International Journal on Smart & Sustainable Intelligent Computing
Vol. 03, Iss. 01, S. No. 02, pp. 11--24, March 2026
ISSN (Online): 3048-8508

Received: 02 Feb 2026, Accepted: 23 Feb 2026, Published: 09 April 2026
Digital Object Identifier: https://doi.org/10.63503/j.ijssic.2026.234

Review Article

Reinforcement Learning from Human and AI Feedback for Large Language Model Alignment: A Review

Tanay Chowdhury

Data Science Lead – Gen AI Center of Innovation, Amazon Web Services

Seattle, USA

tanayz@outlook.com

*Corresponding author: Tanay Chowdhury, tanayz@outlook.com

ABSTRACT

Safe and effective deployment of AI requires that large language models (LLMs) generate it in a way that complies with human values and preferences. The application of Reinforcement Learning from Human Feedback (RLHF) has effectively been applied to fine-tune models on human judgment-based categories, enhancing the helpfulness, coherence, and safety. Nonetheless, RLHF suffers certain limitations, such as using high-quality human labels, being costly, slow to iterate, and not being consistent owing to the subjectivity of annotators. The Reinforcement Learning AI Feedback (RLAIF) has become a scalable and effective method of resolving the challenges. The RLAIF provides an opportunity to use AI-generated preferences, revisions, and reward modeling to automatically fine-tune LLMs without violating ethical and safety standards. This will decrease human efforts, enhance reproducibility, and enhance response harmlessness, uniformity, and ethical compliance. Applications of RLAIF have been successful in dialogue generation, summarization, content personalization and automated reasoning. The review summarizes the recent research of feedback-based reinforcement learning, including underlying mechanisms, practical advantages, constraints, and usage of RLAIF. It points out that AI-based feedback offers a systematic and scalable channel of enhancing alignment, robustness and safety of large-scale language models.

Keywords: *Large Language Models, RLHF, RLAIF, Human Feedback, AI Feedback*

1. Introduction

LLMs including GPT, LLaMA, and PaLM have reshaped the NLP domain by training the models to view patterns, semantics, and contextual connections based on the large volumes of data. In the applications of these models, chatbots, content generation, summarization, code assistance, and decision support systems are increasing in their use [1]. LLM increases human-computer interaction by automating complex tasks in language and offers insights in industries [2]. Nevertheless, the fact that they are so widely accepted creates significant issues of conformity to human values, intentions, and ethical standards. The inappropriate models can produce biased, unsafe, or incorrect information leading to trust loss and creating threats in sensitive areas.

Conventional forms of the training of LLM, including supervised learning and large-scale pretraining, are concerned with making text predictions in view of statistical tendencies. These techniques have a limitation in their approach in modelling human preferences because in most cases, they are not as effective in capturing subtle human needs, morality and contextual needs [3]. Models that are solely trained on existing data can thus perform unwanted behavior, and these include biases, inconsistencies,

and unsafe behaviours. To solve such limitations, more mechanisms are needed, which can actively influence the model behavior to the desired results, which are desired by humans.

Reinforcement Learning (RL) provides a method of behavioural improvement of a model by improvement of actions to increase cumulative reward signals [4]. RLHF has been important in the use of fine-tuning of models in LLMs fined-tuned based on human opinions. On ranking outputs and rewarding desired reactions [5], RLHF improves beneficial Ness, security, and conformity to human expectations. Contrarily, RLHF need human labelling of high quality, which may be expensive, time-consuming, and susceptible to annotator bias. These drawbacks limit scalability, slowness of iteration and can lower reproducibility of the alignment process.

Reinforcement Learning of Artificial Intelligence Feedback (RLAIF) has been suggested as an alternative to these difficulties to be scaled. RLAIF RLAIF empowers the fine-tuning of AI-generated feedback and preference models to achieve efficacy in the context of ensuring ethical and safety measures [6], [7]. Such a strategy increases the use of less human labour, and the reproducibility, increases the harmless nature of model outputs, consistency, and overall quality [8]. With the fast growth of the area of research on RLHF and RLAIF, the review is necessary to consolidate the existing methods, contrast the practical advantages, and point out weaknesses and new challenges. This paper includes a review of feedback-based reinforcement learning of LLM alignment with particular emphasis on human and AI-based feedback and its use in the context of enhancing model behavior, safety, and reliability.

The paper is structured in the following way: Section 2 indicates the principles of feedback-driven learning to align LLM. Section 3 explains the workflow, mechanisms and limitations of RLHF and section 4 includes a detailed discussion of RLAIF with a focus on its scalability, efficiency and ethical benefit. Section 4 examines different applications, and section 6 is a literature review of the recent research. Lastly, Section 7 ends the paper with future directions towards safe and effective application of LLMs.

2. Foundations Of LLM Alignment and Feedback-Based Learning

LLMs are inherently connected to data-driven applications, as they are trained on extensive datasets to learn language patterns and semantics, enabling them to perform tasks such as text generation and reasoning [9]. NLP and other data-driven applications are made possible by these models' robust industry-wide language comprehension [10]. New data may be used to fine-tune LLMs, enabling dynamic modification and ongoing development via feedback, increasing their accuracy and personalisation. Furthermore, LLMs facilitate data analysis by drawing conclusions from unstructured data, which promotes creativity and helps in decision-making.

2.1 Classification of Large Language Models

LLM are neural network-based systems that have been trained on extensive volumes of textual data to understand the fundamental structure of human language. Such models are applicable to a wide range of NLP tasks, including text completion, translation, summarisation, and question responding. When employed in NLP tasks that necessitate human-like values and preferences, a large language model must endure a fine-tuning process.

1) *LoRA*

Low-Rank Adaption (Lora) is a technique for fine-tuning LLMs that uses low-rank decomposition to represent the weight updates by two much smaller matrices, known as update matrices, while the pre-trained model weights remain frozen. This has a noticeable impact when LLM sizes increase since it

enables fine-tuning of pre-trained models without storing the same number of parameters in the new model as in the underlying model [11].

2) *PEFT*

The principle upon which Parameter-Efficient Fine-Tuning (PEFT) operates is similar to that of Lora. PEFT stops adjusting most of the pre-trained model's parameters and barely gets close to adding a handful of new ones [12]. This enables less powerful hardware to train bigger pre-trained models while reducing the storage and computing costs of fine-tuning LLMs. Once the fine-tuning is finished, the learnt additional parameters are applied on top of the pre-trained model.

3) *Transformers*

Vaswani et al. (2017) introduced the Transformers neural network architecture [13]. Machine translation, text summarisation, and text classification are just a few examples of the many natural language processing applications that have seen substantial success using transformers as an architecture. The self-attention system, which recognises the connections between words in a phrase, is Transformers' primary invention. The Transformers architecture is used by successful models like T5, BERT, and GPT models, such GPT3.

2.2 Reinforcement Learning for Language Models

Reinforcement Learning (RL) is a framework where an agent would adapt to perform actions so as to optimize cumulative rewards. Conventional RL is based on reward functions that are well-defined, whereas NLP problems are commonly subjective, and it is difficult to define reward functions in such cases. To do so, there have been preference-based reward models developed, in which output of candidates is assessed by human or AI feedback and surrogate reward functions are formed [14].

This RL is applied to the supervised learning of models where the model behavior can be optimized given these reward signals to yield RLHF or RLAIF models. The common pipeline consists of the factor that produces candidate outputs, which are scored with the help of a reward model that is trained on feedback, and the model is updated with the help of a policy optimization approach like Proximal Policy Optimization (PPO) [15]. Fine-tuning using RL models has played an important role in making models such as GPT-3.5 and generate responses that are coherent, helpful, and safe. This method can be used to supplement supervised fine-tuning by offering scalable consistency to human values and preferences.

2.3 Role of Preference Learning in Model Alignment

Preference learning is a key tool that can be used to adjust LLMs to human or AI-specified values and objectives. Preference learning, rather than taking using explicit labels, does seize relative judgments concerning outputs of candidate model, including rankings or ratings, which reflect which responses are more desirable. Reward models are trained with these preference signals and use the estimation to give feedback on the quality of model output and seeks to guide the fine-tuning of reinforcement learning [16]. The model can be further refined to output behavior that is more expected, such as helpfulness, safety, and coherence, through the process of optimizing the LLM.

This Fig. 1 represents the alignment and evaluation pipeline of the feedback-based learning large language models. This starts with the receipt of feedback where various responses of the candidates to a specific instruction are created and evaluated using rankings and ratings by human or AI critics. These preference signals are used to condition a reward model used to steer the LLM towards desirable behaviours and outputs. Similar ranking and scoring criteria are then used to compare the aligned model responses with reference LLM responses. This pipeline shows that preference learning can be used to assist in stronger alignment and systematic analysis of the performance and quality of LLM.

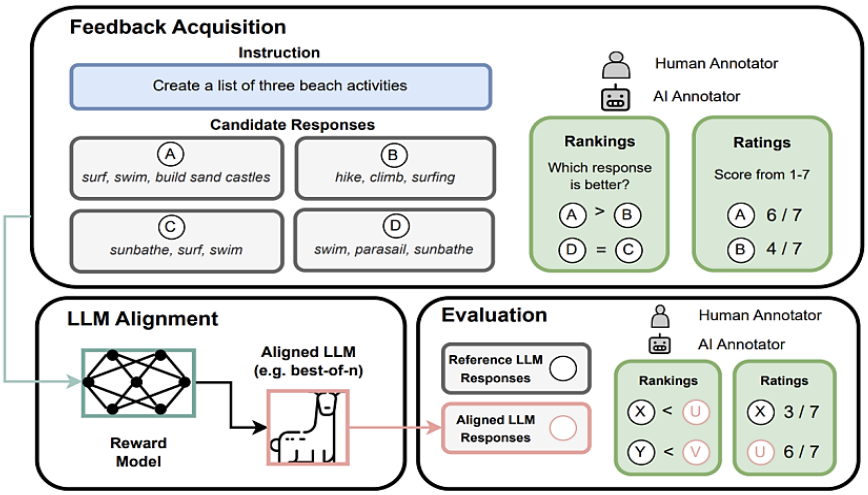

Fig. 1. Alignment and Evaluation Pipeline of Large Language Models (LLMs) [17].

3. Reinforcement Learning from Human Feedback

RLHF has become one of the most popular methods to control LLM to be more aligned to human goals before being deployed. Although it is widely used, models trained by RLHF still have severe shortcomings, such as hallucination of misinformation, sensitive information leakage, biased or forerunner-responses, and vulnerability to pull tricks like jailbreaking and prompt injection [18]. One of the main factors of such inadequacies is in the area of human preference data collection, which is the basis of the RLHF process. Human feedback can be generally acquired by way of pairwise comparisons or ranking of model-generated information regarding subjective evaluation criteria like helpfulness and safety. Such judgments are due to annotator bias, cultural context, task framing, as well as incomplete coverage of an adversarial or long-tail case [19], and result in noisy or incomplete preference signals which can contribute to the spread of misalignment in training.

According to the Fig. 2, the RLHF pipeline starts with human feedback to train a reward model through supervised learning. This reward model then provides scalar rewards to optimize the policy by means of reinforcement learning. The resultant policy is referred to as human-labelled and is repeatedly tested against to create a closed feedback loop. Although successful at enhancing superficial consistency, the capability of preference data and variability limits affect reward modelling and robustness of policies directly, which prevents RLHF in the context of providing reliable and safe behavior in the real world.

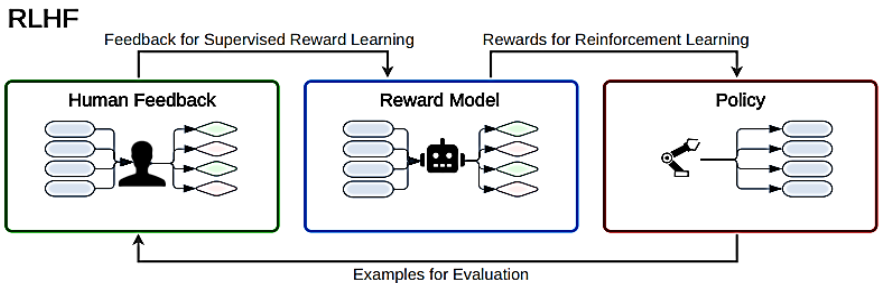

Fig. 2. Reinforcement Learning from Human Feedback [20]

3.1 Reward Model Training

Reward Model (RM) training is an essential part of RLHF in which human preferences are translated into a human learnable reward signal. The system also learns to assess the outputs of models in a manner that is likely to represent human judgment about quality, relevance and even correctness when it is not being assisted by hand-crafted reward functions [21]. During this step, a trained language model will produce many signals to a query. These responses are then compared by human annotators who then classify them as preferred (chosen) or rejected (non-preferred). An example of such a reward model, where one trains a Seq2Seq architecture or transformer architecture, is trained on a pairwise ranking objective or maximum likelihood estimation (MLE). The training promotes the reward model of giving a greater score to the selected responses as compared to rejected responses to the same prompt presented by Fig. 3.

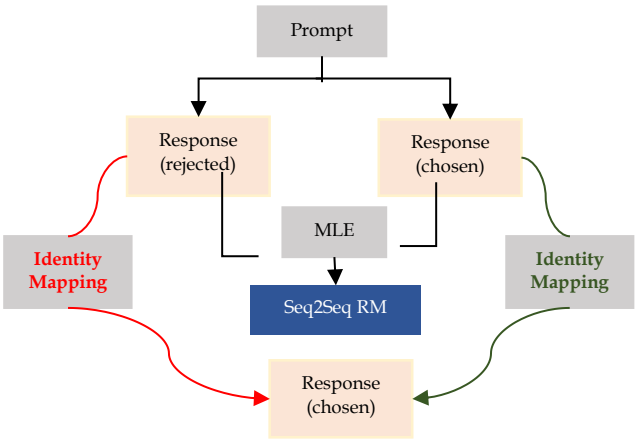

Fig. 3. Reward Modeling in RLHF

Fig. 3 illustrates that one prompt results in multiple responses, both of which are humanly rated as either selected or rejected outputs. The pairs of responses are utilized to construct an MLE-based training procedure to create a reward model Seq2Seq. The trained reward model is trained to give more reward to preferred responses, human preferences which are used in optimizing reinforcement learning.

3.2. Limitations of Human-Centric Feedback

In RLHF, there are several practical and the methodological limitations associated with the use of the human judgements:

- **Scalability Limits:** The process of gathering quality human feedback is go-slow and costly and it is hard to scale to large datasets and complicated tasks [22].
- **Inconsistency and Subjectivity:** Human preferences differ with different individuals and situations and therefore there are noisy, biased or contradictory feedback signals.
- **Annotation Fatigue:** Monotonous ratings: Monotonous rating of the data leads to fatigue, which decreases attention and worsens feedback quality in the long run.
- **Limited Coverage:** Humanity might not predict the existence of edge cases or long-tails, which will lead to model mismatches [23].
- **Bias Propagation:** There is a risk of bias propagation due to the social, cultural or cognitive bias of the annotators being fed into the reward model.

- **Slow Iteration Cycles:** Human in the loop training is a slower type of experimentation and model updating than automated approaches.

4. Reinforcement Learning from AI Feedback

Although there are several limits, RLHF has been a beneficial way for increasing LLM performance, particularly in reducing or stopping the development of unwanted outputs. Scaling up the procedure is quite challenging since the optimum advantage of RLHF requires high-quality human labelling [24]. A method called RLAIF has been put out to remove this bottleneck without sacrificing efficiency.

RLAIF represents a significant advancement in addressing these challenges. This method integrates RL algorithms with feedback from additional AI models (Preference Model (PM)) to facilitate hybrid learning. The RLAIF system employs AI-generated feedback to assist the learning agent make improved judgements [25]. Transitioning from RLHF to RLAIF addresses the issue of restricted human feedback inherent in RLHF. This enhances the effectiveness and scalability of the learning process. Fig. 4 shows the process of RLAIF; an off-the-shelf LLM is used to generate responses, which are rated to generate a reward model (RM). The RM will then lead the RL model to enhance outputs through reinforcing learning.

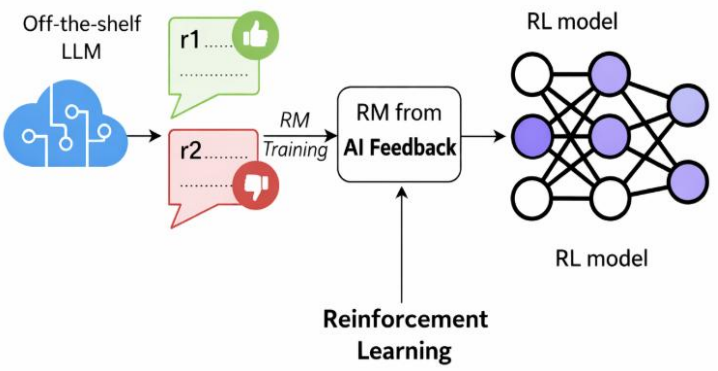

Fig. 4. Reinforcement Learning from AI Feedback

4.2. RLAIF Process Flow

The RLAIF's process are given and also that how it generates and uses Al feedback:

Step 1: Generate Revisions

- **Initial Response Generation:** First, employ a suitable RLHF model to produce responses. Nevertheless, it may occasionally generate detrimental outcomes.
- **Critique and Revision:** Certain standards are used to examine the answer for damaging features, such as unethical or unlawful material. The model then revises the response to remove these damaging features [26].
- **Iterative Process:** Repeat critiquing and revising using random constitutional principles multiple times. This improves the reaction and makes it non-evasive and harmless.

Step 2: Fine-Tune with Revisions

- **Creation of SL-CAI Model:** Develop a pre-trained model through fine-tuning using the datasets of prompts and the revised responses from Step 1. This model is designated as the SL-

CAI (Supervised Learning for Constitutional AI) model. This model serves as the Response Model in the subsequent step and forms the foundation for the final model following the RL phase.

- **Purpose of Fine-tuning:** The SL-CAI model may be fine-tuned to get more accurate results. It also helps reduce the need for later RL training.

Step 3: Generate Harmlessness Dataset

- **Using the SL-CAI Model:** This model produces two responses to prompts that are potentially detrimental. Feedback Model Evaluation: A feedback model assesses responses based on constitutional principles, presented in a multiple-choice format. This process generates a preference dataset, wherein normalised probabilities serve as the scoring metric for each response.

Step 4: Train Preference Model

- **Preference Model Pre-training (PMP):** The PM is prepared for training by being pre-trained on the harmlessness dataset. Data collected to teach Al systems to refrain from responding in an immoral, damaging, or otherwise improper way is referred to as the "Harmlessness Dataset" in the area of AI and ML [27]. This training allows the PM to improve by leveraging knowledge from websites such as Stack Exchange. It is especially beneficial when there is limited information available.
- **Training the PM:** The PM makes advantage of the harmlessness dataset's comparative data. The system is able to prioritise input pairs (prompt/response) by using this feature.

Step 5: Reinforcement Learning

- **Application of Proximal Policy Optimization (PPO):** The SL-CAI model is then trained using Proximal Policy Optimization (PPO). The policy mapping from prompt text to answer text is optimised with the aid of PPO.
- **Using PM as a Reward Signal:** This RL step trains the SL-CAI model using the reward signal from the prior stage, which was the PM's output.

4.2. Benefits of RLAIF

The comparison summarized in Table I, illustrates how RLAIF improves upon RLHF by enhancing harmlessness, ethical alignment, consistency, and scalability, while reducing human bias and subjectivity for more reproducible and efficient training of large AI models.

Table 1. Benefits of RLAIF Over RLHF

Aspect	**RLHF**	**RLAIF**
Interaction Quality	Produces human-like interactions and aligns outputs with human preferences. Effective for conversation, summarization, and general assistance. However, it can occasionally generate harmful or biased responses due to the complexity of human preferences.	Comparable to or exceeding RLHF in tasks such as summarisation and the generation of safe, innocuous dialogue. Ensures helpfulness while enhancing harmlessness, ethical conformity, and the overall safety of responses.

Bias and Subjectivity	Feedback reflects the biases, values, and cultural perspectives of human annotators. There may be discrepancies and a failure to reflect the diversity of human values as a result of this subjectivity.	Minimises subjectivity via the use of AI-generated feedback that is directed by clear guidelines or a predetermined "constitution." Provides more consistent and objective guidance aligned with ethical and safety standards.
Scalability and Cost	The process of gathering high-quality human preference labels is costly, time-consuming, and labour-intensive. There is a limit to scalability, especially for huge and intricate models.	Automates feedback generation with AI, enabling the creation of large-scale datasets with minimal human labor. Enhances scalability and efficiency for training large and complex models.
Reproducibility and Consistency	Human feedback can vary across annotators, time, and context, making results less reproducible and standardization difficult.	AI-generated feedback follows consistent rules and principles, ensuring higher reproducibility across training runs, model updates, and large-scale deployments.

5. Applications Of RLHF and RLAIF

RL is now one of the cornerstones to enhance AI systems, especially in NLP, human-computer interaction, robotics, and content personalization [28]. Two notable versions, RLHF) and RLAIF have different applications in such directions.

5.1. Applications of RLHF

The RLHF is aimed at aligning AI behaviour with human preferences, which is why it is especially effective in relationships that involve delicate understanding, reasoning and situational awareness. Its applications include:

- **Dialogue Generation:** RLHF allows conversational agents to give helpful, safe and context-sensitive replies without being harmful [29].
- **Summarization:** RLHF enhances the quality and coherence of large language model summaries with the help of human feedback so that valuable information can be accurately represented [30].
- **Grammar Error Correction:** RLHF is used by tools such as Trinka AI to provide better grammar correction to non-native English speakers, both in the context of linguistic and situational correctness.
- **Reasoning and Decision-Making:** RLHF boosts reasoning abilities of language models and can support them to be human-oriented assistants that make logical, consistent, and coherent decisions [31].

By making AI-based conversations more natural and intuitive, RLHF contributes to human-computer interaction to a considerable extent. For example, employs the RLHF to preserve the context of the conversation and the customer service chatbots employ it to give relevant, efficient, and the consistent feedback.

5.2. Applications of RLAIF

RLAIF extends the principles of RLHF by using AI-generated feedback, which enables scalable and automated training processes without relying extensively on human annotations. Its applications include:

- **Robotics and Automation:** By analysing data from their surroundings, RLAIF enables robots to improve their autonomous navigation, object handling, and adaptive control capabilities. This makes them capable of dynamic decision-making in complex and uncertain environments.
- **Gaming and Simulations:** In the gaming industry, RLAIF trains intelligent agents that adapt their behaviour based on interactions within virtual worlds, resulting in immersive and challenging gameplay experiences.
- **Personalized Content Recommendations:** Digital platforms implement RLAIF in recommendation systems to continuously refine content suggestions based on user interactions, combining explicit and implicit feedback to enhance engagement and satisfaction [32].
- **Scalable AI Training:** By relying on AI-generated feedback instead of costly human annotations, RLAIF supports large-scale model improvement across various domains [33], from language understanding to autonomous systems.

6. Literature Review

The literature presents RLHF and RLAIF approaches to align LLMs with the human will, solve the problems of robustness and scalability, personalization, and data efficiency, as well as discover difficulties in the modelling of the rewards, human feedback, and multimodal combination.

Shen et al. (2024) note that RLHF which is essential to tailoring LLMs to human preferences is compromised by the weakness of reward models attributed to weaknesses such as errors in human labelling. They suggest that the effectiveness of reward models can be improved by introducing to the existing models a new penalty term known as the contrastive rewards. This procedure comprises of offline sampling to collect baseline responses and a contrastive reward based on the responses, which is incorporated into Proximal Policy Optimization (PPO). The findings suggest that contrastive rewards address reward uncertainty, strengthen, scale based on task difficulty, and dramatically reduce variance which can eventually prove to be better in performance compared to established baselines because of GPTs and human rating [34].

Lee et al. (2023) compare the concept of RLHF and RLAIF. They refer to high-quality human preference labelling as one of the major bottlenecks of RLHF. Their comparison indicates that RLAIF produces similar results to RLHF and, in most cases, (around 70%), the human judges would prefer the result of the two algorithms compared to the performance of a baseline model. Moreover, it was also found to be equally preferred between the RLAIF and RLHF summaries. The findings indicate that RLAIF can perform at a human level, which might solve the problem of RLHF's scalability [35]. Jin et al. (2023) discuss the effectiveness of human feedback implementation in the natural language to enhance large language models (LLMs) such as Falcon-40B-Instruct. With a small dataset of 1000

records or less, containing both critiques and revisions they show that the model is greatly improved in response quality. It is important to note that ChatGPT revised responses won by 56.6% over originals, with this result rising to 65.9% after five rounds of revision, suggesting the possibilities of natural language feedback to make improvements in LLM responses [36].

Ouyang et al. (2022) show how it is possible to align the language models with the intent of the users based on the fine-tuning with human feedback. They collect the data from the labelled prompts and the language model API submissions to the fine-tune GPT-3 through the supervised learning. This is supplemented by human feedback reinforcement learning that is based on rankings in the model output, resulting in the creation of Instruct. Human judgments show that Instruct 1.3B model with much fewer parameters are preferred to the 175B GPT-3 model. Improvements in truthfulness and reduced toxic output have been also recorded in instruct models with little performance regression. The results are that the human feedback fine-tuning of language models is a plausible approach to increased alignment [37].

Bai et al. (2022) show the use of preference modelling and RLHF to finetune language models improves their performance on a wide range of NLP tests and can be trained on specialized skills, including python coding and summarization. Their method presents an online training process based on iteration, training preference models and RL policies each week on new human feedback. The experiment shows that there is a linear relationship between RL reward and square root of KL divergence between the policy initiation. Moreover, the authors perform calibration, competing objectives, and out-of-distribution detection analysis, and compare their models with human writers and offer prompt examples of the same [38].

Amit Kumar Jain (2021) explains the revolutionary role of large-scale language models (LLMs) that benefits personalization to improve user experience in e-commerce, healthcare, education, and entertainment. The recent innovations are new pre-training strategies, fine-tuning approaches, and prompt engineering, which enables LLCs to offer real-time and contextual personalization. As opposed to the old systems, LLCs are highly dynamic and can integrate RLHF so that the responses can correspond to the intentions of the user. Advances in multimodal LLMs allow the processing of text, images, audio, and video, and allow more interactive interactions. The article indicates trends, challenges and future directions in implementing the LLM in personalization, noting constant learning, ethical concerns and fine-tuning adaptations to overcome constraint and promote equality and user trust [39].

Table 2 presents the recent research, summarizing their methods, findings, advantages, challenges, and perspectives in the future to allow the developing a comparative knowledge of the RLHF and AI feedback to achieve effective LLM alignment.

Table 2. Comparative Analysis and Research Gaps in Human and AI Feedback for LLM Alignment

References	**Methods**	**Findings**	**Advantages**	**Challenges**	**Future Work**
Shen et al., (2024)	Introduces contrastive rewards in RLHF; offline sampling + PPO with penalty term	Improves robustness, penalizes reward uncertainty, calibrates task	More stable RLHF, encourages improvement over baseline, better	Sensitive to reward noise; requires baseline sampling	Extend contrastive reward design, test across multiple LLMs and tasks

		difficulty, reduces PPO variance	reward modeling		
Lee et al., (2023)	RL from AI Feedback (RLAIF) vs. RLHF	RLAIF achieves comparable improvements to RLHF on summarization; human evaluators cannot distinguish	Reduces reliance on costly human labels; scalable	Depends on quality of AI feedback; potential bias from LLM labeling	Explore RLAIF in other NLP tasks; hybrid human+AI feedback pipelines
Jin et al., (2023)	Fine-tuning LLMs on natural language human feedback (critiques/revisions)	Small datasets (~1000 records) of natural language feedback can improve strong LLMs; iterative revisions boost win rate	Data-efficient; can refine outputs from existing LLMs like ChatGPT	Limited by quality of critiques; may require multiple iterations	Larger-scale experiments; combining structured and unstructured feedback
Ouyang et al., (2022)	InstructGPT: supervised fine-tuning + RLHF on human-ranked outputs	Smaller model (1.3B) outperforms larger GPT-3 (175B) in human preference; reduces toxicity	Effective alignment with human intent; improves truthfulness	Still prone to simple mistakes; requires human ranking	Apply to multimodal LLMs; improve efficiency of human ranking
Bai et al., (2022)	RLHF with preference modeling; iterated online updates	Alignment improves NLP performance, robust to policy shifts; linear relation between reward and KL divergence	Compatible with specialized skills; robust	Complexity of iterated updates; requires regular human feedback	Explore out-of-distribution (OOD) detection; extend to real-time feedback scenarios
Amit Kumar Jain, (2021)	Review on LLM personalization; RLHF for aligning with human intent; multimodal and federated learning	RLHF improves personalized outputs; multimodal LLMs enhance user experience; federated learning preserves privacy	Real-time, scalable personalization; privacy-preserving	Integration complexity; ethical/fairness concerns	Combine RLHF with multimodal, federated LLMs; continual learning for adaptive personalization

7. Conclusion And Future Work

RL based on feedback is now a fundamental method to enhance the correspondence between large language models and ethical and human-focused goals. Although RLHF is a useful approach to improve model predictions by incorporating the human judgment approach, it suffers the disadvantage of being unscalable, having bias, and being slow in iterating. To deal with those challenges, RLAIF employs AI-generated feedback and allows the automatic and reproducible and efficient fine-tuning of LLMs. This strategy makes it more harmless, more ethical, and more consistent and less reliant on the expensive human annotations. It is demonstrated that RLAIF can be used to run scalable model training in a variety of applications such as dialogue systems, summarization, content personalization and automated reasoning, without reducing the quality of alignments. Although it has several benefits, it has several challenges in developing strong reward models, addressing the biases that may arise as a result of AI and scaling the alignment to multimodal and interactive experiences. Future studies are needed to investigate hybrid pipelines that combine human feedback with AI feedback to make the pipelines more adaptable, widespread application of RLAIF to diverse tasks, and lifelong learning that adapts to the changing needs of the users. Also, it will be necessary to enhance the transparency, interpretability, and ethical governance of reinforcement learning systems based on feedback to facilitate trustful, safe, and socially responsible application of AI.

Conflict of Interest

The authors declare no potential conflict of interest.

References

[1] D. Bill and T. Eriksson, “Fine-tuning a LLM using reinforcement learning from human feedback for a therapy chatbot application,” 2023. :contentReference[oaicite:0]{index=0}
[2] S. Achouche, U. B. Yalamanchi, and N. Raveendran, “Method, apparatus, and computer-readable medium for performing a data exchange on a data exchange framework,” U.S. Patent 10,387,195 B2, 2019.
[3] R. Guha, “Fine-tuning human for LLM projects,” SSRN, 2023.
[4] Y. Dubois et al., “AlpacaFarm: A simulation framework for methods that learn from human feedback,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 36, pp. 30039–30069, 2023.
[5] V. Pal, “Bias detection and mitigation in foundation AI models: A human-centric approach,” TIJER – Int. Res. J., vol. 8, no. 2, pp. 1–7, 2021.
[6] S. K. Chintagunta, “AI in code, testing, and deployment: A survey on productivity enhancement in modern software engineering,” Int. J. Res. Anal. Rev., vol. 10, no. 4, pp. 747–752, 2023.
[7] S. Thangavel, S. Srinivasan, S. B. V. Naga, and K. Narukulla, “Distributed machine learning for big data analytics: Challenges, architectures, and optimizations,” Int. J. Artif. Intell. Data Sci. Mach. Learn., vol. 4, no. 3, pp. 18–30, Oct. 2023, doi: 10.63282/3050-9262.IJAIDSML-V4I3P103.
[8] H. R. Kirk, A. M. Bean, B. Vidgen, P. Röttger, and S. A. Hale, “The past, present and better future of feedback learning in large language models for subjective human preferences and values,” arXiv preprint arXiv:2310.07629, 2023.
[9] C.-A. Cheng, A. Kolobov, D. Misra, A. Nie, and A. Swaminathan, “LLF-Bench: Benchmark for interactive learning from language feedback,” arXiv preprint arXiv:2312.06853, 2023.
[10] D. Patel, “AI-enhanced natural language processing for improving web page classification accuracy,” ESP J. Eng. Technol. Adv., vol. 4, no. 1, 2024, doi: 10.56472/25832646/JETA-V4I1P119.
[11] E. J. Hu et al., “LoRA: Low-rank adaptation of large language models,” in Proc. Int. Conf. Learn. Representations (ICLR), 2022.
[12] S. Paul and S. Manglukar, “PEFT: Parameter-efficient fine-tuning of billion-scale models on low-resource hardware,” Feb. 2023.

[13] Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.
[14] Y. Du et al., “Guiding pretraining in reinforcement learning with large language models,” in Proc. Int. Conf. Mach. Learn. (ICML), 2023, pp. 8657–8677.
[15] T. Carta et al., “Grounding large language models in interactive environments with online reinforcement learning,” in Proc. Int. Conf. Mach. Learn. (ICML), 2023, pp. 3676–3713.
[16] S. Huang, J. Zhao, Y. Li, and L. Wang, “Learning preference model for LLMs via automatic preference data generation,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 9187–9199.
[17] H. Bansal, J. Dang, and A. Grover, “Peering through preferences: Unraveling feedback acquisition for aligning large language models,” arXiv preprint arXiv:2308.15812, 2023.
[18] R. Zheng et al., “Secrets of RLHF in large language models part I: PPO,” arXiv preprint arXiv:2307.04964, 2023.
[19] G. Sarraf, “DeepDefender: High-precision network threat classification using adversarial-resistant neural networks,” Int. J. Adv. Res. Sci. Commun. Technol., vol. 2, no. 1, pp. 596–606, 2022, doi: 10.48175/IJARSCT-3600E.
[20] S. Casper et al., “Open problems and fundamental limitations of reinforcement learning from human feedback,” arXiv preprint arXiv:2307.15217, 2023.
[21] M. Bakker et al., “Fine-tuning language models to find agreement among humans with diverse preferences,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 35, pp. 38176–38189, 2022.
[22] R. Kanchana, K. Phusavat, Z. Pastuszak, A. N. Hidayanto, and J. Majava, “Effects of external feedback on disengagement in a human-centric environment,” Hum. Syst. Manag., vol. 41, no. 6, pp. 685–697, 2022.
[23] J. Abramson et al., “Improving multimodal interactive agents with reinforcement learning from human feedback,” arXiv preprint arXiv:2211.11602, 2022.
[24] S. Höglund and J. Khedri, “Comparison between RLHF and RLAIF in fine-tuning a large language model,” 2023.
[25] K. M. R. Seetharaman and S. Pandya, “Importance of artificial intelligence in transforming sales, procurement, and supply chain processes,” Int. J. Recent Technol. Sci. Manag., vol. 8, no. 7, pp. 1–9, 2023.
[26] V. Gallego, “ZYN: Zero-shot reward models with yes-no questions for RLAIF,” arXiv preprint arXiv:2308.06385, 2023.
[27] G. K.-M. Liu, “Transforming human interactions with AI via reinforcement learning with human feedback (RLHF),” Massachusetts Institute of Technology, 2023.
[28] R. Saxena, S. A. Pushkala, and R. Carvalho, “Systems and methods for rapid processing of file data,” U.S. Patent 9,594,817, Mar. 2017.
[29] H. P. Kapadia, “Generative AI for real-time conversational agents,” Int. J. Curr. Sci., vol. 13, no. 3, pp. 201–208, 2023.
[30] M. Abdullah, A. Madain, and Y. Jararweh, “ChatGPT: Fundamentals, applications and social impacts,” in Proc. 9th Int. Conf. Social Netw. Anal., Manag. Security (SNAMS), IEEE, 2022, pp. 1–8, doi: 10.1109/SNAMS58071.2022.10062688.
[31] V. Verma, “Security compliance and risk management in AI-driven financial transactions,” Int. J. Eng. Sci. Math., vol. 12, no. 7, pp. 1–15, 2023.
[32] Y. Bai et al., “Constitutional AI: Harmlessness from AI feedback,” 2022.
[33] S. Garg, “AI-driven innovations in storage quality assurance and manufacturing optimization,” Int. J. Multidiscip. Res. Growth Eval., vol. 1, no. 1, pp. 143–147, 2020, doi: 10.54660/IJMRGE.2020.1.1.143-147.
[34] W. Shen et al., “Improving reinforcement learning from human feedback using contrastive rewards,” 2024.
[35] H. Lee et al., “RLAIF: Scaling reinforcement learning from human feedback with AI feedback,” 2023.
[36] D. Jin et al., “Data-efficient alignment of large language models with human feedback through natural language,” 2023.
[37] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2022.

[38] Y. Bai et al., “Training a helpful and harmless assistant with reinforcement learning from human feedback,” 2022.
[39] K. Jain, “Advancements in large-scale language models for personalization,” Int. J. Comput. Technol. Electron. Commun., 2021.