Dean&Francis
ISSN 2959-6157

The Comparison of Alignment Fine-Tuning Methods for Large Language Models on Honesty, Helpfulness, and Harmlessness

Haodong Huo

College of Computing, City University Of Hong Kong, Hong Kong, HKG, China
haodonhuo2-c@cityu.edu.hk

Abstract:

The alignment of large language models (LLMs) with human values of Honesty, Helpfulness, and Harmlessness (HHH) is a critical issue for their safe deployment. However, a comparative analysis across these three dimensions remains lacking for primary alignment approaches such as Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF), and Direct Preference Optimization (DPO). This paper aims to conduct a comparative study of the performance of these alignment methods on the HHH dimension through a review and synthesis of relevant research literature. The results indicate that there are performance trade-offs between the different alignment algorithms. In particular, SFT is the foundation of the alignment process, but is generally outperformed by preference-based methods. RLHF excels in usefulness and harmlessness, but involves greater implementation complexity. DPO matches RLHF in terms of usefulness in a cleaner and more efficient way and has an advantage in honesty, but is more sensitive to dataset quality and slightly weaker in harmlessness.

Keywords: Alignment Tuning, Honesty, Harmlessness, Helpfulness, RLHF, DPO

1. Introduction

In recent years, large language models (LLMs) show strong skill in handling many kinds of work and have been used in a wide range of fields. Still, what they produce should not only work well but also fit with human values and adhere to safety standards. To achieve this, alignment tuning has been proposed to mitigate issues like the generation of false, harmful, or toxic content. This process typically follows the HHH (Honesty, Helpfulness, and Harmlessness) principle,and utilizes various approaches, including supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), and direct preference optimization (DPO). Therefore, this paper analyzes the performance and trade-offs of different alignment methods under the HHH principle, exploring their principles, algorithmic processes, and respective

ISSN 2959-6157

strengths and weaknesses, while examining their performance and balance across the three HHH dimensions. Through a literature review and case study analysis, it compares existing methods, summarizes performance differences and applicable scenarios, offers guidance for selecting appropriate alignment algorithms, and supports the advancement of language models that are safer, more dependable, and aligned with human values. The choice of alignment methods should consider specific application scenarios, and this study offers references and insights to enhance the safety and human-value compatibility of large-scale language models.

2. Overview of Alignment Fine-Tuning Methods

Since large language models (LLMs) are trained solely by predicting the next word, their behavioral objectives may diverge from human intent. Therefore, fine-tuning is necessary to improve model performance in specific tasks and to align their behavior with human values. Typically, the fine-tuning process consists of two main stages: First, instruction tuning, which aims to enhance the model's general capabilities by teaching it to understand and execute diverse human commands. Second, alignment tuning, which focuses on shaping the model's behavior to ensure its outputs are consistent with human values and ethical standards.

2.1 Principles and Core Functions of Fine-Tuning

Although LLMs have demonstrated remarkable capabilities in handling many tasks, these models can produce unnecessary content, such as generating harmful or toxic responses that misalign with human values (such as race discrimination or guidance for a crime) generating factually incorrect or unreasonable responses ("hallucinations").

Building on foundational AI safety research by Leike et al., the primary goal of alignment tuning is to obtain models that are aligned with human values [1]. Based on this, institutions like Anthropic that focused on further advance model alignment research proposed and adopted the three core standards of Helpful, Honest, and Harmless (HHH) to evaluate effectiveness of the alignment.

Among these, Honesty requires the model to faithfully reflect its knowledge boundaries, avoiding the fabrication of facts or hallucinations. Helpfulness means the model must accurately and efficiently fulfill user intent, providing valuable information and services. Harmlessness emphasizes that the model must refuse to generate unsafe, unethical, or toxic content, ensuring the safety of the response. The HHH framework proposed by Anthropic provides concrete standards for assessing the degree of model alignment and has become a crucial principle guiding the ethical and safe development of modern large language models.

2.2 Main Alignment Tuning Approaches

Current mainstream alignment tuning methods can be divided into three main categories, which share an iterative and developmental relationship: Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF), and Direct Preference Optimization (DPO). Among these, SFT typically serves as a foundational step for both RLHF and DPO, providing the initial model policy.

2.2.1 . Supervised Fine-Tuning (SFT)

As a foundational step, supervised fine-tuning (SFT) paves the way for more advanced alignment techniques like RLHF and DPO. The core idea of SFT is imitation learning. This method involves training a model on a high-quality dataset consisting of prompt-response pairs crafted by human labelers to model desired behavior. Importantly, the model updates its internal parameters (weights) during training to boost performance on this task. However, the alignment effect of SFT is heavily dependent on data quality; any biases or low quality present in the dataset will be directly learned and reflected in the model's behavior.

2.2.2 . Reinforcement Learning from Human Feedback (RLHF)

The RLHF is a more complex and powerful technology. Its core is to use reinforcement learning to enable a LLM to learn and adhere to human preferences and values, maximizing the generation of responses preferred by humans. Basically, according to Zhao et al., the RLHF system consists of three key components: a language model (policy) using a SFT model as the initial policy, a reward model

(RM) that learns human preferences, and a reinforcement learning algorithm for training the language model.

In particular, a reward model (RM) is built. This is done by providing prompts to the SFT model to generate multiple responses. Human annotators then rank or compare these responses to collect preference data. This preference data is used to train the RM, thereby enabling it to mimic human preferences, automatically assess the quality of any prompt-response pair, and assign a reward score. Besides, the Proximal Policy Optimization (PPO) algorithm optimizes the policy model. The policy model receives a prompt and generates a response, which is fed into the RM to generate a reward signal. The PPO algorithm uses this reward signal to update the policy model's parameters, aiming to maximize the reward achieved by the policy. A KL divergence penalty is also introduced to limit the deviation of the policy from the original SFT model, maintaining the coherence and naturalness of the generated text. For example, OpenAI uses the existing GPT-3 to apply the RLHF and finally gets the InstructGPT. However, RLHF faces many challenges, including high human labeling costs, "reward hacking" caused by possible defects in the reward model, and inherent stability issues in the reinforcement learning process.

2.2.3 . Direct Preference Optimization (DPO)

While RLHF is effective for alignment tuning, its inherent problems of complex workflow, unstable training, and high computational cost have resulted in the emergence of new alignment methods, the most representative of which is DPO. Proposed by Rafailov et al. at Stanford University in 2023, it distinguishes itself by avoiding explicit reward model training and complex reinforcement learning. Instead, it achieves alignment by applying a single loss function to the policy model that directly optimizes the preference probability.

The method typically starts with an initial policy, often obtained through SFT. And a preference dataset is then collected in the RLHF format, comprising pairs of winning responses (preferred) and losing responses (rejected). The final stage is direct optimization training of the DPO. In this phase, the weights of the policy model are updated via backpropagation using a novel loss function derived from alignment target theory. This loss function encourages the model to favor winning responses while suppressing losing responses, effectively turning the challenging preference learning task into a tractable and stable maximum likelihood estimation problem. However, this approach still relies on high-quality human preference datasets, and in theory the alignment effect may be limited by its simplified objective function.

3. Evaluating HHH Metrics for Alignment Fine-Tuning Methods

3.1 Honesty Evaluation

Honesty requires that model outputs be factually accurate and avoid fabrication (i.e., hallucinations). Furthermore, it requires that the model acknowledges its knowledge boundaries and uncertainties [2,3].

Evaluating honesty is a challenging problem, since it is difficult to directly observe the model's internal state. Current evaluation methods primarily compare the model's responses with known correct answers to determine their authenticity. Due to the relative scarcity of dedicated research on honesty evaluation, evaluation methods are still under development. For example, in an OpenAI study, honesty evaluation relied on the TruthfulQA dataset, which is specifically designed to test the truthfulness of LLMs when answering questions that are prone to false confidence. In the TruthfulQA evaluation, answers are considered truthful if they contain no false statements. The evaluation employs a 0-1 scale, where 1 denotes maximal alignment with the truth. Scores above 0.5 are regarded as true, consistent with the standard experimental methodology for this dataset. The formula of the truthful score for a model is as follows:

$$T = \frac{\#Truthfulanswers}{total\#ofquestions} \tag{1}$$

A model's tendency to hallucinate is also used as a measure of honesty. Another research team (GAIR) proposed a new evaluation method for calculating honesty scores using a prudence score and an overconservative score, but this method has not yet been widely used [4]. When comparing SFT and RLHF, it is evident that RLHF generates responses that align with human preferences, sometimes appearing more general or abstract; whereas SFT tends to directly regurgitate training data. While RLHF generally outperforms SFT on QA tasks, they are more likely

ISSN 2959-6157

to acknowledge their limitations in uncertain situations, meaning their answers are more likely to be truthful (but potentially less informative) than confidently misleading. A comparison by Ivison et al. showed that, on the same test set, DPO achieved a 2.5-point higher authenticity score (out of 100) than PPO, indicating that DPO tends to be more truthful [5].

3.2 Helpfulness Evaluation

Basically, according to Askell et al., the helpfulness demands that the AI should make a clear attempt to perform the task or answer the question posed (if this isn't harmful). It should do this as concisely and efficiently as possible [2].

To quantify this concept, most researchers employ human labeler's preference ratings as primary metrics, which directly reflects how humans are satisfied with the response of a model. Other researchers like Rafailfov et al. also used the standardized dataset, such as Anthropic-HH dataset, HH-RLHF, to measure the helpfulness of a model [6,7]. According to OpenAI's research, RLHF is a model that labelers prefer over SFT and is less likely to completely fail to follow the correct instructions, meaning that RLHF is more helpful than SFT [8]. Also, Xu et al. shows that PPO and DPO are much more preferred than SFT model outputs [9]. As for the comparison between the DPO and RLHF, DPO's performance might be influenced by the base model. For example, Rafailfov et al. showed that when using a fine-tuned model on the IMDB dataset (like GPT-2), DPO is notably more efficient and outperformed RLHF on Anthropic-HH dataset, which showed that a simpler algorithm can match or exceed the performance of a more complex one under certain conditions [6]. A study by Xu et al. demonstrated that DPO performed poorly when the base model of the DPO is Llama2-7B fined-tuned by a different dataset called Alpaca [9]. This might be due to the scarcity of the high-quality data of the dataset.

3.3 Harmlessness Evaluation

Based on Askell et al., harmlessness requires AI models to avoid generating any discriminatory or harmful responses, whether direct statements, suggestive content, or implicit systematic biases [2]. Harmlessness evaluation involves multiple dimensions and depends on both human feedback and quantitative metrics, with human assessment remaining the central method. For example, OpenAI relies on human annotators to directly review the safety of model outputs, determining whether they contain inappropriate content such as sex or violence. Quantitatively, researchers use standardized datasets like RealToxicityPrompts to measure the toxicity of responses [10]. This assessment often utilizes the Perspective API, returning a toxicity score within the range [0, 1], with scores closer to 1 indicating higher toxicity.

In addition, emerging datasets such as ToxiGen and XSTest are being used to expand and deepen model safety assessments [11,12]. In a comparison of different alignment methods, OpenAI's research shows that RLHF models achieve higher coherence in output, thus making them more suitable for real-world scenarios like customer assistance. This indirectly indicates that RLHF achieves higher overall harmlessness while maintaining practicality. Though certain toxicity metric tests sometimes show that SFT outputs have lower toxicity values, this "low toxicity" often comes at the expense of helpfulness, as SFT answers are often overly brief and incoherent, generating degenerate or conservative text to avoid triggering harmful content. With respect to novel algorithms, Xu et al. reveal that both DPO and PPO are capable of generating responses that are less toxic compared to the baseline SFT model [9]. However, in a specific experiment by Ivison et al., PPO outputs were less toxic than DPO, but the validity of this conclusion must be limited to the safety assessment metrics and experimental settings they used [5].

4. Optimization of Fine-Tuning Alignment Method and Its Application

4.1 Optimization of Existing Fine-Tuning Methods

Despite RLHF having achieved significant success in alignment tuning, it also faces several problems. A primary issue is training instability, particularly during the PPO optimization phase. The initial Supervised Fine-Tuning (SFT) step merely provides the model with demonstration data, teaching it to mimic the pattern of preferred responses. The subsequent PPO training, which optimizes the model against a reward model, can be unstable.

For instance, the model may learn to "game" the reward system, a problem known as reward hacking. This is often observed when the model produces long, repetitive, or meaningless text simply to achieve a high reward score. To counteract this, alignment researchers typically employ technical countermeasures, such as applying a KL divergence penalty to prevent the model from deviating too far from the SFT model or clipping the reward signal to limit extreme outputs.

Another challenge is the model's sometimes poor understanding of complex human preferences or context. As an optimization method to address this, some researchers have explored multimodal feedback. According to Zhai et al., this approach helps the model better grasp the context by combining information from different modalities, such as text and images, thereby providing more detailed information to understand human preferences [13].

4.2 Practical Applications of Different Alignment Methods

Each alignment method offers distinct trade-offs in terms of data efficiency and computational cost, making them suitable for different real-world scenarios.

SFT (Supervised Fine-Tuning) is ideal for tasks that are clearly defined and have standardized answers, such as code generation or extracting key metrics from financial reports. In these cases, a dataset of high-quality demonstrations is sufficient for the model to quickly learn the correct behavior. Beisdes, DPO (Direct Preference Optimization) is well-suited for building general-purpose conversational agents and AI assistants. It is particularly valuable when computational resources are limited or rapid iteration is needed. By directly optimizing human preference data, DPO can effectively improve a model's politeness, naturalness, and overall usability. In contrast, RLHF (Reinforcement Learning from Human Feedback) remains the most time-consuming and complex method due to its multi-stage process. Nevertheless, it is powerful for tackling complex, domain-specific tasks that require advanced reasoning and the careful optimization of multi-dimensional human preferences, such as balancing helpfulness with harmlessness in nuanced situations.

5. Conclusion

This comparative literature analysis evaluates three key alignment methods: SFT, RLHF, and DPO, in relation to the Honesty, Helpfulness, and Harmlessness (HHH) dimensions. However, no single method emerged as the best. Although SFT serves as a baseline, it is generally outperformed by preference-based methods. RLHF (using PPO) excels in helpfulness and harmlessness, making it robust for safety-critical applications, but it is complex and computationally expensive. DPO offers a simpler, more efficient alternative that performs well on honesty and helpfulness, yet it can be sensitive to data quality and slightly less effective in harmlessness than RLHF. In summary, the optimal choice is context-dependent, requiring a trade-off based on specific application priorities, such as safety (favoring RLHF) or efficiency and honesty (favoring DPO).

References

[1] Leike, J., Kruger, D., Evieritt, T., et al. (2018). Scalable agent alignment via reward modeling: A research direction. https://arxiv.org/abs/1811.07871

[2] Askell, A., Bai, Y., Chen, A., et al. (2021). A general language assistant as a laboratory for alignment. https://arxiv.org/abs/2112.00861

[3] Schulman, J. (2023). Reinforcement learning from human feedback: Progress and challenges. In Berkeley EECS Colloquium. YouTube www. youtube. com/watch.

[4] Yang, Y.Q., Chern, E., Qiu, X.P., Neubig, G., & Liu, P.F. (2023). Alignment for honesty. Advances in Neural Information Processing Systems, 37, 63565-63598.

[5] Ivison, H., Wang, Y., Liu, J., et al. (2024). Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. Advances in neural information processing systems, 37, 36602-36633.

[6] Rafailfov, R., Sharma, A., Mitchell, E., et al. (2023). Direct preference optimization: Your language model is secretly a reward model. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, 53728-53741.

[7] Bai, Y., Jones, A., Ndousse, K., et al. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. ArXiv, abs/2204.05862.

ISSN 2959-6157

[8] Yang, L.O., Wu, J., Jiang, X., Alemeida, D., & L, C. (2022, March 4). Training language models to follow instructions with human feedback. https://arxiv.org/abs/2203.02155

[9] Xu, S.S., Fu, W., Gao, J.X., et al. (2024). Is DPO superior to PPO for LLM alignment? A comprehensive study. ICML'24: Proceedings of the 41st International Conference on Machine Learning, 54983-54998

[10] Gehman, S., Gururangan, S., Sap, M., et al. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3356-3369.

[11] Röttger, P., Kirk, H., Vidgen, B., et al. (2024). XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 5377-5400.

[12] Hartvigsen, T., Gabriel, S., Palangi, H., et al. (2022). Toxigen: Controlling language models to generate implied and adversarial toxicity. In Annual Meeting of the Association for Computational Linguistics, 1.

[13] Zhai, Z.X. (2025). Exploring the Optimization of RLHF and its Variants in Aligning Large Models with Human Preferences. In ITM Web of Conferences, 78, 01038. EDP Sciences.