Teaching Language Models to Self-Improve by Learning from Language Feedback

Chi Hu[1] Yimin Hu[1] Hang Cao[1] Tong Xiao[1,2]* Jingbo Zhu[1,2]

[1]NLP Lab, School of Computer Science and Engineering,
Northeastern University, Shenyang, China
[2]NiuTrans Research, Shenyang, China
huchinlp@gmail.com
{xiaotong,zhujingbo}@mail.neu.edu.cn

Abstract

Aligning Large Language Models (LLMs) with human intentions and values is crucial yet challenging. Current methods primarily rely on human preferences, which are costly and insufficient in capturing nuanced feedback expressed in natural language. In this paper, we present Self-Refinement Tuning (SRT), a method that leverages model feedback for alignment, thereby reducing reliance on human annotations. SRT uses a base language model (e.g., Tulu2) to generate initial responses, which are critiqued and refined by a more advanced model (e.g., GPT-4-Turbo). This process enables the base model to self-evaluate and improve its outputs, facilitating continuous learning. SRT further optimizes the model by learning from its self-generated feedback and refinements, creating a feedback loop that promotes model improvement. Our empirical evaluations demonstrate that SRT significantly outperforms strong baselines across diverse tasks and model sizes. When applied to a 70B parameter model, SRT increases the win rate from 9.6% to 25.8% on the AlpacaEval 2.0 benchmark, surpassing well-established systems such as GPT-4-0314, Claude 2, and Gemini. Our analysis highlights the crucial role of language feedback in the success of SRT, suggesting potential for further exploration in this direction.

1 Introduction

Recent advances in Large Language Models (LLMs) have revolutionized the field of natural language processing. These models have demonstrated remarkable capabilities in various tasks such as open-ended generation, question answering, and mathematical reasoning (Brown et al., 2020; Chowdhery et al., 2023; Ouyang et al., 2022; Bubeck et al., 2023). However, despite their impressive performance, LLMs occasionally generate content that can be untruthful or harmful. This
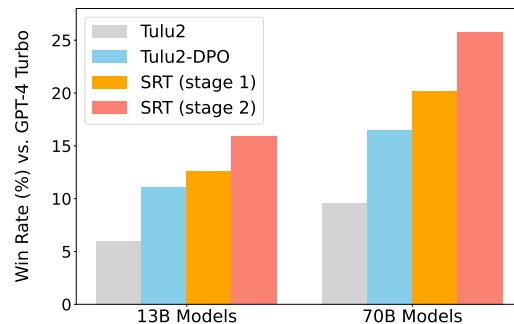
*Corresponding author.

Figure 1: Results on AlpacaEval 2.0. SRT significantly boosts the performance of the base Tulu2 models. We report the win rates against GPT-4 Turbo.

highlights the need to align language models with human intentions and values to ensure safe and controllable deployment (Weidinger et al., 2021; Bai et al., 2022b; Perez et al., 2022).

Many current methods to align LLMs, such as Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO), rely on human preferences (Christiano et al., 2017; Jaques et al., 2019; Stiennon et al., 2020; Rafailov et al., 2023). These preferences are typically expressed through rankings or scores. However, there are two main challenges with these methods. Firstly, annotating human preferences is expensive and time-consuming, which makes it difficult to scale these techniques. For instance, Scheurer et al. (2023) spent $40k to annotate approximately 60K feedback instances for a single summarization task. Secondly, simple rankings or scores may not fully capture the subtlety and complexity of human preferences. This can limit the depth of feedback provided to models for improvement (Myers et al., 2021; Casper et al., 2023). Humans, on the other hand, can derive insights from minimal language feedback, indicating that there is potential for more sophisticated and efficient alignment methods (Scheurer et al., 2022; Chen et al., 2023).

In response to these challenges, we introduce a new method for aligning language models, which we call Self-Refinement Tuning (SRT). SRT obviates the need for human preferences with a two-stage learning process. In the first stage, SRT employs a powerful model like GPT-4 to critique and refine the outputs from a base model, such as Tulu2 (Ivison et al., 2023). The base model is then fine-tuned on critiques and refinements, enabling self-evaluation and improvement. In the second stage, SRT further boosts the model by learning from self-feedback. Specifically, SRT uses the fine-tuned model to generate model preferences, i.e., pairs of outputs and refinements. Subsequently, SRT optimizes the model on these preferences using DPO (Rafailov et al., 2023).

We evaluate SRT on open-ended generation, question answering, and mathematical reasoning. Empirical results show that SRT consistently outperforms baseline models of sizes from 7B to 70B, with an average performance enhancement of 3.7 to 4.0 points. These improvements are obtained using merely 22K feedback instances annotated by GPT-4 Turbo. Figure 1 shows that the SRT significantly improves Tulu2 models using model-generated feedback. The strongest model trained with SRT attains a 25.8% win rate against GPT-4 Turbo on the AlpacaEval 2.0 benchmark. This performance surpasses established systems such as GPT-4 0314, Mistral Medium, Claude, and Gemini. Our analysis confirms that the success of SRT primarily stems from its language feedback feature, which identifies weak areas and offers valuable suggestions for improvement.

2 Related Work

This section briefly reviews two critical areas in the alignment of LLMs: learning from AI feedback and learning to self-improve. Our work intersects with these domains and addresses existing challenges.

Learning from AI Feedback. Learning from human feedback is the key to the success of state-of-the-art language models such as GPT-4 (OpenAI, 2023) and Gemini (Google, 2023). However, acquiring high-quality human feedback is both costly and time-consuming (Christiano et al., 2017; Jaques et al., 2019; Stiennon et al., 2020; Rafailov et al., 2023). This has led to a growing interest in harnessing AI-generated feedback to enhance LLMs (Lee et al., 2023; Roit et al., 2023; Hu et al., 2024). For instance, Hu et al. (2024)

employs LLMs to annotate ranking-based preferences. Our research diverges from this approach by utilizing a more comprehensive range of feedback, encompassing identified weaknesses and proposed improvements. A similar approach is Reinforcement Learning from AI feedback (RLAIF, Bai et al., 2022b), which uses LLMs to generate critiques and refinements. However, RLAIF encounters challenges such as computational inefficiency and unstable training dynamics due to the inherent complexities of reinforcement learning techniques (Casper et al., 2023; Shen et al., 2023). We address these challenges by unifying the generation of feedback and refinement into instruction-following, thereby simplifying the training process.

Learning to Self-Improve. Using Large Language Models (LLMs) to identify and correct their own errors has become increasingly popular. This self-improvement capability has significantly enhanced performance across various tasks, such as question-answering, code generation, and mathematical reasoning (Shinn et al., 2023; Madaan et al., 2023; Chen et al., 2024). Pan et al. (2023) conducted a comprehensive survey of this field. However, these approaches rely on critique and refinement skills that are generally lacking in open-source language models (Valmeekam et al., 2023; Huang et al., 2023). This underscores the urgent need to train open-source models for self-improvement. Unlike previous methods that train separate models for generation, critique, and refinement (Yasunaga and Liang, 2020; Welleck et al., 2022), our approach employs a single model for all tasks, facilitating knowledge transfer across them. Additionally, our method is designed for the general alignment of LLMs, in contrast to prior methods that are task-specific (Wang and Li, 2023; Yu et al., 2023; Lu et al., 2023).

3 Methodology

Figure 2 presents the two-stage Self-Refinement Tuning (SRT) process. SRT aims to enhance the capabilities of a base language model (denoted as M_{base}) by harnessing learning from AI feedback (LAIF), thereby reducing the reliance on human feedback. In the first stage, we train M_{base} to self-improve by learning from feedback and refinements annotated by more powerful models. In the second stage, we further optimize the model using its self-generated feedback.
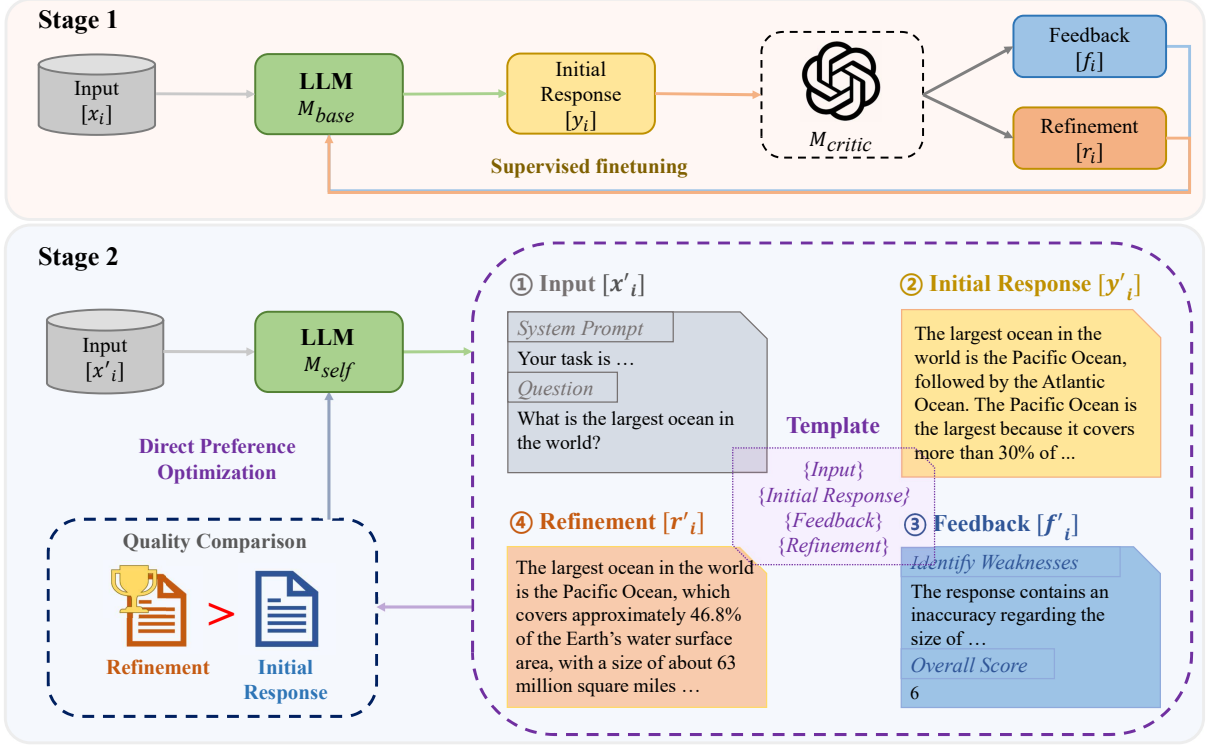
Figure 2: An overview of Self-Refinement Tuning (SRT). In the first stage (above), SRT teaches the base model to self-improve by fine-tuning it on the feedback and refinements from a powerful critic model. In the second stage (bottom), SRT enables the model to learn from its self-generated feedback and refinements.

3.1 Training Models for Self-Improvement

3.1.1 Collecting Feedback and Refinements

In the first stage, we collect the training data for self-improvement from the interaction of two models. Specifically, M_{base} interacts with a stronger model, denoted as M_{critic}. Given a set of instructions $\mathbf{x} = [x_1, ..., x_N]$, M_{base} generates initial responses $\mathbf{y} = [y_1, ..., y_N]$. M_{critic} then provides feedback $\mathbf{f} = [f_1, ..., f_N]$ on these responses. As delineated in Table 1, we instruct M_{critic} to generate 1) analyses of the weaknesses, 2) overall quality scores ranging from 1 to 10, and 3) suggestions for enhancing the responses. Subsequently, M_{critic} generates the refinements (i.e., improved responses) $\mathbf{r} = [r_1, ..., r_N]$ according to its previous feedback.

The critique-refinement process can be iterative: after the first iteration, we obtain a set of instructions, initial outputs, critiques, and refinements. We then use the refinements as the inputs for the next iteration and continue the process until the quality of the outputs no longer improves. Intriguingly, we find that a *single iteration* of refinement is typically adequate, with additional iterations contributing only marginal improvements. See Section 4.1 for more details of the process of collecting the feedback and refinement.

3.1.2 Self-Improvement Training

After collecting the feedback and refinements, we fine-tune M_{base} to identify and rectify its own errors or undesirable behaviors, thereby improving its performance. To facilitate this, we reformat the collected data into sequences of sentences as follow:

$Instruction \rightarrow Response \rightarrow Feedback \rightarrow Refinement$

By doing so, we can train the model using a language modeling objective. This method is somewhat akin to Chain of Hindsight (Liu et al., 2023), where the model is trained to produce outputs from provided instructions and feedback. However, our approach differs in training the model to generate the feedback and refinements sequentially instead of merely predicting the outputs. More formally, we want to optimize:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1} \log P(y_i, f_i, r_i | x_i) \quad (1)$$

where N is data size, x_i is the input instruction, y_i is the initial output, f_i is the feedback, and r_i is the best refinement with the highest score. The input instruction is masked from the loss calculation during the training phase. This results in a

Instruction
{Instruction}
Response
{Response}
Requirements
As a AI assistant instructor, your task is to provide constructive feedback on responses generated by the assistant. Ensure that your feedback focuses on the original instruction and do not introduce new requirements. Follow these steps to ensure your feedback is effective and beneficial:
1. Identify Weaknesses: Carefully review the response and pinpoint any areas where it falls short.
2. Offer Actionable Advice: Provide specific suggestions on how the response can be improved.
3. Conclude with a Rating: After providing feedback, score the response on a scale from 1 to 10, with 1 being the lowest and 10 the highest. Use the format: "Overall Score: [[1-10]]".
4. Show an Improved Response: Offer an improved version that incorporates your feedback. Clearly indicate the enhanced response with the heading: "### Improved Response".
Feedback

Table 1: The template we used for obtaining feedback from LLMs. It instructs LLMs to evaluate and refine the response to a given instruction.

self-improvement model, denoted by M_{self}, which can provide feedback on its outputs and refine them until they are satisfactory. In Section 4.2, we show that M_{self} significantly outperforms M_{base}.

3.2 Scaling SRT through Self-Feedback

We now describe how M_{self} can be used to scale the learning from AI feedback. Specifically, we leverage M_{self} to generate preference data and for further model optimization. Our approach is similar to Self-Rewarding (Yuan et al., 2024) but deviates from the strategy for acquiring superior responses. Rather than sampling multiple responses and selecting the optimal one, we directly generate an improved response via self-refinement, thereby enhancing efficiency.

More concretely, we use model M_{self} to generate initial responses $\mathbf{y'} = [y'_1, ..., y'_N]$ from a distinct instruction set $\mathbf{x'} = [x'_1, ..., x'_N]$. The model critiques and refines these responses into feedback $\mathbf{f'}$ and refinements $\mathbf{r'}$. We filter to ensure refinements exceed initial response quality, based on model-assigned scores. This yields preference pairs of initial responses and refinements, which can used for DPO training (Rafailov et al., 2023).

4 Experiments

4.1 Implementation Details

In this subsection, we outline the specifics of SRT training, which includes the choice of base models, the procedure for feedback annotation and cleaning, and the comprehensive model training process.

Models. To evaluate the efficacy of SRT, we employ three fine-tuned LLaMA2 models as our base models. These include Tulu2-7B, Tulu2-13B, and Tulu2-70B, all trained on the Tulu-2 Mixture (Ivison et al., 2023). Given that these datasets are partially distilled from models such as GPT-4, they can also be referred to as **dSFT**, an acronym for distilled Supervised Fine-Tuning (Tunstall et al., 2023). In the first stage of SRT, we use Tulu2-7B to generate initial responses and employ GPT-4 Turbo (gpt-4-1106-preview) to generate language feedback and refinements. The GPT-4 critic exhibits a 78.9% agreement rate with human preferences on the HH-RLHF dataset (Bai et al., 2022a). We then fine-tune all base models on these refinements. In the second stage, these fine-tuned models independently generate feedback and refinements for DPO training, which we refer to as **sDPO** (self-feedback DPO). To measure the effectiveness of self-feedback, we further compare our models with other DPO variants trained on the UltraFeedback (Cui et al., 2023) dataset. These include Tulu2-DPO-7B, Tulu2-DPO-13B, and Tulu2-DPO-70B. We refer to these models as **dDPO**, an acronym for distilled DPO (Tunstall et al., 2023).

Details on Feedback and Refinements Annotation. In our two-stage process, we initially select 25,000 instructions from the Tulu-2 Mixture, followed by 75,000 instructions for the second stage. We adopt the approach of Li et al. (2023a), choosing examples of a single conversation turn to streamline the refinement process. For generating responses, we use a sampling temperature of 0.7, while for feedback and refinement generation, we set the temperature to 0 and restrict the maximum new tokens to 2,048. We use the prompt template shown in Table 1 for annotating feedback and refinements. Figure 3 shows the score distribution of the initial outputs of Tulu2-7B and their subsequent refinements. On average, the refinements enhance the score by 1.5 points compared to the initial outputs. Table 2 shows one iteration of critique-and-refinement is sufficient.
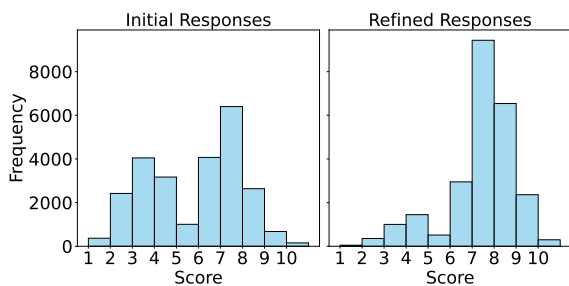
Figure 3: The score distribution of 25K initial responses (left) and refined responses (right). The initial responses are generated by Tulu2-7B and are then refined and scored by GPT-4 Turbo using the template presented in Table 1.

Iteration	Average Score
Initial Responses	5.42
Iteration 1	6.93
Iteration 2	7.04
Iteration 3	6.91
Iteration 4	6.85

Table 2: The average score of the refinements generated by GPT-4 Turbo at different iterations. The results are obtained from 1,000 samples from the Tulu-2 Mixture dataset. The initial responses are generated by Tulu2-7B.

Post-processing of Feedback and Refinements. Our goal is to compile instances that form a consistent sequence of instruction-response-feedback-refinement. However, even sophisticated models like GPT-4 Turbo occasionally fail to conform strictly to the format requirements. As a result, we apply the following filtering rules:

Rule #1: The feedback should include potential weaknesses, overall scores, and suggestions[1].

Rule #2: The quality of the refinement should not be lower than that of the initial response.

Note that we evaluate the initial responses and their subsequent refinements independently. This is important as the critic sometimes gives higher scores when it has been conditioned with prior feedback, regardless of the quality of the refinement. After applying these filters, we obtain 22K and 63K valid feedback and refinements for the initial and subsequent stages, respectively.

Training Details. The base models are trained using the Tulu-2 Mixture dataset, following the same settings as Ivison et al. (2023). In the first stage of SRT, we fine-tune the base models for five epochs. We concatenate each instruction's initial response, feedback, and refinement into sequences. Following Ivison et al. (2023), we set the maximum sequence length to 8,192 but use a smaller global batch size of 32. We use a cosine learning rate scheduler with a peak learning rate of 1e-5 and 10% warmup steps. In the second stage, we train our DPO models for one epoch with a β value of 0.01. We maintain the same optimizer and learning rate scheduler as the first stage but adjust the maximum learning rate to 5e-7 to ensure training stability.

We use the HuggingFace TRL library[2] to train our models, and use FlashAttention-2 (Dao, 2023) and DeepSpeed Zero-3 (Rajbhandari et al., 2020) to speedup training.

Evaluation Details. We evaluate SRT across open-ended generation, reasoning, and question-answering. For open-ended generation, we test SRT on the AlpacaEval benchmark (Li et al., 2023b), which consists of 805 instructions. We report the win rate of our models compared to GPT-3.5 (text-davinci-003). Since this baseline may be outdated, we also compare our models with GPT-4 Turbo (gpt-4-1106-preview). We use the default configuration provided in the official library[3]. For reasoning, we evaluate our models on GSM8K (Cobbe et al., 2021) and Big-Bench-Hard (BBH, Suzgun et al., 2022). We report the exact match score (EM) on the test sets, using few-shot chain-of-thought prompting strategies identical to those used by Ivison et al. (2023). For question-answering, we test with TydiQA (Clark et al., 2020), a multilingual benchmark that covers 11 different languages. In line with Ivison et al. (2023), we report the F1 score under the Gold Passage (GP) setting, where the passage containing the answer is provided. To assess the "self-evaluation" capability of SRT, we test our models on the HH-RLHF dataset (Bai et al., 2022a), comparing model-predicted scores with human preferences. We evaluate using a sample of 500 data points from the HH-RLHF test set. Unless stated otherwise, we always select the refined responses as the outputs for the SRT models. We also limit the output refinement to one iteration, as additional iterations provide minimal benefits and significantly slow down the model's generation.

---

[1] We identify these elements by looking for keywords such as '###Feedback,' '###Overall Score,' and '###Improved Response.'

[2] https://github.com/huggingface/trl

[3] We use 'alpaca_eval_gpt4' as the judge for comparing GPT-3.5 and use 'weighted_alpaca_eval_gpt4_turbo' for GPT-4 Turbo.

Model	Type	GSM8k 8-shot CoT, EM	BBH 3-shot CoT, EM	TydiQA 1-shot, F1	AlpacaEval % Win	Average -
Tulu2 7B	dSFT	34.0	48.5	46.4	73.9	50.7
Tulu2-DPO 7B	dDPO	34.5	45.5	44.5	85.1	52.4
SRT 7B (*stage 1*)	dSFT	**35.7**	**49.2**	**47.9**	84.6	**54.4**
SRT 7B (*stage 2*)	sDPO	34.2	47.3	47.2	**85.3**	53.4
Tulu2 13B	dSFT	46.0	49.5	53.2	78.9	56.9
Tulu2-DPO 13B	dDPO	49.5	49.4	39.7	89.5	57.0
SRT 13B (*stage 1*)	dSFT	48.2	50.4	56.4	88.7	60.9
SRT 13B (*stage 2*)	sDPO	**49.7**	**50.9**	**57.9**	**91.6**	**62.5**
Tulu2 70B	dSFT	73.0	68.4	53.6	86.6	70.4
Tulu2-DPO 70B	dDPO	71.5	66.0	35.8	95.1	67.1
SRT 70B (*stage 1*)	dSFT	72.3	**70.2**	60.9	93.1	74.1
SRT 70B (*stage 2*)	sDPO	**73.9**	69.7	**61.8**	**95.2**	**75.1**

Table 3: Comparisons of the performance on four benchmarks. All models are fine-tuned from LLaMA2 pre-trained models. The term **dSFT** stands for distilled supervised fine-tuning, **dDPO** represents direct preference optimization with distilled feedback, and **sDPO** signifies direct preference optimization with self-generated feedback. The table presents the average scores for each benchmark, with the highest performing models of equal size highlighted in **bold**. The AlpacaEval results are win rates compared against GPT-3.5.

4.2 Main Results

Table 3 summarizes the experimental results on four benchmarks. The results indicate that SRT consistently improves model performance across various tasks and model sizes. In the first stage, models trained with language feedback from GPT-4 significantly outperform the Tulu2 baselines. On average, the first stage of SRT enhances the performance by a margin of 3.7 to 4.0 across different model scales. Models trained during this stage are comparable to the Tulu2-DPO baselines on AlpacaEval but use considerably fewer feedback instances (22K vs 64K). These findings underscore the efficacy of training models to assess and refine their outputs by learning from the feedback of more advanced models.

In the second stage, SRT further boosts performance by scaling feedback with self-generated responses and refinements. The most notable improvement is observed on the AlpacaEval task. SRT's second stage is more effective with larger models (13B and 70B) but less with smaller ones (7B), which see an average drop of 1.0 points. This suggests that the inherent capabilities of a model may restrict its ability to learn from self-feedback.

Additionally, the enhancements observed in reasoning tasks, including GSM8K and BBH, are relatively slight, especially for the 7B and 13B models. This could be attributed to the limited reasoning capabilities present in these models.

Model	Win Rate (%)	Length
GPT-4 0314	22.1	1371
Mistral Medium	21.9	1500
Claude 2	17.2	1069
Gemini Pro	16.9	1315
Tulu2 13B	6.0	993
Tulu2-DPO 13B	11.1	1440
SRT 13B (*stage 1*)	12.6	1307
SRT 13B (*stage 2*)	15.9	1285
Tulu2 70B	9.6	960
Tulu2-DPO 70B	16.5	1480
SRT 70B (*stage 1*)	20.2	1536
SRT 70B (*stage 2*)	**25.8**	1663

Table 4: AlpacaEval 2.0 results. We report the win rates compared to GPT-4 Turbo and the output length.

4.3 Results on AlpacaEval 2.0

GPT-3.5 may be a relatively weak baseline for evaluating our strongest models. For example, both SRT 70B and Tulu2-DPO 70B models achieve win rates over 95%. Therefore, we further evaluate these models on the more challenging AlpacaEval 2.0 benchmark, which employs GPT-4 Turbo as the baseline for calculating win rates. Table 4 presents the results of 13B and 70B models and several proprietary models. We can see that SRT models significantly outperform Tulu2 models. Notably, our strongest model, SRT 70B (stage 2), achieves a win rate of 25.8%, outperforming Tulu2-DPO 70B by a substantial margin of 9.0%. The model

Model	Human Agreement (%)
GPT-4 Turbo	78.9
SRT 7B (*stage 1*)	74.4
SRT 7B (*stage 2*)	68.2
SRT 13B (*stage 1*)	72.2
SRT 13B (*stage 2*)	68.9
SRT 70B (*stage 1*)	72.8
SRT 70B (*stage 2*)	72.3

Table 5: Feedback accuracy of different models. We report the level of agreement with human preferences based on 500 samples from the HH-RLHF test set.
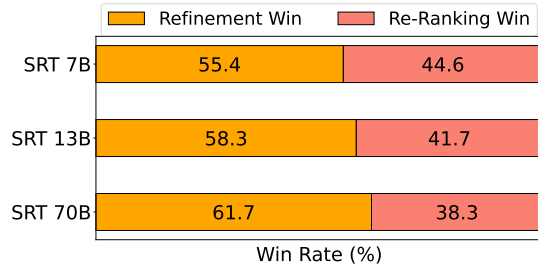

Figure 4: Self-Refinement vs. Re-Ranking over 16 candidates. The results are obtained on the AlpacaEval test set using models trained at the first stage of SRT.

Model	Win Rate (%)	Drop
SRT 13B (*stage 1*)	88.7	-
— *score*	87.6	1.1
— *suggestion*	85.9	2.8
— *weakness*	85.5	3.2
— *all feedback*	83.6	5.1
Refinement Only	84.2	4.5

Table 6: Ablation study on the language feedback components. We report the win rates compared to GPT-3.5 using the same settings as Section 4.2.

surpasses well-established models such as GPT-4 0314, Mistral Medium, Claude 2, and Gemini Pro. We also find that SRT encourages models to produce more verbose outputs, which could influence the observed performance improvement. However, despite generating shorter responses, our SRT 13B models achieve superior results compared to Tulu2-DPO models.

4.4 Results on HH-RLHF

Table 5 presents a comparison of feedback accuracy among different models on the HH-RLHF test set. Feedback accuracy is evaluated based on the agreement of model outputs with human preferences. Specifically, for each pair of responses, the feedback is considered 'correct' if the score of the 'chosen' response exceeds that of the 're-jected' response. Remarkably, GPT-4 Turbo surpasses all other models, achieving an agreement rate of 78.9%. Intriguingly, during the initial stage of SRT, the 7B model outperforms larger models, reaching an agreement rate of 74.4%. However, after the second-stage fine-tuning of SRT, the accuracy of the 7B and 13B models significantly declines, while the 70B models show slight changes.

5 Analysis

This section provides an in-depth examination of the factors that affect the performance of our methods. Initially, we evaluate the efficacy of self-refinement in SRT by contrasting it with the widely adopted re-ranking approach. Subsequently, we explore the influence of various elements in SRT, encompassing language feedback, the quality of refinements, and the quantity of training data. Through these analyses, we aim to deepen the understanding of our methods.

5.1 Self-Refinement vs. Re-Ranking

Re-ranking is a widely adopted technique to bolster the performance of text generation systems. Given that our models can precisely assess the quality of their outputs, they can be effectively leveraged for re-ranking. In this study, we contrast the efficacy of self-refinement and re-ranking using the AlpacaEval test set. For self-refinement, we employ greedy decoding and refine the response once. For re-ranking, we sample 16 distinct responses using a temperature of 0.7. These responses are then re-ranked based on the scores predicted by the model itself. We then task GPT-4 for comparing the two sets of responses using the settings described in Section 4.1. Figure 4 illustrates a direct comparison between self-refinement and re-ranking across 16 candidate responses. It is evident that self-refinement consistently surpasses re-ranking in performance across various models. This advantage becomes more pronounced with an increase in model size. Given the significant cost associated with generating multiple responses, self-refinement emerges as a more efficient and cost-effective approach for improving the performance of language models.
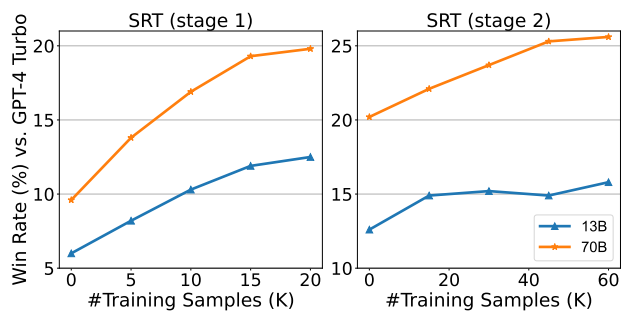
Figure 5: Win rates against GPT-4 Turbo by varying numbers of training samples for SRT models.

Figure 6: AlpacaEval results (vs. GPT-3.5). The low-, medium-, and high-quality refinements have scores of 6, 7, and 8, respectively.

5.2 Impact of Language Feedback

A unique characteristic of SRT is its integration of language feedback, which offers insights into potential weaknesses, overall scores, and suggestions for improvement. Here, we ablate these components to investigate their influence on the final performance. We evaluate 13B models trained during SRT's first stage on AlpacaEval and compare these models with GPT-3.5. The results are presented in Table 6. We find that removing any part of the feedback leads to a drop in performance. Also, weaknesses and suggestions are almost equally important, while the score seems less necessary. Removing all feedback results in a significant performance drop of 5.1 points, which is even inferior to the performance achieved by training solely with refinement. To summarize, these findings emphasize the crucial role of language feedback in enhancing SRT's performance.

5.3 Impact of Training Data Size

In our primary experiments, we use 22K and 63K training samples for the first and second stages of SRT, respectively. We now delve further into the impact of data volume on the performance of SRT. We experiment with the challenging AlpacaEval 2.0 benchmark and replicate the training settings from Section 4.1. Figure 5 illustrates the results of 13B and 70B models. As depicted, the performance almost monotonically increases with the variation in the number of training samples. Models of different sizes and stages exhibit diverse convergence speeds. Increasing the training data results in a more significant performance gain on the 70B models. However, we also find the 13B model's performance slightly deteriorates when the data volume increases from 36K to 48K. We hypothesize that this could be due to the presence of noise. To validate this hypothesis, we further analyze the impact
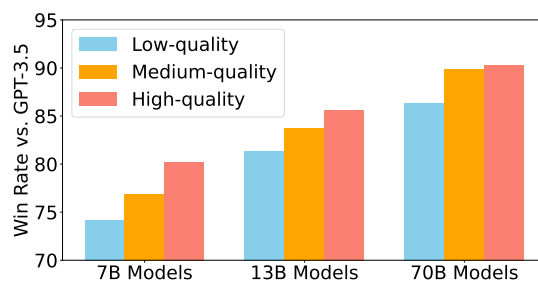
of data quality on the performance of SRT.

5.4 Impact of Refinement Quality

In our primary experiment, we merely filter refinements that are of lower quality (reflected by overall scores) than the initial responses. Here, we delve deeper into this issue by conducting experiments with more fine-grained control over the refinement quality. To investigate this, we train models with varying levels of refinement quality and compare their performances. Specifically, we select samples of 2,000 refinements each from low-, medium-, and high-quality categories, corresponding to scores of 6, 7, and 8, respectively. As illustrated in Figure 6, there is a clear monotonic increase in the performance of SRT in line with the quality of refinements. The results hold with varying model sizes, suggesting that SRT can be further augmented by enhancing the refinement quality.

6 Conclusion

We have presented Self-Refinement Tuning (SRT) for aligning language models using language feedback. SRT initially teaches language models to assess and enhance their outputs by learning from feedback provided by more advanced models. Subsequently, SRT optimizes these models further by learning from their self-generated feedback. The primary benefits of SRT are twofold: (1) it obviates the need for human-annotated preferences, and (2) it obtains promising performance across a wide range of tasks and model sizes. In our analysis, we find that both high-quality language feedback and refinements are crucial for language models to learn to self-improve. Collectively, we demonstrate that SRT is an effective and efficient strategy for improving the performance of language models.

Acknowledgement

Limitations

While SRT has demonstrated promising results across a variety of tasks, it is not without its limitations. A significant constraint lies in its dependency on a powerful critic model to provide feedback and refinements. Even state-of-the-art language models, such as GPT-4-Turbo, can sometimes generate feedback or refinements that are inaccurate or suboptimal. For instance, the refined responses from GPT-4 yield an approximately average score of 7, indicating substantial room for improvement.

Moreover, one of the inherent limitation of self-refinement is it greatly increase the output length of language models. Although we find a single iteration of self-refinement is typically sufficient (in Table 2), it still approximately doubles the output length of our models. The lengthy outputs lead to higher computational costs during decoding.

In the future, we will develop more accurate and efficient feedback mechanisms, optimize the self-refinement process to control output length, and explore ways to improve the quality of refinements.

Ethical Considerations

SRT enhances language models by incorporating feedback and refinements from both open-source and commercial models. This approach minimizes the reliance on human annotations but also might introduces potential risks. Notably, SRT can lead to unintended issues such as overfitting and the reinforcement of existing biases in the initial models. Therefore, it is essential to develop more robust and trustworthy alignment methods to ensure these language models are safe, fair, and controllable. Additionally, as the model improves, it may become increasingly complex and difficult to interpret, posing challenges in applications where transparency and interpretability are crucial. To mitigate these issues, future work could integrate ethical guidelines or external validation tools into the SRT process.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, John Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, E Perez, Jamie Kerr, Jared Mueller, Jeff Ladish, J Landau, Kamal Ndousse, Kamilė Lukoiūtė, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noem'i Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, T. J. Henighan, Tristan Hume, Sam Bowman, Zac Hatfield-Dodds, Benjamin Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom B. Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback. *ArXiv preprint*, abs/2212.08073.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani,

Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research.* Survey Certification.

Angelica Chen, Jérémy Scheurer, Tomasz Korbak, Jon Ander Campos, Jun Shern Chan, Samuel R. Bowman, Kyunghyun Cho, and Ethan Perez. 2023. Improving code generation by training with natural language feedback.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations.*

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *ArXiv preprint*, abs/2310.01377.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *ArXiv preprint*, abs/2307.08691.

Google. 2023. Gemini: A family of highly capable multimodal models.

Chi Hu, Yuan Ge, Xiangnan Ma, Hang Cao, Qiang Li, Yonghua Yang, Tong Xiao, and Jingbo Zhu. 2024. Rankprompt: Step-by-step comparisons make language models better reasoners.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *ArXiv preprint*, abs/2310.01798.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew E. Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hanna Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *ArXiv preprint*, abs/2311.10702.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Àgata Lapedriza, Noah J. Jones, Shixiang Shane Gu, and Rosalind W. Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *ArXiv preprint*, abs/1907.00456.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *ArXiv preprint*, abs/2309.00267.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023a. Self-alignment with instruction backtranslation. *ArXiv preprint*, abs/2308.06259.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Hao Liu, Carmelo Sferrazza, and P. Abbeel. 2023. Chain of hindsight aligns language models with feedback. *ArXiv preprint*, abs/2302.02676.

Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Fei Mi, Baojun Wang, Weichao Wang, Lifeng Shang, and Qun Liu. 2023. Self: Self-evolution with language feedback.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Vivek Myers, Erdem Biyik, Nima Anari, and Dorsa Sadigh. 2021. Learning multimodal rewards from rankings. In *Conference on Robot Learning, 8-11 November 2021, London, UK*, volume 164 of *Proceedings of Machine Learning Research*, pages 342–352. PMLR.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Liangming Pan, Michael Stephen Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *ArXiv preprint*, abs/2308.03188.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv preprint*, abs/2305.18290.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, page 20. IEEE/ACM.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, L'eonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos, Piotr Stańczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Annual Meeting of the Association for Computational Linguistics*.

J'er'emy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with language feedback.

J'er'emy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2023. Training language models with language feedback at scale. *ArXiv preprint*, abs/2303.16755.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *ArXiv preprint*, abs/2309.15025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan J. Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *ArXiv preprint*, abs/2009.01325.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *ArXiv*, abs/2310.16944.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *ArXiv preprint*, abs/2310.08118.

Danqing Wang and Lei Li. 2023. Learning from mistakes via cooperative study assistant for large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Laura Weidinger, John F. J. Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zachary Kenton, Sande Minnich Brown, William T. Hawkins, Tom Stepleton, Courtney Biles, Abeba

Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *ArXiv preprint*, abs/2112.04359.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *ArXiv preprint*, abs/2211.00053.

Michihiro Yasunaga and Percy Liang. 2020. Graph-based, self-supervised program repair from diagnostic feedback. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10799–10808. PMLR.

Xiao Yu, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhou Yu. 2023. Teaching language models to self-improve through interactive demonstrations. *ArXiv preprint*, abs/2310.13522.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *ArXiv preprint*, abs/2401.10020.