Auto-Guideline Alignment: Human Preference Alignment for Large Language Models Without Fine-Tuning

Li-Ni Fu, Chien-Hua Chen, Chang-Chih Meng, and I-Chen Wu*
Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan
*icwu@cs.nycu.edu.tw

***Abstract*— In this work, we propose a novel technique, Auto-Guideline Alignment (AGA), that enables large language models (LLMs) to achieve human preference alignment without relying on resource-intensive parameter tuning. This approach allows models to attain comparable performance without fine-tuning and is particularly well-suited for application in black box language models such as GPT-4o. Additionally, the method is highly interpretable, providing insights into what LLMs focus on when aligning with human preferences, and the explainability of LLMs' underlying behaviour. Moreover, we intend to utilize the method to collect well-aligned datasets, further enhancing the ability to fine-tune and validate LLMs. AGA represents a significant step forward in the pursuit of more ethical and effective deployment of LLMs in real-world settings.**

I. Introduction

Large language models (LLMs) have become central to advancing the capabilities of artificial intelligence in understanding and generating human-like text. However, aligning these models with human values and preferences remains a challenge, especially when the model outputs can significantly impact user experience [1][2]. References [3] and [4] rely on extensive human feedback and iterative parameter tuning, which can be time-consuming and resource-intensive.

In order to solve these issues, we introduce a novel approach, Auto-Guideline Alignment (AGA), inspired by [5]-[7], to enhance the alignment of LLMs with human preferences without the need for fine-tuning. Our method leverages the concept of auto prompting, a technique that dynamically generates and modifies guidelines to guide the model aligning through limited human-annotated data.

Our experiments show that this method not only reduces the dependence on extensive human-labeled datasets but also achieves remarkable accuracy in preference alignment without fine-tuning. Furthermore, the inherent explainability of this method supports continuous improvement and adaptation of models to align with evolving human preferences, thereby enhancing user trust and interaction quality.

We summarized the contributions of this work as follows:

- **Reduction of Resource-Intensive Processes**: We introduce a novel framework that minimizes the reliance on resource-intensive human feedback and parameter tuning by leveraging auto prompting techniques.
- **Application to Black Box Models**: Our approach is particularly well-suited for application in black box models. This extends the utility of our method to a wider range of LLMs, including those with restricted access.
- **Enhanced Explainability**: The method provides high explainability, offering insights into how models align with human preferences. This transparency supports the systematic understanding and continuous improvement of LLMs for evolving human values and preferences.

II. Related Work

A. Preference Learning

Preference learning (PL) plays a crucial role in aligning LLMs. Consider an LLM that has undergone pre-training and fine-tuning, enabling it to follow instructions and have basic conversations. However, the model's responses can sometimes be unsafe or unhelpful. To address these issues, we often use PL techniques such as [3] and [4] to ensure the model aligns more closely with human preferences.

PL relies heavily on the generation of human preference data. Reference [8], [9] uses AI feedback instead of humans to annotate preference data, significantly reducing the time and cost required.

B. Auto Prompting

Auto prompting [5]-[7] dynamically generates and modifies prompts to "adjust" the model. The advantage of this approach is that humans can observe the generated prompts to understand the rules that the model deems important, making AI decisions less of a black box.

III. Methodology

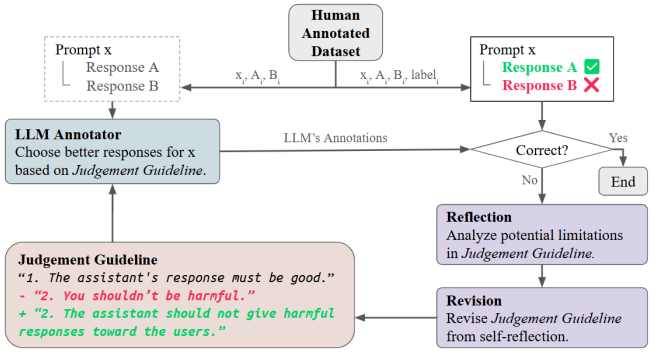

Fig. 1 A complete workflow of Auto-Guideline Alignment (AGA)

In this section, we outline the framework of AGA. The complete process is illustrated in Fig. 1. Our objective is to enable LLMs to generate a set of evaluative guidelines that can make LLMs more accurately determine whether the model's output is safe and harmless, thereby aligning the model towards more appropriate outputs.

A. Problem Definition

Let C represent the set of dialogues with n data. All $c \in C$ contains four factors: $\{x, A, B, label\}$, where A and Brepresent two possible responses based on prompt x, with $label$ indicates which one is the correct response, A or B.

The initial guideline g, as shown in Fig. 1, is set to one simple principle which intends to provide minimal constraints during subsequent guideline revisions.

B. Annotation Process

For each training epoch, we sample n pairs of data from the training dataset. The LLM annotator is then tasked with selecting the preferred response from each pair.

We evaluate the model's probabilities for both options, like [10]-[12], and select higher probability as the model's choice.

C. Reflection and Revision

If the accuracy has no improvement, the incorrectly annotated data are forwarded to the LLM analyst. The analyst identifies deficiencies in the current g that led to the incorrect selections. Based on this reflection, the model revises g, and re-evaluates the responses according to the updated guideline.

D. Evaluation

The final guideline g is applied to the testing dataset for annotation. The accuracy for evaluation is based on the guideline's selection accuracy across the entire testing dataset.

IV. Experiments

In this section, we conduct experiments to evaluate the efficiency and effectiveness of the proposed method, and generate judgemental guideline prompts with high qualities.

A. Dataset

We adopt Golden HH [13] for both training and evaluation purposes. This dataset is an enhanced version of the original HH dataset [14]. The modifications were made by rewriting positive responses in HH using *GPT-4* [15] for higher quality.

B. Models

We use *Meta LLaMA3-8B-Instruct* [16], *OpenAI GPT-4o* [17], and *OpenAI GPT-4o-mini* [18] as the LLMs in the architecture mentioned above and compare their performance. We run all of these models with 1000 pairs of training data.

C. Results

We evaluate the accuracy of these models when employing the proposed technique. The results are shown in Table 1.

TABLE I
Performance Comparison of Different Models on Golden HH Dataset

Methods	**LLaMA-3-8B-Instruct**	**GPT-4o**	**GPT-4o-mini**
Baseline	0.59	0.50	0.50
Ours	**0.93**	**0.90**	**0.95**

V. Conclusions

AGA is an advanced approach for human preference alignment. It significantly reduces the heavy reliance on human feedback and hardware resources. This work gives us a glance of how LLMs reflect on past mistakes on annotation, and adjust their strategies to align with correct positions in the spectrum of human preferences.

Future work will focus on collecting aligned preference data based on generated guidelines, with the intention of providing a benchmark of preference alignment, enhancing the performance of LLMs in subsequent fine-tuning efforts.

References

[1] H. Lee *et al.*, "RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback," *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2309.00267.

[2] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2203.02155.

[3] P. Christiano *et al.*, "Deep reinforcement learning from human preferences," *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/1706.03741.

[4] R. Rafailov *et al.*, "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2305.18290.

[5] M. Yuksekgonul *et al.*, "TextGrad: Automatic 'Differentiation' via Text," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2406.07496.

[6] T. Shin *et al.*, "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts," *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2010.15980.

[7] Zhengbao Jiang *et al.,* "How Can We Know What Language Models Know?" *Transactions of the Association for Computational Linguistics, vol. 8, pp. 423–438, 2020.*

[8] Y. Bai *et al.*, "Constitutional AI: Harmlessness from AI Feedback," *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2212.08073.

[9] L. Tunstall *et al*, "Zephyr: Direct Distillation of LM Alignment," *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2310.16944.

[10] D. Hendrycks *et al.*, "Measuring Massive Multitask Language Understanding," *arXiv preprint arXiv:2009.03300*, 2021. [Online]. Available: https://arxiv.org/abs/2009.03300.

[11] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering," *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2203.14371.

[12] BIG-bench authors, "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=uyTL5Bvosj.

[13] T. Cai *et al.*, "ULMA: Unified Language Model Alignment with Demonstration and Point-wise Human Preference," *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2312.02554.

[14] Y. Bai *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2204.05862.

[15] OpenAI *et al.*, "GPT-4 Technical Report," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2303.08774.

[16] A. Dubey *et al.,* "The Llama 3 Herd of Models," *arXiv preprint arXiv:2407.21783,* 2024.

[17] OpenAI, "Hello gpt-4o," [Online]. Available: https://openai.com/index/hello-gpt-4o/, 2024.

[18] OpenAI, "GPT-4o mini: advancing cost-efficient intelligence," [Online]. Available: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024.