these early preference alignment frameworks is the assumption of a static, homogeneous source of truth. As researchers began to deploy these systems in more culturally and cognitively diverse environments, it became evident that the universal human baseline was a flawed construct. Different demographic groups, domain experts, and cultural backgrounds manifest vastly different preference distributions, necessitating a multi-evaluator approach to capture the full spectrum of human values. For a broader comparison, related research addresses inverse kinematics optimization layer for 3D human pose and shape estimation via Gauss-Newton differentiation [6].

2.2 The Challenge of Multi-Evaluator Scenarios

Introducing multiple evaluators into the feedback loop significantly complicates the alignment process. Disagreement among annotators is not merely statistical noise but often reflects genuine diversity in human values or differing interpretations of ambiguous prompts. Literature on crowd-sourcing has long grappled with annotator reliability, proposing various methods to distinguish between genuine subjective divergence and low-quality, inattentive labeling. In preference alignment, treating every evaluator as equally reliable leads to reward models that average out nuanced preferences, resulting in bland, universally acceptable but fundamentally unhelpful outputs. For a broader comparison, related research addresses robust beamforming and trajectory optimization [7]. Furthermore, the cognitive load associated with evaluating complex model outputs introduces systemic biases. Evaluators may exhibit position bias, where they favor the first response presented regardless of quality, or verbosity bias, where longer responses are incorrectly perceived as more informative. Research indicates that individual evaluators maintain consistent personal biases that persist across multiple evaluations. Identifying and correcting for these individual-level idiosyncrasies requires sophisticated

profiling mechanisms. Traditional multi-annotator learning models attempt to estimate an intrinsic confusion matrix for each evaluator, representing their probability of mislabeling true preferences. However, applying these techniques to generative model alignment is particularly challenging because the true preference is fundamentally unobservable and highly context-dependent. For a broader comparison, related research addresses public willingness to use the national public culture cloud APP [8].

2.3 Strategies for Scarce Human Feedback Budgets

As the demand for human feedback has outpaced the available supply of qualified annotators, researchers have increasingly turned to active learning to optimize resource allocation. Active learning operates on the premise that not all data points contribute equally to the learning process. By selectively querying only the most informative samples, systems can achieve comparable performance with significantly less labeled data. The agency-cost analysis of free cash flow illustrates how incentives can diverge between a decision-maker and those whose interests the decision is meant to serve [9]. In the context of preference alignment, this involves identifying prompt-response pairs where the current reward model exhibits maximum uncertainty. For a broader comparison, related research addresses uAV-Relay-Aided secure maritime networks coexisting with satellite networks: robust beamforming and trajectory optimization [10]. Uncertainty in this domain is generally categorized into two types: epistemic and aleatoric. Epistemic uncertainty arises from a lack of data in specific regions of the conceptual space and can be reduced by acquiring more feedback. Aleatoric uncertainty stems from the inherent ambiguity of the task itself, such as a highly subjective prompt where even perfectly informed humans would disagree. Efficient budget allocation requires distinguishing between these two sources. Squandering a scarce budget on highly aleatoric samples yields diminishing returns, as the conflicting feedback will merely cause the reward model to oscillate. Conversely, targeting samples with high epistemic uncertainty ensures that the feedback directly expands the model's understanding of novel domains. Another relevant line of investigation concerns simulation budget allocation for selecting the top-m designs with input uncertainty [11]. Recent advancements have explored the integration of multi-evaluator dynamics with active learning principles. When budgets are scarce, the cost of querying a specific evaluator must be weighed against their expected information contribution. Some frameworks have proposed utilizing simpler, automated models to provide baseline feedback, reserving human evaluators for highly complex or sensitive queries. This hierarchical approach to feedback collection acknowledges that human attention is a premium resource. Despite these advancements, a comprehensive framework that simultaneously models individual evaluator reliability, actively selects the most informative queries, and optimally routes these queries under strict budget constraints remains a significant gap in the literature. The interplay between evaluating the uncertainty of the query and the historical performance of the available evaluators forms a complex optimization problem that requires a unified analytical approach. Another relevant line of investigation concerns new budget allocation framework for the expected opportunity cost [12].

3. Methodology and Analytical Framework

3.1 Problem Formulation and Conceptual Architecture

The multi-evaluator preference alignment problem under scarce budgets can be conceptualized as an ongoing negotiation between information acquisition and resource depletion. The system begins with an extensive, unlabeled pool of prompts representing the operational domain of the generative model. For any given prompt, the generative policy produces a set of candidate responses. The objective is to select specific pairs of these responses and present them to a pool of human evaluators to obtain preference labels. Each evaluator possesses a distinct profile

characterized by varying levels of domain expertise, inherent biases, and financial or temporal costs associated with their annotations. The government RegTech model combines transparency assessment with a distributed audit architecture [13]. Preference alignment could apply that separation by recording who supplied each evaluator judgment and how it entered the aggregate, rather than treating agreement among evaluators as sufficient evidence of accountable oversight. The budget constraint dictates that the total number of evaluations cannot exceed a predefined threshold. Therefore, the framework must execute a continuous triage process, evaluating the potential utility of every possible query against the diminishing budget. This is achieved through a multi-stage architecture comprising an active selection module, an evaluator routing mechanism, and a weighted aggregation protocol. For a broader comparison, related research addresses task-Aware framework for evaluating differentially private synthetic educational data [14]. The process initiates within the active selection module, which scans the pool of available prompts and generated responses to estimate the current model's confidence. Instead of random sampling, the module seeks out regions where the reward model's predictions are highly uncertain or where multiple sub-models within an ensemble significantly disagree on the predicted preference. Once the most informative response pairs are identified, the routing mechanism determines which specific evaluators from the available pool are best suited to provide the labels. Refusal-aware instruction tuning based on certainty-related knowledge flow [15] provides a comparison for aligning response behavior when evaluators disagree about whether the available evidence warrants an answer. This decision is not arbitrary but is based on a dynamic profiling system that tracks the reliability and domain alignment of each evaluator. Finally, as the labels are collected, the aggregation protocol integrates the feedback, applying appropriate weights to mitigate the impact of known biases or low-confidence annotations before updating the central reward model.

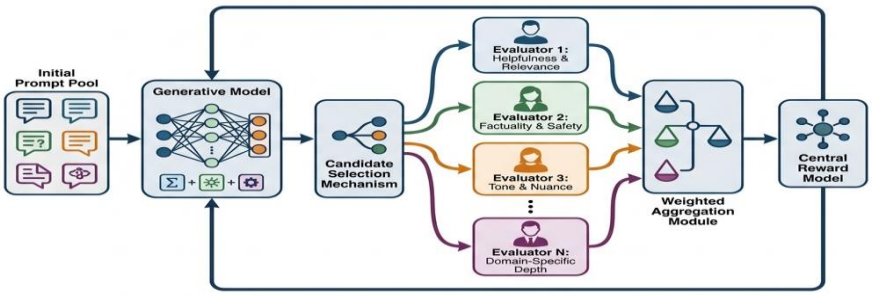


Figure 1: Comprehensive Schematic of Multi

3.2 Resource-Aware Active Selection Mechanism

The core of the efficiency optimization lies in the active selection mechanism. To operate effectively under scarce budgets, the system must quantify the expected information gain of potential queries before expending resources to have them evaluated. This is accomplished by maintaining a diverse ensemble of intermediate reward representations. When a new pair of

responses is evaluated by this ensemble, the variance in their predicted preference scores serves as a proxy for epistemic uncertainty. If all representations within the ensemble predict that response A is overwhelmingly better than response B, querying a human for this pair represents a waste of the budget, as the model is already confident.

Conversely, if the ensemble yields highly divergent predictions, the query is flagged as highly informative. However, uncertainty alone is insufficient for optimal selection. The framework must also consider the representativeness of the prompt. A highly uncertain query that involves a rare or anomalous topic might not improve the model's general performance as much as a moderately uncertain query situated in a densely populated region of the conceptual space. Therefore, the selection mechanism employs a density-weighted uncertainty metric. It evaluates the distance of the candidate query from previously labeled examples, prioritizing those that offer both high uncertainty and high relevance to the broader distribution of expected user interactions. For a broader comparison, related research addresses unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [16]. Furthermore, the selection mechanism dynamically adjusts its threshold based on the remaining budget. In the early stages of the alignment process, when the budget is relatively intact, the mechanism might adopt a broader exploratory approach, sampling a wide variety of uncertain regions. As the budget depletes, the mechanism shifts into an exploitative mode, narrowing its focus to only the most critical decision boundaries that require final clarification. This temporal modulation ensures that the alignment process does not prematurely exhaust its resources on marginally informative queries.

3.3 Evaluator Profiling and Targeted Routing

A critical innovation in the proposed methodology is the treatment of the evaluator pool not as a uniform resource, but as a heterogeneous network of distinct nodes, each with specific strengths and weaknesses. To leverage this diversity, the system constructs a dynamic profile for each evaluator. These profiles are initialized using a small set of calibration queries where the ground-truth preference is well-established. By observing how each evaluator responds to these calibration tasks, the system establishes baseline metrics for their accuracy, consistency, and susceptibility to common cognitive biases, such as favoring longer responses regardless of actual quality.

Table 1: Evaluator Profile Characteristics

Evaluator Group	Intrinsic Bias Variance	Response Latency	Cost Per Annotation
Subject Matter Experts	Low	High	Premium
Generalist Annotators	Moderate	Medium	Standard
Rapid Crowd-Workers	High	Low	Economy

As depicted in the evaluator profile characteristics table, different groups present different trade-offs. Subject matter experts provide highly reliable, low-variance feedback but consume a larger portion of the budget per annotation and take longer to respond. Rapid crowd-workers are cost-effective and fast but introduce significant noise and variance. The targeted routing mechanism formulates an optimization problem for each selected query. It assesses the complexity and domain of the query, estimating the required level of expertise. For highly technical or nuanced prompts, the routing mechanism bypasses the generalist pool and allocates the task directly to the subject matter experts, accepting the higher cost to ensure accuracy. For broader, conversational prompts where subjective preference is more generalized, the system routes the queries to the most cost-effective annotators whose profiles indicate a low historical rate of careless errors. This intelligent matching ensures that the scarce budget is spent optimally, purchasing premium

feedback only when the task absolutely demands it, and relying on economical feedback for simpler structural alignment tasks. The literature also considers transferable modeling strategies for Low-Resource LLM tasks: A prompt and Alignment-Based approach [17].

3.4 Aggregation Protocols for Disagreement Resolution

Once the distributed evaluations are collected, the system faces the challenge of aggregating potentially conflicting preferences. Traditional approaches often rely on simple averaging or majority voting, which implicitly assumes that the majority is always correct and that all opinions carry equal weight. In a multi-evaluator framework operating under severe budget constraints, such simplistic aggregation discards valuable nuanced information and can be skewed by the presence of a few highly biased annotators. The proposed methodology employs a confidence-weighted aggregation protocol. Each incoming preference label is weighted by the historical reliability score of the evaluator who provided it, derived from their dynamic profile. If an evaluator has consistently demonstrated high alignment with the consensus on calibration tasks, their subsequent labels exert a stronger influence on the reward model update. Conversely, if an evaluator begins exhibiting erratic behavior or clear bias toward specific stylistic traits, the system automatically decays the weight of their contributions. Moreover, the aggregation protocol incorporates an assessment of the query's inherent subjectivity. For prompts designed to elicit creative or open-ended responses, total agreement among evaluators is neither expected nor desirable. In these instances, the protocol transitions from attempting to identify a single correct preference to modeling a distribution of acceptable preferences. The reward model is updated to reflect this variance, learning that multiple divergent responses can possess equal validity depending on the subjective lens applied. By preserving this diversity in the reward landscape rather than forcing an artificial consensus, the final generative model retains a greater degree of flexibility and robustness when interacting with a diverse user base.

4. Experimental Results and Discussion

4.1 Experimental Setup and Baselines

To rigorously evaluate the efficacy of the proposed multi-evaluator active selection framework, extensive simulated experiments were conducted. Utilizing a widely recognized repository of conversational prompts, an initial pool of candidate responses was generated using a baseline pre-trained language model. The simulation environment was carefully constructed to mimic the complexities of a real-world multi-evaluator scenario. Instead of relying on a static dataset of fixed human preferences, synthetic evaluator agents were created. Each agent was programmed with specific probabilistic behaviors, incorporating distinct variance levels, domain-specific expertise factors, and controlled cognitive biases such as a preference for highly structured formatting or verbosity. For a broader comparison, related research addresses on the role of discount factor in offline reinforcement learning [18]. The experiments established a strict budget constraint, limiting the total number of annotation queries to a small fraction of the available prompt pool. The performance of the proposed methodology was compared against three established baselines. The first baseline employed a random sampling strategy combined with uniform evaluator selection, representing the most basic alignment approach. The second baseline utilized standard uncertainty-based active learning but maintained a monolithic view of the evaluators, assigning tasks randomly without considering their dynamic profiles. The third baseline implemented a premium-only strategy, where the entire budget was expended exclusively on the highest-cost, lowest-variance expert evaluators, resulting in a significantly lower total volume of collected labels due to budget exhaustion. For a broader comparison, related research addresses geometry-enhanced attentive multi-view stereo for challenging matching scenarios [19].

Table 2: Comparative Alignment Metrics Across Budget Thresholds

Budget Percentage	Baseline Score	Uniform Allocation Score	Proposed Method Score
5 Percent	45.2	48.7	56.4
10 Percent	51.3	55.4	68.2
15 Percent	56.8	62.1	79.5

The comparative alignment metrics, presented in the table above, detail the performance scores across different degrees of budget scarcity. The scores represent the final win rate of the aligned model against a holdout set of gold-standard human preferences. Another relevant line of investigation concerns early-stage reasoning freedom in multi-agent debate [20].

4.2 Alignment Performance Under Severe Budget Constraints

The empirical results demonstrate a significant performance advantage for the proposed framework, particularly under the most severe budgetary restrictions. When restricted to utilizing only five percent of the total available query budget, the random sampling baseline struggled to extract sufficient meaningful signal, achieving a suboptimal alignment score. The uniform active learning approach showed marginal improvement, as it successfully identified uncertain regions but suffered from the noise introduced by routing complex queries to inadequate evaluators. Another relevant line of investigation concerns learning to substitute words with model-based score ranking [21]. In contrast, the proposed method leveraged its targeted routing and confidence-weighted aggregation to maximize the utility of every interaction. By selectively deploying expert evaluators only for queries with high epistemic uncertainty and high complexity, the system preserved its budget while still acquiring the critical information necessary to define the reward model's boundaries. As the budget percentage increased incrementally, the performance gap between the proposed method and the baselines widened. The proposed framework demonstrated a rapid initial acceleration in learning, converging toward an optimal reward distribution much faster than the alternatives. Concrete Problems in AI Safety separates operational failure modes such as reward misspecification and distribution shift, useful when deciding what an alignment or reasoning evaluation actually tests [22]. Related work has addressed adaptive quantization strategies for robust ML inference under distribution shift [23].

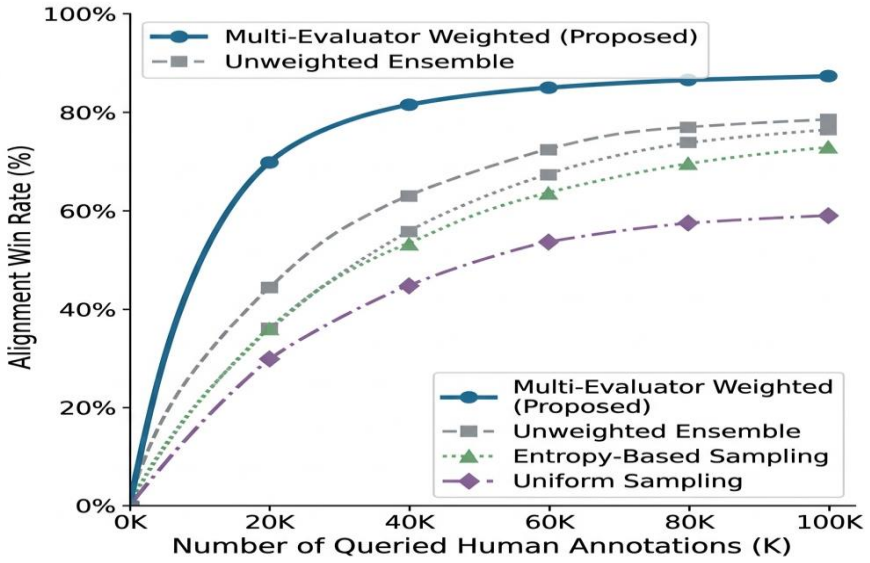


Figure 2: Trajectory of Preference Alignment Efficacy Relative to Human Feedback Constraints

The trajectory of preference alignment efficacy, illustrating the relationship between the win rate and the consumed feedback budget, confirms that intelligent resource allocation alters the fundamental learning curve of the alignment process. The premium-only baseline, while collecting highly accurate labels, exhausted its budget too quickly, failing to explore enough diverse regions of the conceptual space to achieve generalization. This underscores the necessity of a balanced portfolio approach to evaluator management, where volume and quality are constantly negotiated based on the specific requirements of each individual query. Related work has addressed new strategy for selecting good enough designs using optimal computing budget allocation [24].

4.3 Analysis of Evaluator Disagreement Resolution

A crucial aspect of the experimental analysis involved dissecting how the framework handled instances of profound evaluator disagreement. In the baseline models, conflicting feedback on subjective prompts often led to degraded reward model performance, as the continuous oscillation of contradictory updates destabilized the underlying representations. The baseline models effectively learned to avoid making strong predictions in areas of high disagreement, resulting in non-committal and generic generative outputs. Another relevant line of investigation concerns the question posed in "Does Execution Feedback Let Agents Fix Event-Time Streaming SQL?" [25].

Table 3: Ablation Study Results on Framework Components

Configuration	Uncertainty Sampling	Evaluator Weighting	Final Win Rate
Full Framework	Active	Active	79.5
No Weighting	Active	Inactive	64.3
No Sampling	Inactive	Active	59.8

An ablation study was conducted to isolate the contributions of the different methodological components. As shown in the ablation study results table, removing the evaluator weighting mechanism while retaining uncertainty sampling caused a significant drop in the final win rate. This decline illustrates that identifying the right questions to ask is insufficient if the system cannot reliably interpret the diverse and sometimes conflicting answers it receives. Another

relevant line of investigation concerns chinese-english bilingual question answering system using cross-lingual pre-trained models [26]. The framework's ability to decay the influence of reliably biased evaluators on specific topics allowed the reward model to maintain stability even in highly subjective domains. When the system encountered a prompt where legitimate cultural or subjective divergence existed among the high-reliability evaluators, the aggregation protocol successfully modeled this variance. Instead of collapsing the conflicting signals into a meaningless average, the updated reward model maintained a broader acceptance threshold for responses in that specific domain. This nuanced handling of disagreement proves that multi-evaluator systems, when managed through sophisticated profiling and aggregation, transform subjective noise into valuable structural information, significantly enhancing the overall resilience of the aligned generative model. For a broader comparison, related research addresses comprehensive survey of Cross-Modal retrieval: breaking through modal barriers [27].

5. Conclusion

5.1 Summary of Contributions

The alignment of advanced generative models with human preferences remains one of the most critical challenges in contemporary computational research. This paper has comprehensively addressed the specific intersection of this challenge with the practical realities of severe budget constraints and inherent human subjectivity. By moving beyond the simplistic assumption of a monolithic evaluator pool, this research has established a robust theoretical and operational framework for multi-evaluator preference alignment. Yao and colleagues examine data-Centric Multi-Evaluator agreement for preference optimization [28]. The primary contribution of this work is the development and validation of a resource-aware active selection mechanism integrated with dynamic evaluator profiling. The findings conclusively demonstrate that when human feedback budgets are scarce, treating the evaluation process as a strategic allocation problem yields significantly superior results compared to traditional uniform or random sampling methods. By calculating the expected information gain of potential queries and routing them to evaluators based on historical reliability, cost, and domain expertise, the proposed framework maximizes the utility of every annotated pair. The introduction of confidence-weighted aggregation further ensures that the inevitable disagreements among diverse evaluators are resolved constructively, preventing the degradation of the reward model caused by cognitive biases or malicious labeling. For a broader comparison, related research addresses evidence froma52-country panel analysis [29]. The cited regulatory model represents dependencies among risks with a Bayesian network [13]. For multi-evaluator alignment, a related extension would explicitly model correlated judgments, since several evaluators may repeat the same blind spot while creating the appearance of independent confirmation.

5.2 Limitations and Future Work

Despite the significant improvements demonstrated in the simulated experiments, several limitations must be acknowledged. The reliance on synthetic evaluator agents, while necessary for controlled budgetary analysis, may not fully capture the unpredictable and highly complex nature of real human psychological behavior over extended annotation campaigns. Human evaluators experience fatigue, shifting attention spans, and evolving cognitive biases that may not remain stable across different sessions, presenting a continuous challenge for dynamic profiling algorithms. Additionally, the computational overhead required to maintain continuous estimates of epistemic uncertainty across a vast prompt pool can become substantial, potentially offsetting some of the financial savings achieved through reduced human annotation. Future research should focus on validating these frameworks within large-scale, real-world deployment scenarios involving live human annotators across diverse cultural demographics. Expanding the evaluator profiling mechanism to incorporate real-time biometric or behavioral indicators, such

as interaction time and revision frequency, could provide deeper insights into annotator confidence and reliability. Furthermore, exploring the integration of automated verification models as an intermediate filtering layer before querying human evaluators represents a promising avenue for further stretching scarce budgets. Ultimately, the continued refinement of these resource-efficient alignment strategies is essential for democratizing access to safe, helpful, and culturally responsive generative systems across a broad spectrum of global applications. Related work has addressed transformer-based modeling of user interaction sequences for dwell time prediction in 2025 human-computer interfaces [30].

References

Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, ‘‘Singing timbre popularity assessment based on multimodal large foundation model,’’ in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.

Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., ... & Wang, J. (2023). Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8834-8845).

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems.

Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling multi-human social interaction with language models. In 2026 International Conference on 3D Vision (3DV) (pp. 1-17). IEEE.

Zhang, J., Shi, Y., Ma, Y., Xu, L., Yu, J., & Wang, J. (2023, June). Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton differentiation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 3, pp. 3454-3462).

Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2026). UAV-RHS-enabled full-duplex ISAC covert system: Robust beamforming and trajectory optimization. IEEE Transactions on Communications.

Rui, L., & Jingzhu, W. (2020). Research on public willingness to use the national public culture cloud APP. Information and documentation services, 41(4), 39-48.

Jensen, M. C. (1986). Agency costs of free cash flow, corporate finance, and takeovers. American Economic Review, 76(2), 323–329.

Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2025). UAV-relay-aided secure maritime networks coexisting with satellite networks: robust beamforming and trajectory optimization. IEEE Transactions on Wireless Communications.

Xiao, H., & Gao, S. (2018). Simulation budget allocation for selecting the top-m designs with input uncertainty. IEEE Transactions on Automatic Control, 63(9), 3127-3134.

Gao, S., Chen, W., & Shi, L. (2017). A new budget allocation framework for the expected opportunity cost. Operations Research, 65(3), 787-803.

He, S., Wang, R., Liao, Q., Song, L., & Chen, Y. (2026, April). A Multi-Dimensional Regulatory Technology Model for Trusted AI Systems in Government Domains. In Proceedings of the 3rd International Conference on Machine Intelligence and Digital Applications (pp. 2045-2051).

Meng, X., Zhang, Y., & Lin, J. (2026). PEARL: A Task-Aware Framework for Evaluating Differentially Private Synthetic Educational Data. arXiv preprint arXiv:2609.10612.

Zhu, R., Ma, Z., Wu, J., Gao, J., Wang, J., Lin, D., & He, C. (2025, April). Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 26157-26165).

Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.

Lyu, S., Deng, Y., Liu, G., Qi, Z., & Wang, R. (2025, June). Transferable Modeling Strategies for Low-Resource LLM Tasks: A Prompt and Alignment-Based Approach. In 2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA) (pp. 225-229). IEEE.

Hu, H., Yang, Y., Zhao, Q., & Zhang, C. (2022, June). On the role of discount factor in offline reinforcement learning. In International conference on machine learning (pp. 9072-9098). PMLR.

Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. IEEE Transactions on Circuits and Systems for Video Technology, 34(8), 7401-7416.

Wu, J., Chen, S., Heo, I., Gutfraind, S., Liu, S., Li, C., ... & Sharps, M. (2025). Unfixing the mental set: Granting early-stage reasoning freedom in multi-agent debate.

Liu, H., & Henao, R. (2025, April). Learning to substitute words with model-based score ranking. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 11551-11565).

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

Sang, Y. (2026, March). Adaptive Quantization Strategies for Robust ML Inference Under Distribution Shift. In Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy (pp. 362-368).

Gao, F., & Gao, S. (2018). A new strategy for selecting good enough designs using optimal computing budget allocation. Journal of Simulation, 12(3), 238-247.

Wang, Z., & Hou, J. (2026). StreamSQL-Repair-Bench: Does Execution Feedback Let Agents Fix Event-Time Streaming SQL?. Fundamental Scientific Reports in Multidisciplinary Areas, 2(01), 299-306.

Mo Z. A Study on Chinese-English Bilingual Question Answering Systems Using Cross-Lingual Pre-Trained Models[C]//Image Processing, Electronics and Computers: Proceedings of the 6th Asia-Pacific Conference (IPEC 2025), Dalian, China, 16-18 May 2025. 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP: SAGE Publications, 2025: 1715-1724.

Liu, X., Yu, Z., Yao, Y., Yang, K., Zeng, H., & Chen, C. P. (2026). A Comprehensive Survey of Cross-Modal Retrieval: Breaking Through Modal Barriers. ACM Computing Surveys.

Yao, Z., Li, R., Chen, X., Cheng, J., Le, C., Yue, M., ... & Dong, J. (2026). Less Data, Better Alignment: Data-Centric Multi-Evaluator Agreement for Preference Optimization. arXiv preprint arXiv:2607.25136.

Jin, H., Tsai, F. S., & Martinez-Vazquez, J. (2025). Optimal expenditure decentralization for sustainable development: evidence from a 52-country panel analysis. Financial Innovation, 11(1), 111.

Liu, R., Zhang, R., & Wang, S. (2025, December). Transformer-based modeling of user interaction sequences for dwell time prediction in human-computer interfaces. In 2025 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST) (pp. 352-356). IEEE.