REINFORCEMENT LEARNING FROM AI FEEDBACK: A CROSS-MODEL ANALYSIS OF PERFORMANCE, SCALABILITY AND BIAS

Lê Văn Nguyễn [1], Rory Sie [2]

[1] University of Wollongong, Australia
[2] Nib group, Australia

ABSTRACT

Reinforcement Learning with Human Feedback (RLHF) has significantly enhanced the performance of large language models (LLMs) in tasks such as summarization, dialogue generation, and content moderation. However, the reliance on human-annotated data makes RLHF expensive and difficult to scale. To address these challenges, Reinforcement Learning from AI Feedback (RLAIF) has emerged as a promising alternative. In RLAIF, AI-generated preference labels replace human feedback, offering a more cost-effective and scalable solution while maintaining competitive performance. Despite its success in single-model families, RLAIF's generalizability across diverse model architectures and scales remains unclear. This study extends the evaluation of RLAIF by applying it to three different model families—T5, Phi-3.5, and LLaMa 3.2— representing a variety of model sizes and architectures. We compare RLAIF with traditional supervised fine-tuning (SFT) and examine the impact of model size on its effectiveness. Our findings reveal that RLAIF improves model alignment across all architectures, although the extent of the improvement varies depending on the model type. The research contributes to the broader discussion on improving the efficiency and scalability of reinforcement learning techniques for LLM alignment. By evaluating RLAIF across multiple architectures, our work provides practical guidance for implementing AI feedback-based alignment techniques that are applicable to a wide range of LLMs, advancing the field of AI model fine-tuning.

KEYWORDS

Reinforcement Learning, AI Feedback, Large Language Models, Alignment, Scaling

1. INTRODUCTION

Recent advances in large language models (LLMs) have greatly benefited from Reinforcement Learning from Human Feedback (RLHF), a method that improves model performance in tasks such as summarization, dialogue generation, and content moderation [1], [2]. However, the reliance on high-quality human annotations presents significant scalability challenges, rendering RLHF both costly and time-consuming. As the demand for more scalable methods grows, an alternative approach—Reinforcement Learning from AI Feedback (RLAIF)—has emerged. In RLAIF, human-generated preference labels are replaced by AI-generated ones, effectively reducing both costs and the time required for training, while maintaining comparable performance in various tasks [3]. Recent work by Lee et al. [4] demonstrated that RLAIF performs on par with RLHF across multiple language generation tasks.

David C. Wyld et al. (Eds): SIGI, CSTY, AI, NMOCT, BIOS, AIMLNET, MaVaS, BINLP – 2025
pp. 197-211, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.151916

Despite these advancements, several questions remain unanswered regarding RLAIF's effectiveness across different model architectures and scales. Previous evaluations focused primarily on models from the same family (e.g., PaLM 2), leaving a gap in understanding how RLAIF generalizes to diverse architectures. Additionally, AI-generated preference labels are known to suffer from position bias, where the order in which response candidates are presented can skew model preferences [5]. While Lee et al. [4] identified this issue, their study did not systematically address methods for mitigating position bias in RLAIF-generated preferences.
The motivation for this study stems from the need to understand how RLAIF performs across different model families and at various scales, especially in comparison to traditional Supervised Fine-Tuning (SFT). Given the scalability challenges of RLHF and the potential of RLAIF to address them, it is essential to explore how well RLAIF adapts to diverse architectures and model sizes. Our study aims to address these gaps by systematically evaluating RLAIF's effectiveness across T5, Phi 3-5, and LLaMa 3.2, representing a range of model architectures. Specifically, we aim to investigate the following objectives:

1. **Performance Variation Across Model Families:** How does RLAIF improve model alignment in different architectures compared to traditional SFT?
2. **Scaling Effects:** How does RLAIF's effectiveness vary with model size? Does it show better performance with larger models, or is its effectiveness independent of model size?

Through these evaluations, we aim to offer deeper insights into the applicability of RLAIF across different model scales and architectures. Our findings will contribute to advancing the discussion on scaling reinforcement learning techniques to align large language models effectively and efficiently, providing a clearer path for future research in this domain.

2. LITERATURE REVIEW

2.1. Encoder-Decoder Architecture

The original Transformer model [8] follows an encoder-decoder architecture, commonly used for sequence-to-sequence tasks such as machine translation and summarization. The encoder maps an input sequence to a continuous representation, which the decoder then processes to generate an output sequence. The encoder consists of multiple identical layers, each with self-attention and feed-forward neural networks, while the decoder includes an additional cross-attention mechanism to attend to encoder outputs.

The introduction of attention mechanisms, particularly the self-attention mechanism in Transformers mitigated these challenges by enabling models to focus on different parts of the input sequence simultaneously. Self-attention assigns varying importance to different tokens in the input, allowing for a more dynamic representation of contextual relationships. It addresses the long-range dependency problems that Long Short-Term Memory (LSTM) networks [6] and Gated Recurrent Networks (GRUs) [7] suffer from.

2.2. T5 (223M) – Encoder-Decoder Architecture (Seq2Seq Model)

T5 was originally created by Raffel et al [6]. It is a transformer-based encoder-decoder model, which makes it particularly suitable for translation and summarisation tasks. The pre-training objective for these models is commonly a denoising autoencoder uses masked span prediction. Masked span prediction involves randomly masking spans of text within an input sequence and training the model to predict the missing content. This helps the model learn strong contextual

representations and improves its ability to generate meaningful text.

2.3. Encoder-Only models

Encoder-Only architectures such as BERT [7] are designed primarily for representation learning and downstream classification tasks. These models utilize bidirectional self-attention, allowing them to capture context from both past and future tokens in a sentence. As a result, they excel in tasks such as text classification, named entity recognition and question answering.

2.4. Decoder-Only models

Decoder-only architectures, such as GPT [9] focus on autoregressive text generation. These models use a unidirectional self-attention mechanism, where each token attends only to previous tokens, making them suitable for generative tasks like language modeling and text completion. Unlike the encoder--decoder setup, decoder - only models generate text iteratively, predicting one token at a time. To enforce and prevent tokens from attending to future token, masked self- attention is applied. This ensures that a token at position t can only attend to position ≤ t, maintaining the autoregressive nature of the model.

2.4.1. Phi-3.5 – Compact, Instruction-Tuned Autoregressive Model

Phi-3.5 (3.8B)[10] is a decoder-only Transformer model, optimized for instruction-following and reasoning tasks. Unlike standard language models trained only on generic text, Phi-3.5 undergoes instruction tuning, where it is fine-tuned on datasets designed to enhance its ability to follow prompts, answer complex queries, and generate structured responses. The model employs causal language modeling (CLM) as its primary training objective. Phi-3.5 is particularly effective for structured text generation, including summarization tasks.

2.4.2. LLaMA 3.2 – A Scalable Autoregressive Transformer

LLaMA 3.2 (1B), developed by Meta AI [11], is a decoder-only Transformer model designed for high-quality text generation. It follows the causal language modeling (CLM) paradigm, where the model predicts the next token in a sequence given the previous context. This autoregressive nature enables LLaMA 3.2 to generate coherent, contextually relevant text by progressively extending input sequences. Due to its lightweight design, LLaMA 3.2 (1B) is well-suited for chatbots, summarization, and multilingual text processing, offering a balance between efficiency and language modeling capability.

2.5. Enhancing model performance

To improve the performance of decoder-only models in various NLP tasks, multiple strategies have been developed. These techniques focus on optimizing model outputs, fine-tuning for specific use cases, and improving generalization capabilities.

2.5.1. Prompting (Zero-shot prompting and few-shot prompting)

Zero-shot prompting [12] refers to leveraging a model's pre-trained knowledge to generate responses without additional fine-tuning. The model is provided with a prompt and must generate an appropriate output solely based on its training data. While effective in many scenarios, zero- shot prompting can sometimes produce inaccurate or overly generic responses, as the model has not been explicitly trained for specific tasks. In few-shot prompting [12], the model is provided with a small set of example inputs and outputs to guide its response generation. By including these

examples within the prompt, the model can better understand the desired format and produce more relevant answers. This technique improves accuracy compared to zero-shot prompting but requires well-crafted examples to function optimally.

2.5.2. Supervised Fine-Tuning

Supervised fine-tuning (SFT)[13] involves training a model on a labeled dataset to improve its performance on specific tasks. In this process, the model is fine-tuned on domain-specific data, enabling it to generate more accurate and contextually appropriate responses. SFT helps models adapt to structured tasks like summarization, dialogue generation, and information retrieval, but requires high-quality labeled datasets and significant computational resources.

Low-Rank Adaptation (LoRA)[14] is a parameter-efficient fine-tuning technique that enables adaptation without modifying all model weights. Instead of updating the entire network, LoRA inserts low-rank matrices into existing weight layers and fine-tunes only these additional parameters. This reduces memory and computational costs, making LoRA a practical alternative to full fine-tuning, especially for large models.

2.5.3. Reinforcement Learning

2.4.3.1 Reinforcement Learning with Human Feedback (RLHF)

Reinforcement Learning (RL) techniques have been increasingly used to optimize large language models (LLMs) by incorporating human feedback or reward-based mechanisms. RLHF [1], [2] enhances model alignment with human preferences by leveraging human annotations to refine generated outputs. The process consists of pretraining, where the model is trained using standard supervised learning on large text corpora, followed by reward model training, in which human annotators rank multiple responses generated by the model to create a dataset used to train a reward model that assigns a score to new outputs. The final stage, policy optimization, fine-tunes the model using Proximal Policy Optimization (PPO) [15] or similar reinforcement learning algorithms to maximize the reward score assigned by the trained model. However, RLHF has limitations, including biases in human preferences that may lead to skewed outputs, high computational expenses associated with training and optimizing, and potential mode collapse, where the model generates overly safe or generic responses, reducing diversity and creativity in text generation.

Reward modeling [16] is an essential component of RLHF, where a separate neural network is trained to predict human preference scores. Instead of direct reinforcement learning, the reward model serves as an intermediary, guiding the primary language model to produce better responses. Reward modeling helps mitigate human annotation costs by automating the evaluation process, though it remains susceptible to biases introduced during training.

2.4.3.2 Reinforcement Learning with AI Feedback (RLAIF)

Reinforcement Learning with AI Feedback (RLAIF, Figure 1) [3] replaces human annotators with an AI-based reward model to reduce dependence on human labor. Instead of using human-ranked outputs, a secondary AI system evaluates and assigns rewards based on predefined criteria. This method offers scalability and efficiency advantages over RLHF but comes with its own challenges, such as model alignment issues where AI evaluators may introduce unintended biases, leading to suboptimal reward assignment, and the loss of human intuition, as AI feedback may struggle with subjective or context-dependent assessments [4].

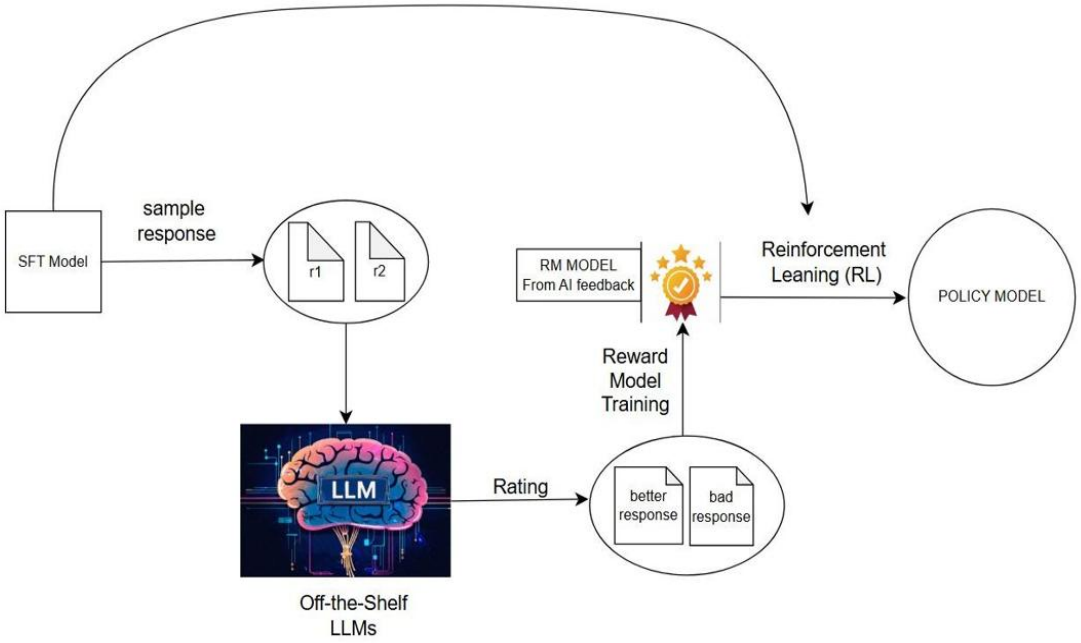

Figure 1. RLAIF technique in detail.

3. METHODOLOGY

This section describes the techniques used to generate preferences, the rationale for models of different scales, the reinforcement learning setup and evaluation metrics.

3.1. Data

We used the following datasets for our experiment:

- Reddit TLDR-17 [17] - comprehensive corpus compiled from Reddit posts between 2006 and 2016 accompanied by the summaries of the post (See Figure 2).
- Reddit TLDR-17 preferences [17] - a dataset created from a subset of Reddit TLDR-17. Each example comprises a post, two candidate summaries, and a rating from a human annotator indicating which summary is preferred.

In this study, we explored how RLAIF performs using a dataset rich in real–world language, such as Reddit posts. To maintain comparability and validate their findings across different model architectures, we use the same type of dataset as used by Lee et al [4]. This ensures that any observed differences in performance are due to variations in model scale rather than inconsistencies in data.

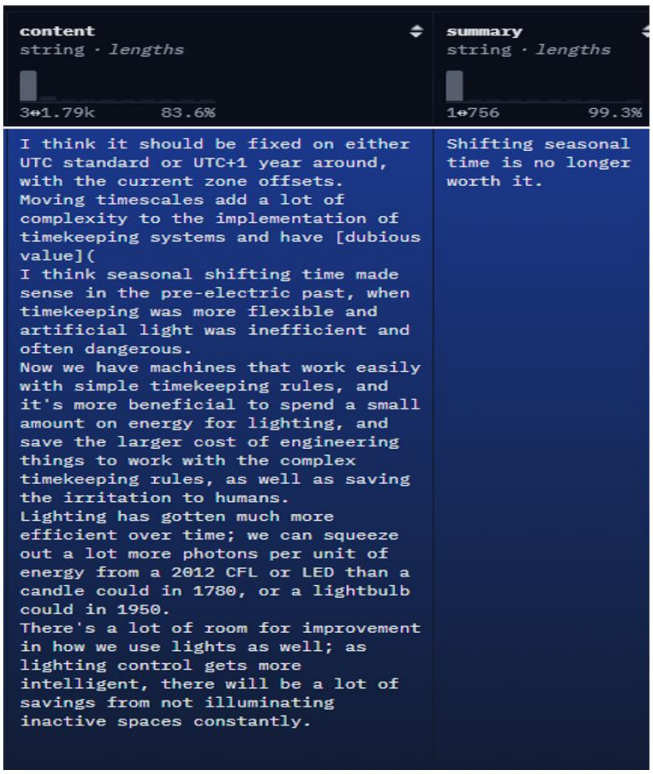

Figure 2. An example row of the TLDR-Reddit dataset. This corpus contains preprocessed posts from the Reddit dataset (Webis-TLDR-17). The dataset consists of 3,848,330 posts with an average length of 270 words for content, and 28 words for the summary. Content is used as document and summary is used as summary.

3.2. Model Selection

To evaluate the effectiveness of Reinforcement Learning from AI feedback (RLAIF) across different model architecture scales, we selected three distinct language models: T5, Phi-3.5 and Llama 3.2. These models were chosen to analyze how RLAIF performs under varying training paradigms, model sizes and architectures.

T5 was included in our selection, because unlike the other models (e.g Palm XS in Lee et al [4]), it follows a Seq2Seq structure rather than a decoder-only design. This allowed us to assess how RLAIF performs on a non-autoregressive model, as Seq2Seq architectures excel in structured tasks such as summarization and translation. Additionally, T5 is widely used for structured NLP applications, making it an ideal candidate for evaluating the impact of RLAIF beyond free-form text generation.

LLaMA 3.2 was chosen as a mid-scale model to assess RLAIF's impact on an extensively pre-trained and well-aligned architecture. By comparing its performance to both a smaller (T5) and a larger (Phi-3.5) model, we were able to evaluate whether RLAIF scales effectively and provides meaningful improvements across different model sizes.

Phi-3.5, the largest model in our selection, represents a high-resource setting optimized for instruction-following and structured text generation. Since it has already undergone fine-tuning with AI-generated data, evaluating RLAIF on Phi-3.5 helps determine whether additional feedback-

driven refinement leads to further performance gains. By comparing it to LLaMA 3.2 and T5, we were able to assess whether RLAIF offers greater benefits to larger-scale models or if its impact is more pronounced in smaller architectures.

3.3. Prompting

We adopted the Detailed + Chain-of-Thought (CoT) Zero-Shot prompting method (see Figure 3), which achieves the highest accuracy of 78.0% for summary tasks across our three models [4]. This approach enhances reasoning by guiding the model through intermediate steps while requiring no task-specific examples. By leveraging detailed instructions combined with CoT reasoning, we improve performance on complex tasks without additional fine-tuning.

Preamble	A good summary is a shorter piece of text that has the essence of the original. It tries to accomplish the same purpose and conveys the key information from the original post. Below we define four evaluation axes for summary quality: coherence, accuracy, coverage, and overall quality. Coherence: This axis answers the question "how coherent is the summary on its own?" A summary is coherent if it's easy to understand when read on its own and free of English errors. A summary is not coherent if it's difficult to understand what the summary is trying to say. Generally, it's more important that the summary is understandable than it being free of grammar errors. Accuracy: This axis answers the question "does the factual information in the summary accurately match the post?" A summary is accurate if it doesn't say things that aren't in the article, it doesn't mix up people, and generally is not misleading. Coverage: This axis answers the question "how well does the summary cover the important information in the post?" A summary has good coverage if it mentions the main information from the post that's important to understand the situation described in the post. A summary has poor coverage if someone reading only the summary would be missing several important pieces of information about the situation in the post. A summary with good coverage should also match the purpose of the original post (e.g. to ask for advice). Overall quality: This axis answers the question "how good is the summary overall at representing the post?" This can encompass all of the above axes of quality, as well as others you feel are important. If it's hard to find ways to make the summary better, the overall quality is good. If there are lots of different ways the summary can be made better, the overall quality is bad. You are an expert summary rater. Given a piece of text and two of its possible summaries, explain which summary best adheres to coherence, accuracy, coverage, and overall quality as defined above.
Sample to annotate	Text – {text} Summary 1 – {summary1} Summary 2 – {summary2}
Ending	Consider the coherence, accuracy, coverage, and overall quality of each summary and explain which one is better. Rationale:

Figure 3. Example of an AI feedback prompt with Chain-of-Thought Zero-shot prompting.

3.4. Model Training

All SFT models were trained using the LoRA [14] method to reduce computational overhead while maintaining performance because of our limited resources. Fine-tuning was performed on the training set of Reddit TLDR-17 with a batch size of 64 for 80 epochs. We used the Adafactor

[18] optimizer with a learning rate of 10−5. The LoRA method was applied to the attention layers, reducing the number of trainable parameters. The maximum input and output lengths were set to 1024 and 64 tokens, respectively.

Reward models (RMs) were initialized from a T5 checkpoint, ensuring consistency across all models. We fine-tuned the RM on the full training split of a preference dataset, where labels reflect AI preferences for AI feedback RMs. Training followed a ranking loss approach with a sigmoid activation function [16], effectively optimizing the RM to differentiate between preferred and less preferred outputs by maximizing their log-sigmoid score differences. We used the Adafactor optimizer with a learning rate of 10−5 and a batch size of 64, with a maximum input length of 1024

tokens. Training continued until the loss and accuracy curves plateau, typically within 3–4 epochs.

For reinforcement learning, we initialized each task with the corresponding SFT model as the initial policy. All models were trained using Proximal Policy Optimization (PPO)[15]. To encourage exploration, we sampled from the language model policies with a temperature of T=0.9. Training was conducted for 10 epochs with a batch size of 128 and a learning rate of 10^{-5}. We applied a KL divergence penalty with β=0.05 to balance optimization stability. This setup followed the experimental framework of Lee et al. [4].

3.5. Measurements

To evaluate the effectiveness of Reinforcement Learning from AI Feedback (RLAIF) across different model scales, we structured our experiment around two key aspects: performance, scaling effects. We utilized ROUGE and BERTScore as evaluation metrics. Below, we explain how we measure and analyze each aspect.

3.5.1. Performance

We assessed model performance using the following automatic evaluation metrics:

1. ROUGE-1 [19]: Measures the overlap of unigrams (single words) between the generated summary and the reference summary.
2. ROUGE-2 [19]: Measures the overlap of bigrams (two consecutive words) between the generated summary and the reference summary.
3. ROUGE-L [19]: Captures the longest common subsequence (LCS) between the generated summary and the reference summary, reflecting fluency and coherence.
4. BERTScore [20]: Uses embeddings from BERT to compare the semantic similarity between generated and reference summaries, providing a more contextualized assessment of quality.

We compared the performance of models fine-tuned with Supervised Fine-Tuning (SFT) against those trained using RLAIF. We analyzed improvements in scores across different model architectures, highlighting how RLAIF impacts different model families and their ability to generate high-quality summaries. We also examined whether performance gains were consistent across different ROUGE variants and BERTScore, assessing improvements in lexical overlap and semantic similarity.
To calculate the percentage of the improvement between the Supervised Fine-Tuning (SFT) and RLAIF models, we used the following equation 1:

$$Improvement\ (\%) = \frac{RLAIF\ Score - SFT\ Score}{SFT\ Score} * 100 \quad (1)$$

3.5.2. Scaling Effects

We investigated the impact of scaling by comparing models of varying sizes: T5, Phi 3-5, and LLaMA 3.2. The key measurements include:

5. Changes in ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore as model size increases.
6. The relative improvement of RLAIF over SFT across different model sizes and architecture.

We analyzed whether larger models benefit more from RLAIF compared to smaller models. By observing trends in performance across T5, Phi 3-5, and LLaMA 3.2, we determine if scaling

improves the effectiveness of reinforcement learning from AI feedback.

4. RESULTS

4.1.Performance

Table 1. Results show performance improvements across different architectures after applying the RLAIF process.

Metric	ROUGE-1	ROUGE-2	ROUGE-L	BERT Score
SFT T5	0.0433	0.0036	0.0389	0.3288
RLAIF T5	**0.0768 (+77.2%)**	**0.0053 (+47.2%)**	**0.0684 (+75.8%)**	**0.4030 (+22.6%)**
SFT Llama 3.2 (1B)	0.0593	0.0059	0.0701	0.3688
RLAIF LLama (1B)	**0.1063 (+79.3%)**	**0.0111 (+89%)**	**0.0932 (+32.9%)**	**0.4129 (+12.0%)**
SFT Phi-3.5 (3.8B)	0.0593	0.0299	0.1093	0.3705
RLAIF Phi-3.5 (3.8B)	**0.0826 (+39.3%)**	**0.0388 (+29.8%)**	**0.2090 (+91.1%)**	**0.4285 (+15.6%)**

4.1.1.ROUGE-1 (Unigram Content Coverage)

The transition from SFT to RLAIF results in the most substantial improvement for the LLaMA 3.2 model, with RLAIF boosting the ROUGE-1 score by 79.3% (from 0.0593 to 0.1063) (Table 1). This substantial gain indicates that RLAIF significantly enhances the LLaMA 3.2 model's ability to capture key individual terms from reference summaries, pointing to an improved content selection capability. In contrast, the improvements for Phi 3-5 and T5 were more modest, at 39.3% and 77.2%, respectively.

4.1.2.ROUGE-2 (Phrase-Level Accuracy)

All models exhibit relatively low ROUGE-2 scores (Table 1), except for Phi 3-5, which achieves a decent score of 0.0389. However, the most significant improvement is observed in the LLaMA

3.2 model, where RLAIF enhances performance by 89% (from 0.0059 to 0.0111). This substantial gain suggests that RLAIF significantly improves the model's ability to retain multi-word expressions, advancing from simply recognizing individual words to effectively preserving meaningful phrases.

4.1.3.ROUGE-L (Sequential Coherence)

The Phi-3.5 model demonstrates the most remarkable improvement in ROUGE-L when trained with RLAIF, showing a 91.1% increase (from 0.1093 to 0.2090) (Table 1). This substantial gain indicates that RLAIF dramatically enhances Phi-3.5's ability to maintain coherent sequences that match reference summaries, suggesting improved narrative flow and structural coherence.

4.1.4.BERTScore (Semantic Similarity)

All models show substantial improvement in BERTScore when using RLAIF (Table 1). The T5 model demonstrates the largest relative improvement with a 22.6% increase (from 0.3288 to 0.4030), highlighting how RLAIF considerably enhances T5's semantic understanding capabilities. However, the Phi-3.5 model achieves the highest absolute BERTScore (0.4285) after RLAIF training, representing a 15.7% improvement over its SFT baseline. This indicates that while T5 shows the greatest relative semantic gains, Phi-3.5 ultimately delivers superior semantic fidelity in its generated outputs.

4.2.Scaling effects

Table 2. Results demonstrate improvements across models in relation to scaling effects.

Model	ROUGE-1 Improvement	ROUGE-2 Improvement	ROUGE-L Improvement	BERTScore Improvement
T5 (Small; 738M)	+77.2%	+47.2%	+75.8%	+22.6%
LLaMA 3.2 (Medium; 1B)	+79.3%	+89%	+32.9%	+12.0%
Phi- 3.5(Large, 3.8B)	+39.3%	+29.8%	+91.1%	+15.6%

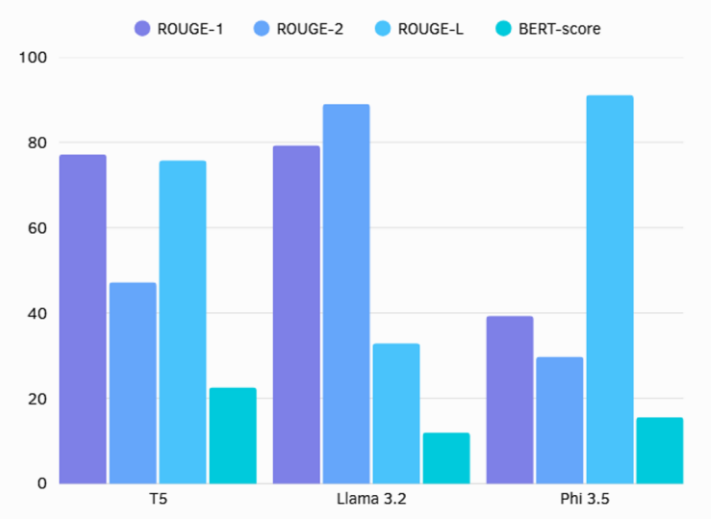

Figure 4. Results per model across all three evaluation metrics.

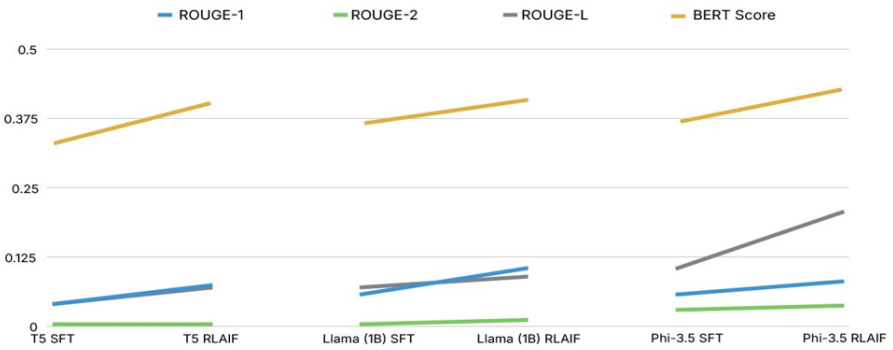

Figure 5. The diagrams depict how performance and improvement vary with scale.

4.3.Cross-Metric Analysis

LLaMA 3.2 shows the most dramatic improvements in ROUGE-1 and especially ROUGE-2, suggesting that RLAIF particularly enhances this model's lexical precision and ability to maintain important phrases from reference texts (See Figure 4 and 5).

Phi-3.5 demonstrates the most significant gains in ROUGE-L and achieves the highest absolute BERTScore, with only modest improvements in ROUGE-1 and ROUGE-2. This suggests that RLAIF has substantially enhanced the model's ability to generate coherent and fluent sequences while preserving the semantic integrity of the text. The model excels in maintaining overall meaning and structural alignment, even when it does not match individual words or phrases exactly. These attributes are crucial for high-quality summarization, where the focus is on conveying the essence of the content, rather than merely replicating specific word choices (See table 2).

T5 shows the largest relative improvement in BERTScore (+22.6%), while demonstrating consistent improvements across other metrics. This suggests that RLAIF particularly enhances T5's semantic understanding capabilities, even though its absolute performance remains below that of larger models. (See table 2)

5. DISCUSSION

The results reveal that while RLAIF improves performance across all model scales, the nature and magnitude of improvement varies significantly by architecture. The most dramatic transformations occur in:

1. LLaMA 3.2 with RLAIF for lexical precision (ROUGE-1) and especially phrase preservation (ROUGE-2).
2. Phi-3.5 with RLAIF for sequential coherence (ROUGE-L) and highest absolute semantic similarity (BERTScore).
3. T5 with RLAIF for relative improvement in semantic understanding (largest percentage gain in BERTScore).

These findings suggest that larger models like Phi-3.5 particularly benefit from RLAIF in aspects related to higher-order language understanding, such as coherence and semantic fidelity. Phi-3.5's use of synthetic data helps fine-tune its performance in these areas, ensuring strong results in structural alignment and semantic preservation. Besides, Phi-3.5 has more parameters, which means they have a greater capacity to learn complex patterns and relationships in data. This makes them particularly adept at tasks that require higher-order language understanding, such as

maintaining structural coherence and semantic fidelity over long text sequences. The additional parameters allow the model to capture nuances in language and better maintain logical flow and meaning across sentences, paragraphs, or even entire documents. Meanwhile, medium-scale models with lower parameters like LLaMA 3.2 show remarkable improvements in lexical and phrasal accuracy, while smaller models like T5 demonstrate significant relative gains in semantic understanding when trained with RLAIF, despite starting from a lower baseline.

Notably, T5 starts with the lowest ROUGE variants and the lowest BERTScore among the three models, yet it shows significant improvement across all metrics. After training, T5 achieves a BERTScore that is quite close to those of the other models, highlighting the efficiency of RLAIF in enhancing semantic understanding in smaller models. This is particularly evident in BERTScore, where T5 shows the highest relative improvement (+22.6%). This is expected given T5's encoder-decoder architecture, which is particularly suited for tasks like summarization. Despite its initially lower performance, T5 benefits greatly from the RLAIF approach, making substantial strides in semantic similarity, thereby demonstrating that even smaller models can achieve considerable improvements in understanding and summarization tasks when trained effectively.

In contrast, larger models like LLaMA 3.2 and Phi-3.5 achieve high performance in their respective strengths: Phi-3.5 excels at maintaining structural coherence and semantic fidelity (ROUGE-L and BERTScore), benefiting from synthetic data training, while LLaMA 3.2 makes substantial progress in lexical precision (ROUGE-1) and phrase retention (ROUGE-2). The findings suggest that RLAIF can significantly enhance different aspects of language generation depending on the model size and architecture, with smaller models benefiting the most from semantic improvements, while larger models excel in coherence and structural alignment.

6. CONCLUSION AND FUTURE WORK

In this paper, we set out to explore how well Reinforcement Learning from AI Feedback (RLAIF) works across different types of language models and sizes, and how it compares to traditional supervised fine-tuning (SFT). We tested RLAIF on three models—T5, Phi 3-5, and LLaMA 3.2—which differ in both architecture and scale, to understand how RLAIF impacts each model's performance.

Our results showed that RLAIF improves performance for all model sizes, but the improvements varied depending on the architecture. LLaMA 3.2 saw major gains in lexical accuracy and phrase retention, Phi 3-5 improved the most in structural coherence and semantic accuracy, and T5 made the biggest leap in semantic understanding. These findings suggest that RLAIF is effective across the board, but works differently depending on the model. Smaller models like T5 showed the most relative improvement in understanding and summarization tasks, while larger models like Phi-3.5 performed better in tasks requiring structural coherence.

1. Performance Across Different Models: RLAIF improved model performance across all three architectures. For Phi-3.5, the biggest gains were seen in maintaining coherence and preserving meaning, while T5 made impressive strides in understanding and summarizing text. LLaMA 3.2 showed the most progress in lexical precision and keeping phrases intact.

2. Effect of Model Size: The size of the model influenced how RLAIF worked. Larger models like Phi-3.5 excelled at maintaining coherence and structure, while

LLaMA 3.2 showed solid gains in lexical precision. T5, which is a smaller model, saw the biggest improvement in semantic understanding, with the largest relative boost in BERTScore.

RLAIF proves to be highly efficient but not consistent across different model architectures, with each architecture benefiting from tailored improvements suited to their strengths, such as summarization for T5 and coherence for Phi-3.5. This highlights the need for a deeper exploration of how RLAIF interacts with various model types and sizes. Additionally, while RLAIF demonstrates immediate improvements, it remains unclear whether these gains will hold consistently across time or diverse tasks, which requires further validation.

In short, RLAIF works well for all model sizes, but the type of improvement varies depending on the model's strengths. Larger models improve in aspects like structure and coherence, while smaller models benefit the most in terms of understanding and semantic accuracy.

In future work, we aim to explore various directions to improve model performance and robustness. One such direction is fully training the model parameters instead of using LoRA (Low-Rank Adaptation) to evaluate whether this approach yields significant gains. Fully training may unlock the model's full potential, improving accuracy and efficiency, especially in summarization and content generation tasks.

We also plan to train for more epochs to determine whether extended training enhances performance. This will reveal if the model has reached optimal performance or if further training refines its coherence and semantic accuracy.

Future research will aim to address the limitations and explore several directions to further validate and enhance the effectiveness of RLAIF. Given that the results varied depending on the model architecture, an important step would be to test RLAIF on a broader range of model families, including newer architectures and those optimized for specific domains. This will help in understanding whether RLAIF's performance improvements can be generalized to other LLMs, especially those fine-tuned for specialized tasks.

Moreover, while RLAIF shows promise in improving performance across different architectures, further research is needed to investigate the long-term stability of these improvements and their generalizability across various tasks. Extending the scope to newer or emerging models, such as Claude 4 and other advanced architectures, will provide valuable insights into the scalability and applicability of RLAIF in a wider context.

In terms of performance, models trained with direct Reinforcement Learning from AI Feedback (d-RLAIF) are expected to outperform those trained with traditional RLAIF, especially in tasks that demand high-quality text generation. Unlike standard RLAIF, which typically relies on a reward model trained using human preferences, d-RLAIF bypasses this step by directly using AI- generated feedback to optimize the model's responses. This direct reward learning strategy simplifies the pipeline and may lead to more efficient training, enabling faster convergence and potentially improving output quality by ensuring a tighter alignment between the model's training objectives and the intended task. However, improvements depend on model architecture, task complexity, and configuration. Thus, thorough evaluation using appropriate accuracy and quality metrics is essential.

By incorporating d-RLAIF across different models, we aim to assess whether it provides measurable advantages over traditional RLAIF. Comparing identical architectures trained with both approaches will help determine the effectiveness of d-RLAIF in enhancing generation quality and

training efficiency. These efforts will offer insights into the scalability of RLAIF and its potential to boost the performance of models like Claude 4 across diverse applications.

These future efforts will provide valuable insights into the scalability and adaptability of RLAIF, allowing us to explore its potential for broader use cases and more advanced model architectures.

REFERENCES

[1] N. Stiennon et al., "Learning to summarize from human feedback," Feb. 15, 2022, arXiv: arXiv:2009.01325. doi: 10.48550/arXiv.2009.01325.

[2] L. Ouyang et al., "Training language models to follow instructions with human feedback," Mar. 04, 2022, arXiv: arXiv:2203.02155. doi: 10.48550/arXiv.2203.02155.

[3] Y. Bai et al., "Constitutional AI: Harmlessness from AI Feedback," Dec. 2022, [Online]. Available: http://arxiv.org/abs/2212.08073

[4] H. Lee et al., "RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback," 2024.

[5] P. Pezeshkpour and E. Hruschka, "Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions," Aug. 22, 2023, arXiv: arXiv:2308.11483. doi: 10.48550/arXiv.2308.11483.

[6] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Sep. 19, 2023, arXiv: arXiv:1910.10683. doi: 10.48550/arXiv.1910.10683.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[8] A. Vaswani et al., "Attention Is All You Need," in 31st Conference on Neural Information Processing Systems (NIPS 2017), Jun. 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[9] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training".

[10] "Phi-3.5 SLMs," TECHCOMMUNITY.MICROSOFT.COM. Accessed: Mar. 17, 2025. [Online]. Available: https://techcommunity.microsoft.com/blog/azure-ai-services-blog/discover- the-new-multi-lingual-high-quality-phi-3-5-slms/4225280

[11] "Llama 3.2: Revolutionizing edge AI and vision with open, customizable models," Meta AI. Accessed: Mar. 17, 2025. [Online]. Available: https://ai.meta.com/blog/llama-3-2-connect- 2024-vision-edge-mobile-devices/

[12] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.11903

[13] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," May 23, 2018, arXiv: arXiv:1801.06146. doi: 10.48550/arXiv.1801.06146.

[14] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 16, 2021, arXiv: arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685.

[15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," Aug. 28, 2017, arXiv: arXiv:1707.06347. doi: 10.48550/arXiv.1707.06347.

[16] H. Zhou, C. Wang, Y. Hu, T. Xiao, C. Zhang, and J. Zhu, "Prior Constraints-based Reward Model Training for Aligning Large Language Models," Sep. 18, 2024, arXiv: arXiv:2404.00978. doi: 10.48550/arXiv.2404.00978.

[17] M. Völske, M. Potthast, S. Syed, and B. Stein, "TL;DR: Mining Reddit to Learn Automatic Summarization," in Proceedings of the Workshop on New Frontiers in Summarization, Copenhagen, Denmark: Association of Computational Linguistics, Sep. 2017, pp. 59–63.

[18] N. Shazeer and M. Stern, "Adafactor: Adaptive Learning Rates with Sublinear Memory Cost," Apr. 11, 2018, arXiv: arXiv:1804.04235. doi: 10.48550/arXiv.1804.04235.

[19] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in Text Summarization Branches Out, Barcelona, Spain, Jul. 2004, pp. 74–81.

[20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," Feb. 24, 2020, arXiv: arXiv:1904.09675. doi: 10.48550/arXiv.1904.09675.

[21] Wang, H. (2025). Efficient and robust reinforcement learning from human feedback. Proceedings of the AAAI Conference on Artificial Intelligence, 39(27), 28730–28730.

https://doi.org/10.1609/aaai.v39i27.35123
[22] Lande, J. (2023, September 7). Google research explores: Can AI feedback replace human input for effective reinforcement learning in large language models? MarkTechPost. https://www.marktechpost.com/2023/09/07/google-research-explores-can-ai-feedback-replace-human-input-for-effective-reinforcement-learning-in-large-language-models
[23] Curuksu, J. (2025, April 4). Fine-tune large language models with reinforcement learning from human or AI feedback. AWS Machine Learning Blog. https://aws.amazon.com/blogs/machine-learning/fine-tune-large-language-models-with-reinforcement-learning-from-human-or-ai-feedback