Alignment Methods for Large Language Models Based on Human Feedback

Nianlin Li *

School of Sports Engineering, Beijing Sport University, Beijing, 100084, China

* Corresponding Author Email: 2024011960@bsu.edu.cn

Abstract. In recent years, artificial intelligence has developed rapidly, and large language models have been widely used. However, as the capabilities of the model continue to improve, there is a risk that its outputs may contain inaccurate information, be misleading, or deviate from human values. To ensure that the model is safe, reliable, and adheres to ethical standards, it is particularly crucial to guide the model's behavior using human feedback alignment techniques. This article systematically organizes relevant alignment methods, clarifies the core connotation of human feedback, further analyzes the key links in the alignment process, categorizes and discusses based on the principles and implementation technologies of the methods, organizes commonly used datasets and evaluation systems, and finally provides a summary. The significance of this research lies in constructing a complete knowledge system, clarifying the technical implementation path, and providing theoretical support and guidance for large language models to better align with human intentions and be safely applied in critical areas.

Keywords: Human Feedback, Large Language Models, Alignment, Reinforcement Learning.

1. Introduction

In recent years, there have been significant breakthroughs in the field of artificial intelligence, with Large Language Models (LLMs) such as ChatGPT, Gemini, and DeepSeek rapidly developing. These models are trained on vast amounts of data, demonstrating exceptional natural language processing capabilities. They have reached human average levels in various fields such as mathematical reasoning and programming, and they are widely used in scenarios like information retrieval and creative generation, significantly advancing the innovation of language processing technology [1]. However, as the performance of models continues to improve, issues such as the accuracy of their output and ethical compliance are becoming increasingly prominent. Ensuring that model behavior aligns with human expectations has become a key challenge.

In order to address the above challenges, human feedback alignment technology has emerged. Before delving into the technology, it is essential to first understand the core meaning of human feedback. Human feedback is essentially the evaluation of behavior and results regarding model outputs, specifically manifested in preference selection, correctness judgment, and ethical recognition or questioning [2]. As a key basis for model optimization, it conveys values and provides important references for correcting biases. Human feedback can be classified into natural language feedback expressed in text and scalar feedback presented in numerical values, levels, or rankings[2]; by scenario, it can be divided into static feedback that is offline pre-labeled and does not change with interaction, and dynamic feedback generated in real-time interactions; by visibility, it can be categorized into explicit feedback that is directly expressed without inference and indirect behavioral signals that need to be inferred through user behavior analysis.

Based on a systematic understanding of human feedback, human feedback alignment technology reconstructs the model's decision-making mechanism, ensuring that the outputs not only maintain high accuracy but also adhere to ethical principles, reduce bias, and align with social values and ethical standards [3]. Its core lies in collecting human evaluations of model outputs, integrating feedback into the training process as a basis for optimization, thereby adjusting the model's behavioral logic to align with human expectations. The necessity of human feedback alignment stems from the fact that the authority over value judgments and ethical standards belongs to human society. Human

direct ratings or implicit behavioral responses convey true preferences, standards of right and wrong, and moral boundaries, providing a reliable basis for model optimization.

This article systematically explores relevant alignment methods, analyzing the technical pathways, alignment method systems, and data sets and evaluation systems from three dimensions, to outline theoretical frameworks and technical systems, aiming to provide a reference for improving the reliability and compliance of model outputs.

2. Core Process of Alignment Based on Human Feedback

Sorting out the core process of human feedback alignment can provide an intuitive and systematic perspective for understanding the mechanism by which large language models achieve alignment with human expectations based on human feedback, as illustrated in Figure 1.

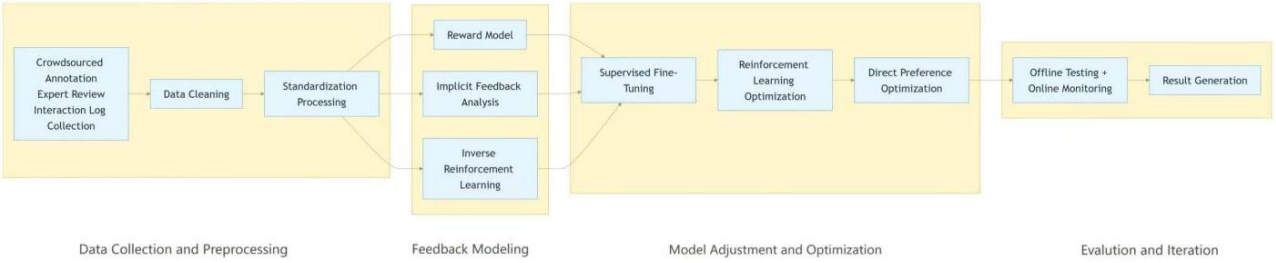

Figure 1. Alignment process based on human feedback (Picture credit: Original)

2.1. Feedback Collection and Preprocessing

To ensure data comprehensiveness, feedback collection needs to integrate multiple channels, specifically by completing large-scale basic annotations through crowdsourcing platforms, relying on expert reviews to ensure the quality of feedback in specialized fields, and extracting valuable user behavior data from real-time interaction logs of the model [3]. The preprocessing stage requires systematic handling of raw data, including cleaning invalid annotations, duplicate content, and extreme bias evaluations, as well as standardization operations, such as converting natural language evaluations into structured labels and normalizing scores with different dimensions, laying a high-quality data foundation for subsequent modeling.

2.2. Feedback Modeling

This section unfolds from two aspects: first, constructing a reward model to quantify the value of feedback, training a scoring model based on labeled data so that it can automatically generate ratings aligned with human preferences; second, for implicit feedback such as click-through rates and dwell time, using methods like behavior sequence analysis and attention modeling to infer user preferences, thereby capturing needs not covered by explicit evaluations and achieving comprehensive utilization of feedback. In addition, by using inverse reinforcement learning techniques to infer the reward mechanisms behind human behavior, the dimensionality of modeling is further expanded.

2.3. Model Tuning and Optimization

To improve model performance, it is necessary to adopt a multi-stage optimization strategy based on feedback analysis results. In the early stages, high-quality feedback data can be directly incorporated into the training process through supervised fine-tuning, thereby correcting fundamental biases; The core phase adopts reinforcement learning methods (such as the PPO algorithm), using reward model scores as the optimization goal, and continuously improving generation quality through policy iteration; In recent years, direct preference optimization techniques have gradually emerged, this method can directly optimize model parameters based on preference data, not only simplifying the process but also effectively reducing the bias transfer problem caused by using reward models[3].

2.4. Evaluation and Iteration

Evaluating the alignment effectiveness requires a combination of offline testing and online monitoring, with a focus on the model's performance in terms of usefulness, honesty, and harmlessness [4]. According to the assessment results, the dataset can be dynamically updated, additional samples can be added for weak links, and corresponding adjustments can be made to modeling and optimization strategies. By constructing a closed-loop mechanism of 'collection-modeling-optimization-evaluation', we can continuously improve the matching accuracy between the model and human values, promoting sustained progress in the safety and reliability of large language models.

3. Classification and Principles of Human Feedback Alignment Techniques

Human feedback alignment technology is a key method for achieving precise matching of AI systems with human intentions and values, by analyzing human feedback signals on model output to optimize model behavior. Based on the presentation form and acquisition method of the feedback, it can be divided into the following three categories.

3.1. Based on the Method of 'Clear Feedback'

The method based on 'clear feedback' takes the preference judgments expressed directly by humans as the core optimization basis, converting subjective evaluations into quantifiable training objectives. The basic principle adopts a closed-loop process of 'human preference-reward signal-model update': the model generates diverse candidate outputs, which are compared or scored by humans, training the reward model based on feedback, and ultimately adjusting model parameters through reinforcement learning, as shown in Figure 2. This mechanism ensures that the model continues to learn and aligns with human preferences, optimizing output quality.

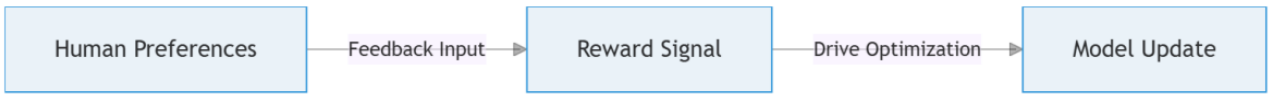

Figure 2. Basic principles of clear feedback (Picture credit: Original)

In typical technologies, Reinforcement Learning from Human Feedback (RLHF) employs a three-stage process, namely supervised fine-tuning, reward model training, and reinforcement learning optimization. Adjusting output strategies based on human preference data significantly enhances the consistency between model outputs and human preferences [3, 5]. Direct Preference Optimization (DPO) simplifies the reinforcement learning process by directly using human preference data to train the model, reducing computational costs while maintaining performance. This method optimizes the model by maximizing the probability of outputs that align with human preferences, avoiding the complex steps of traditional reinforcement learning [6]. Optimization methods based on learning to rank utilize human ranking information of multiple model outputs for training, aiming to optimize the ranking loss function so that the model generates outputs that align better with human preferences [7].

Some studies also explore preference optimization methods based on contrastive learning, by constructing positive and negative sample comparisons, allowing the model to learn the deeper patterns of human preferences through contrast, thereby achieving a more efficient alignment with human needs [8].

3.2. Methods Based on 'Implicit Feedback'

Methods based on 'implicit feedback' do not rely on direct evaluation labels, but instead infer potential preferences by analyzing indirect signals such as the behavioral trajectories in the interaction between humans and models. Its principle utilizes big data analysis and behavioral modeling techniques to extract effective features from unstructured interaction data, capturing the implicit value

tendencies in human behavior, which are then transformed into training signals, as shown in Figure 3. This method can fully utilize the massive amounts of data generated during natural interaction processes to achieve continuous learning about human preferences.

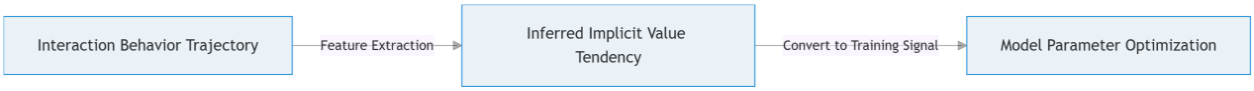

Figure 3. The basic principle of implicit feedback (Picture credit: Original)

Typical techniques include behavioral cloning and inverse reinforcement learning. Behavioral cloning trains models by imitating human demonstration behavior, learning human-edited dialogue responses in dialogue systems. Inverse reinforcement learning combines imitation learning and reinforcement learning, deducing the underlying reward function from human or expert behavior trajectories and optimizing the model's decision-making strategy using reinforcement learning algorithms[9].In addition, collaborative filtering and preference prediction models based on user metrics such as click volume and duration of stay are also widely used, for example, mining implicit preferences through ShareGPT conversation logs and Stack Overflow voting data.

Some studies explore implicit feedback analysis methods based on attention mechanisms, attempting to more precisely infer users' deep preferences by analyzing behavioral patterns such as clicks, dwell times, and interaction sequences [10]. Methods that combine federated learning with implicit feedback are also gradually emerging, achieving data sharing and model training while protecting privacy [11].

3.3. Mixed Feedback Method

The hybrid feedback method combines the advantages of explicit and implicit feedback, achieving more precise alignment through multi-source data fusion. The core principle is to construct a multi-layer fusion framework: using explicit feedback to calibrate core output targets, supplementing fine-grained preference information with implicit feedback, dynamically adjusting weights to resolve feedback conflicts, and forming a complementary enhancement of training effects, as shown in Figure 4. This method balances feedback accuracy and data richness, achieving good alignment in complex scenarios.

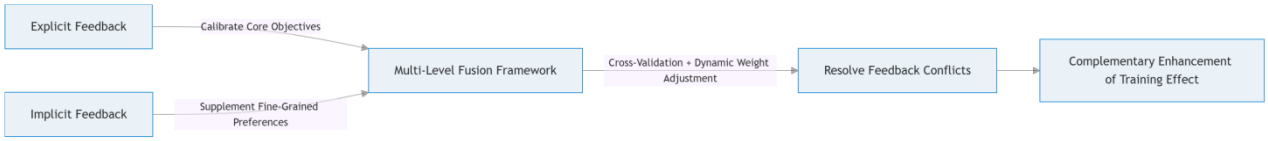

Figure 4. The basic principle of mixed feedback (Picture credit: Original)

Typical technologies include multimodal feedback fusion models and adaptive weight adjustment mechanisms. Multimodal feedback fusion combines human explicit ratings with interaction behavior data for joint modeling. The adaptive weight adjustment mechanism dynamically allocates feedback weights based on the task scenario, increasing explicit feedback weights in safety criterion scenarios and enhancing the proportion of implicit feedback in open domain interaction scenarios.

In research, end-to-end hybrid feedback fusion models based on deep learning have received attention, as they can learn feature representations of different feedback and adjust weights through attention mechanisms to achieve efficient feedback fusion. The mixed feedback method that combines domain knowledge has also been developed, using expertise to guide training and improve alignment accuracy in specific fields. The combination of online learning and mixed feedback enables the model to process feedback in real time, achieving dynamic optimization.

4. Human Feedback Datasets and Evaluation Systems

4.1. Typical Human Preference Dataset

The human preference dataset covers both explicit and implicit feedback types, and Table 1 provides a brief introduction to each dataset. As can be seen from the table, the explicit feedback dataset is characterized by precise annotations and clear preference signals, widely used in summary model training, dialogue alignment, and multi-dimensional preference learning [3]; while the implicit feedback dataset originates from natural interaction behaviors, reflecting user preferences through indirect signals, commonly used in dialogue pattern learning and long-text question answering support. These two types of datasets jointly promote advancements in alignment technology and provide an important data foundation for model training and evaluation.

Table 1. Typical Human Preference Dataset

Category	Dataset Name	Characteristics	Data scale	Application Scenario
Explicit feedback dataset	Summarize from Feedback	In the field of text summarization, manual binary selection is used for optimization	179,000 data samples	Train a summary reward model to match human preferences.
	HH - RLHF	Choose one of the dialogue scenarios, focusing on 'harmlessness and usefulness'	About 160,000 human preference data points	Optimize the harmlessness / usefulness of the model, aligning with human values.
	UltraFeedback	250,000 conversations, preference labeling, finely graded scoring across 4 dimensions, 17 models selected to generate replies.	250,000 dialogues preference labeling	Multi-dimensional preference learning to enhance instruction/model diversity.
Implicit feedback dataset	ShareGPT conversation log	User's interaction history with AI, including interaction behavior trajectory.	Over 100,000 dialogue tracks	Learn natural dialogue patterns to optimize model interaction alignment.
	Web Q&A interaction data (such as ELI5)	Reddit Q&A forums sorted by upvotes indirectly reflect preferences.	ELI5 contains over 50,000 Q&A	Optimize the alignment of long text abstraction problem answers by borrowing like ranking.

4.2. Alignment Evaluation Method Based on Human Feedback

In terms of content alignment, the evaluation focuses on authenticity and harmlessness. The former assesses factual accuracy through 'TruthfulQA', while the latter tests the model's response to malicious prompts using 'RealToxicityPrompts' [2]. Behavioral-level evaluation covers multiple dimensions, including the instruction-following ability assessed through the 'MT-Bench', as well as the context consistency measured using benchmarks like 'LongBench'; additionally, it systematically includes assessments of discriminatory biases (such as Winogender) and ethical considerations (such as the ETHICS benchmark) [2].

In terms of evaluation methods, manual evaluation typically relies on professional annotators to score, rank, or conduct qualitative analysis. Although this approach can effectively capture complex dimensions such as text fluency, it has practical limitations including high cost, low efficiency, and difficulty in ensuring scoring consistency [2, 3]. In contrast, automated evaluation primarily utilizes pre-trained models or similarity calculations to achieve efficient assessment [2, 3]. Common metrics include accuracy and BLEU scores, and they can be specifically categorized into single-answer evaluation, comparative evaluation, and reference-answer evaluation. However, this method still lacks stability and often requires improvement through strategies such as position swapping and multi-evidence generation [3]. Therefore, to balance assessment efficiency and result accuracy, a

hybrid strategy of 'automated preliminary screening combined with manual review' is commonly used in practice.

5. Conclusion

Human feedback alignment technology effectively guides AI models to achieve a dynamic balance between capability and safety by accurately capturing human preference signals, building a technical system centered on explicit, implicit, and hybrid feedback. This technology has established a complete closed loop from data collection to multidimensional evaluation, significantly enhancing the practicality, safety, and compliance of models in applications such as dialogue, becoming a key pathway to instill human values into AI systems and promote the sustainable development of artificial intelligence. Despite its significant effectiveness, the technology still faces multiple challenges, including the subjective bias of human feedback and the high cost of labeling, as well as the limited generalization ability of reward models and the unstable performance of reinforcement learning-based policy models in practical deployment.

Future efforts should focus on technologies for obtaining low-cost feedback, actively promote methods such as RLAIF and SALMON in combination with debiasing techniques; drive technological innovations in dynamic reward models, self-supervised learning, and multimodal modeling; meanwhile, strengthen safety and ethical construction, enhance model interpretability, establish dynamic datasets and interdisciplinary supervision mechanisms, and systematically ensure the robustness and controllability of technological development. Through these initiatives, promote technology to develop in a direction that is smarter, more robust, and more universal, thereby better serving humanity.

References

[1] Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv preprint arXiv: 2303.12712, 2023.

[2] Zhang Yuying, Yun Jing, Liu Xueying, et al. A review of feedback-based methods for aligning content and behavior of large language models. Computer Engineering and Applications, 2025, 1 - 37. https: //link-cnki-net-s.v.bsu.edu.cn/urlid/11.2127.tp.20250522.1435.011.

[3] Liu Kunlin, Qu Xinji, Tan Fang, et al. A survey of alignment research in large language models. Telecommunications Science, 2024, 40 (6): 173 - 194.

[4] Askell A, Bai YT, Chen AN, et al. A general language assistant as a laboratory for alignment. arXiv preprint arXiv: 2112.00861, 2021.

[5] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 2022 (35): 27730 - 27744.

[6] Muldrew W, Hayes P, Zhang M, et al. Active preference learning for large language models. arXiv preprint arXiv: 2402.08114, 2024.

[7] Xiong Deyi. Research on information retrieval models based on ranking learning algorithms. Chongqing University of Posts and Telecommunications, 2018. DOI: 10.27675/d.cnki.gcydx.2018.000554.

[8] Yang C, Xie L. Data augmentation and contrastive learning based on large language models. Advances in Engineering Innovation, 2025, 16 (7): 109 - 118.

[9] Chen Xiliang, Cao Lei, He Ming, et al. A review of research on deep inverse reinforcement learning. Computer Engineering and Applications, 2018, 54 (5): 24 - 35. DOI: 10.3778/j.issn.1002 - 8331.1711 - 0289.

[10] Zhu Jinghua, Guo Xu, et al. Neural network recommendation model based on user vector representation and attention mechanism. In: 2018 National High-Performance Computing Academic Annual Conference; China Computer Society, 2018.

[11] Minto L, Haller M, Haddadi H, et al. Stronger privacy for federated collaborative filtering with implicit feedback. 2021. DOI: 10.48550/arXiv.2105.03941.