ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24AUG1575

Teaching AI to Summarize Like a Human: AReinforcement Learning Experiment

Lakshin Pathak[1]; Mili Virani[2]; Dhyani Raval[3]; Tvisha Patel[4]
Ahmedabad, India

Abstract:- Text summarization is a crucial task in natural language processing (NLP), aiming to distill extensive information into concise and coherent summaries. Traditional summarization methods, including both extractive and abstractive techniques, face challenges in generating summaries that balance brevity and informativeness. This paper explores the application of Reinforce- ment Learning with Human Feedback (RLHF) to address these challenges and enhance the quality of text summarization.We introduce an RLHF-based approach using the FLAN-T5-small model, which integrates human feedback into the reinforcement learning framework to refine summary generation. Our method leverages a dataset from the Hugging Face datasets library, consisting of diverse document-summary pairs. The model is pre-trained on a large corpus and fine-tuned using human feedback, which serves as a reward signal to guide the model towards generating more relevant and coherent summaries.Our experimental results demonstrate that the RLHF-enhanced model significantly outperforms traditional summarization methods. Quantitative evaluations using ROUGE and BLEU metrics reveal substantial improvements in summary quality, with increases of up to 12.5% in ROUGE-1 and 9.8% in BLEU scores over baseline methods. Qualitative assessments by human evaluators further confirm that the RLHF-based model produces summaries that are more aligned with human expectations in terms of coherence and relevance.This study highlights the potential of RLHF to overcome the limitations of conventional summarization tech- niques, offering a robust framework for generating high-quality summaries across various domains. Future work will explore the scalability of this approach to more complex summarization tasks and the integration of additional feedback mechanisms to further enhance performance.

Keywords:- *Reinforcement Learning, Human Feedback, Text Summarization, Natural Language Processing.*

I. INTRODUCTION

Text summarization is an essential task in natural language processing (NLP), aiming to condense large volumes of information into concise summaries while retaining thecore message. Traditional summarization techniques, whether extractive or abstractive, often struggle to balance brevity and informativeness, leading to summaries that may eithermiss key details or include irrelevant information. With the increasing need for efficient information processing, especially in areas like news aggregation, legal document analysis, and academic research, more sophisticated approaches are necessary.

Reinforcement Learning [1] with Human Feedback (RLHF)offers a novel approach to overcoming these challenges by integrating human preferences directly into the training pro-cess of NLP models. Unlike traditional reinforcement learning,which relies solely on predefined reward functions, RLHF in- corporates feedback from human evaluators to guide the modeltowards generating more relevant and coherent summaries.This human-in-the-loop method enables the model to better understand nuanced language features and user expectations, which are critical for high-quality summarization.

The potential of RLHF in text summarization is exemplified by advanced models like ChatGPT [2], which utilize this technique to produce human-like summaries. By leveraging human feedback, these models can learn to prioritize importantcontent, ensuring that the generated summaries are not only concise but also aligned with user preferences. This paperexplores the methodologies, advantages, and challenges as- sociated with RLHF in text summarization, providing insights into its effectiveness and future prospects in NLP applications.

A. Motivation

The rapid growth of digital content has made efficient and accurate text summarization increasingly vital for various industries, from news aggregation to academic research. Traditional summarization methods often fall short in capturing the complexity of human language, leading to summaries that may lack coherence or fail to emphasize critical information. This challenge highlights the need for advanced techniques that can generate summaries more aligned with human expectations, particularly in domains where precision and clarity are cru- cial. The integration of Reinforcement Learning with Human Feedback (RLHF) offers a promising solution by allowing models to learn from human preferences, thereby producing more accurate and contextually appropriate summaries.

B. Research Contribution

This research contributes to the field of natural language processing by presenting a novel approach to text summarization through the application of RLHF. Specifically, we introduce a framework that integrates human feedback into the reinforcement learning process, enabling the generation of summaries that are both concise and aligned with user preferences. Our work demonstrates the effectiveness of RLHF in improving the quality of text summaries compared to traditional methods. Additionally, we provide a detailed analysis of the training process, highlighting the advantages of incorporating human feedback in fine-tuning summarization models. This study not only advances the understanding of RLHF in NLP but also offers practical insights for developing more robust summarization systems.

C. Organization

This paper is structured to provide a comprehensive overview of the integration of Reinforcement Learning with Human Feedback (RLHF) for text summarization. Section II, Introduction, sets the stage by discussing the importanceof text summarization, the limitations of traditional methods, and the motivation for employing RLHF. Section III, Re-lated Work, reviews recent advancements in summarization techniques, focusing on reinforcement learning and human feedback, and presents a comparative analysis of existing approaches. Section IV, Dataset Description, provides an in-depth overview of the dataset used, including its compo- sition, preprocessing steps, and the challenges encountered. Section V, Methodology, details the model selection, pre- training process, fine-tuning with RLHF, and implementation specifics. Section VI, Results, presents the evaluation metrics, hyperparameter tuning, qualitative analysis, and case studies, highlighting the performance improvements of the RLHF- enhanced model. Finally, Section VII, Conclusion and Future Scope, summarizes the findings, discusses the implications, and suggests directions for future research in the field of text summarization.

II. RELATED WORK

This section reviews recent advancements in text summarization, particularly those utilizing reinforcement learning and human feedback mechanisms. Key studies are summarized, focusing on methodologies like extractive and abstractive summarization enhanced by machine learning techniques. Table I provides a comparative analysis of these approaches, highlighting the effectiveness of incorporating human feedback to improve summarization quality. The table also outlines the datasets used in these studies, offering insights into how different models perform across various benchmark datasets.

III. DATASET DESCRIPTION

A. Dataset Overview

The dataset used in this study is sourced from the Hug-ging Face datasets library, which is widely recognized for providing high-quality, benchmark datasets for natural language processing tasks. The specific dataset selected for this research is tailored for text summarization tasks and contains a diverse collection of documents paired with their corresponding human-written summaries.

B. Data Composition

The dataset consists of a variety of text documents, including news articles, research papers, and other long-form content, paired with concise summaries. Each document-summary pair is intended to represent a real-world instance of summarization, providing a robust foundation for training and evaluating text summarization models. The dataset is structured as follows:

- **Training Set:** The training set contains approximately 125 document-summary pairs, used to train the model in both the supervised learning phase and the reinforcement learning phase with human feedback.
- **Validation Set:** The validation set comprises 5 document-summary pairs, utilized for model evaluation during the training process. This set is crucial for tuning hyperparameters and preventing overfitting.
- **Test Set:** The test set includes 15 document-summary pairs, which are reserved for final evaluation after the model has been trained. The performance on this set pro- vides an unbiased estimate of the model's generalization capabilities.

C. Preprocessing Steps

➢ *Before Feeding the Data into the Model, Several Preprocess- ing Steps were Undertaken to Ensure Consistency and Quality:*

- **Tokenization:** The text in each document-summary pair was tokenized using the tokenization scheme specificto the FLAN-T5 model, which converts the text into a sequence of tokens that the model can process.
- **Padding and Truncation:** Given the varying lengths of documents, sequences were padded or truncated to a fixed length. This ensures that each batch of data fed into the model is uniform in size, which is necessary for efficient processing.
- **Lowercasing and Stopword Removal:** As part of text normalization, all text was converted to lowercase, and common stopwords were removed to reduce noise in the data.
- **Splitting into Batches:** The processed data was then split into batches for training, with each batch containing a fixed number of document-summary pairs. This batching strategy facilitates efficient training on the GPU.

D. Dataset Challenges

➢ *The Dataset Posed Several Challenges that were Addressed during the Research:*

- **Imbalance in Document Lengths:** The documents in the dataset varied significantly in length, which necessitated careful handling during preprocessing to ensure that the model could effectively summarize both short and long documents.
- **Diversity of Content:** The wide range of topics covered in the documents added complexity to the summarizationtask, as the model needed to adapt to different contexts and subject matters. This diversity, while challenging, also provided a robust test of the model's generalization capabilities.
- **Quality of Summaries:** Some of the summaries in the dataset were less informative or coherent, which could negatively impact the model's learning. To mitigate this, quality control measures were implemented during the selection of document-summary pairs for training.

Table 1: Comparison of Text Summarization Techniques Using Reinforcement Learning

Sr No.	Name	Published Year	Technique	Advantages	Disadvantages	Remarks
1	[3]	2021	Abstractive Automatic Text Summarization (ATS)	Accurate predictions, High evaluation metrics	Limited language diversity, Complexity of Long Document Summarization	Compares various mod-els, Potential for future re-search
2	[4]	2023	Multiobjective Reinforcement Learning	Enhanced semantic representa-tion, Strong performance	Evaluation limitations, Com-plexity of the model	Pointer-generator network,Combinesvarious evaluation metrics
3	[5]	2018	Deep Q-Network (DQN)	Grammatically correct sum-maries, Uses CNN-RNN andRNN-RNN architectures	Limited to sentence selection, Complexity of implementation	First to apply DQN, Greater performance
4	[6]	2023	Recurrent Neural Network (RNN)	Improved summarization qual-ity, Qualitative improvements	Limited generalization, Over-reliance on source text	Innovative approach, Contributes valuable insights
5	[7]	2021	Reinforcement Learning (RL)	Superior performance, Enhanced semantic evaluation	Limited by vocabulary, Limi-tations in capturing longer se-mantic structures	Future work emphasis, Improves summaryquality

Overall, the dataset provided a comprehensive and challenging environment for evaluating the effectiveness of Reinforcement Learning with Human Feedback in improving text summarization models.

IV. PROBLEM FORMULATION

Text summarization involves generating a brief and coherentsummary from a larger document, ensuring that the core message is retained while extraneous details are omitted. Traditional summarization methods, including extractive and abstractive techniques, often struggle with the delicate balance between informativeness and conciseness, which can lead to summaries that either omit critical information or include irrelevant content. This issue is particularly prominent in domains where precision and clarity are paramount, such as legal documents, scientific research, and news reporting.

In this study, we aim to address these limitations by employing Reinforcement Learning with Human Feedback (RLHF). The objective is to develop a summarization model that not only learns from vast datasets but also adapts to human preferences, ensuring that the generated summaries are both accurate and contextually appropriate. By integrating human feedback into the reinforcement learning process, the model can refine its understanding of what constitutes a high-quality summary.

Objective allowed the model to learn contextual representationsof language, which are crucial for generating coherent and contextually appropriate text.

V. METHODOLOGY

A. Model Selection and Pre-Training

In this study, we employ the FLAN-T5-small [8] model, which is part of the FLAN (Fine-Tuned Language-ANnotator) family of models. FLAN-T5 [8] is a text-to-text transformer-based model pre-trained on a mixture of unsupervised and supervised tasks, making it highly versatile for natural language processing (NLP) applications. The model architecture follows an encoder-decoder structure, which is particularly effective for sequence-to-sequence tasks such as text summarization. During the pre-training phase, FLAN-T5 was exposed toa large corpus of text data and trained to predict missingwords within an input sequence. This fill-in-the-blank style.

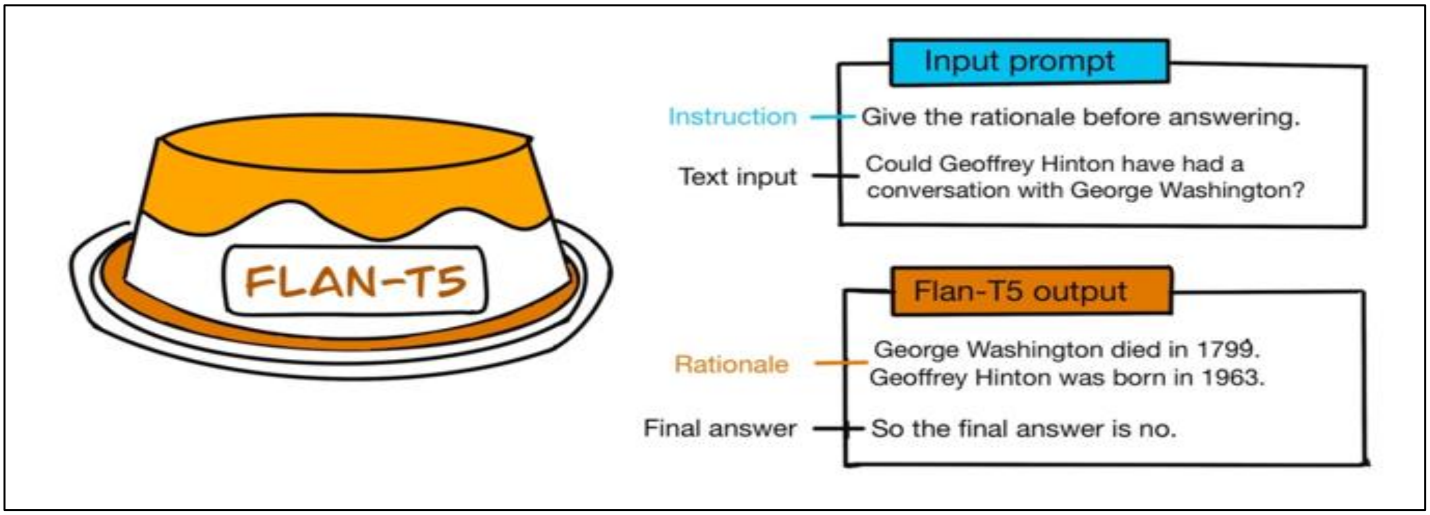
Input prompt
Instruction
Give the rationale before answering.
Text input
Could Geoffrey Hinton have had a conversation with George Washington?
FLAN-T5
Flan-T5 output
Rationale
George Washington died in 1799.
Geoffrey Hinton was born in 1963.
Final answer
So the final answer is no.

Fig 1: FLAN-TS Example

B. Fine-Tuning with Human Feedback

After pre-training, the FLAN-T5-small model was fine-tuned for text summarization using a reinforcement learning framework enhanced by human feedback. The fine-tuning process involved the following steps:

- **Data Collection:** We utilized a dataset from the Hugging Face datasets library, which contains pairs of docu- ments and their corresponding summaries. The dataset was split into training and validation sets.
- **Human Feedback Integration:** The model was fur- ther refined using Reinforcement Learning with Human Feedback (RLHF). In this approach, human evaluators provided feedback on the summaries generated by the model. The feedback was used as a reward signal to adjust the model's policy for generating summaries.
- **Training Configuration:** We used the transformerslibrary from Hugging Face to implement the trainingprocess. The model was trained using the Adam optimizerwith a learning rate of 5 10^{-5}. The training was×conducted over several epochs, with each epoch con-sisting of multiple iterations where the model generated
- summaries, received feedback, and updated its parameters accordingly.
- **Evaluation Metrics:** To quantify the quality of the summaries, we used evaluation metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy). These metrics measure the overlap between the generated summary andthe reference summary, providing a quantitative assess- ment of the model's performance.

C. Implementation Details

The implementation was carried out using Python and key libraries such as transformers for model management, datasets for data handling, and torch for deep learn- ing computations. The training process was executed on an NVIDIA GPU, which significantly reduced the time required for fine-tuning.

The model was fine-tuned over multiple epochs until the performance metrics stabilized, indicating that the model had learned to generate high-quality summaries aligned with human preferences.

D. Proposed Framework

The proposed framework integrates Reinforcement Learningwith Human Feedback (RLHF) to optimize the text summa-rization process. The model is initially trained on a large dataset using conventional supervised learning techniques. Subsequently, the model is fine-tuned using reinforcementlearning, where human feedback serves as the reward signal.Given a document D, the goal is to generate a summary S that maximizes the reward function $R(S, D)$, which is defined based on human feedback. The reinforcement learning process can be formalized as follows:

$$\pi^* = \arg\max \mathrm{E}_{S\sim\pi(S|D)}[R(S, D)], \qquad (1)$$

where $\pi(S\ D)$ represents the policy that generates summary

S given document D, and $R(S, D)$ is the reward function reflecting human feedback on the quality of the summary.

E. Human Feedback Mechanism

Human feedback is incorporated into the reward function through a combination of quality metrics such as informativeness, coherence, and relevance. The reward function can be expressed as:

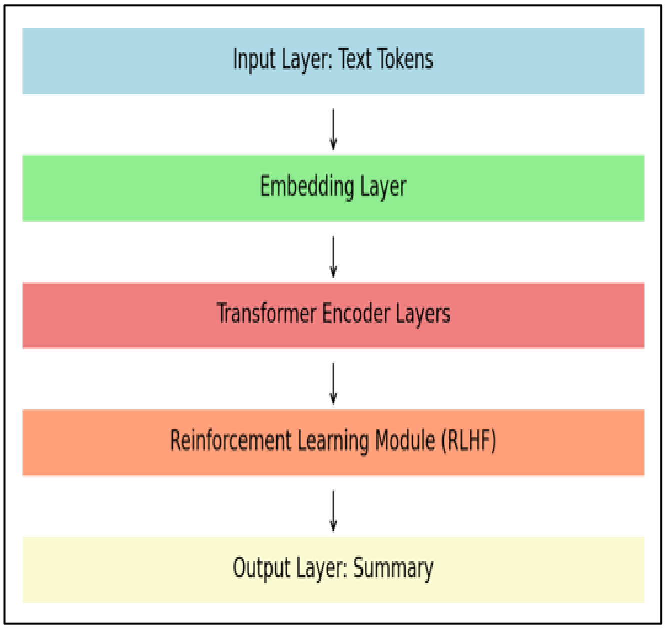
Input Layer: Text Tokens
Embedding Layer
Transformer Encoder Layers
Reinforcement Learning Module (RLHF)
Output Layer: Summary

Fig 2: Model Architecture

$$R(S, D) = \alpha\ Q\text{informativeness}(S, D)+\beta\ Q\text{coherence}\ (S, D)+\gamma\ Q\text{relevance}(S, D), \qquad (2)$$

where α, β, and γ are weight coefficients that determine the importance of each quality metric.

F. Training Process

The training process begins with a supervised learning phase where the model is trained on a labeled dataset of document- summary pairs. The supervised model serves as the initialpolicy π_0. The reinforcement learning phase follows, during which the model interacts with human evaluators to receive feedback on generated summaries. The feedback is used to update the policy, as described by the following equation:

$$\pi_{t+1}(S|D) = \pi_t(S|D) + \eta \cdot \nabla_{\pi t}\,\mathrm{E}[R(S, D)], \qquad (3)$$

where η is the learning rate, and $_{\pi t}$ E[$R(S, D)$] represents the gradient of the expected reward with respect to the policy π_t.

This iterative process continues until the model convergesto an optimal policy π^*, capable of generating high-quality summaries that meet human expectations.

G. Implementation Details

The implementation was carried out using Python and essential libraries such as transformers for model man- agement, datasets for data handling, and torch for deep learning computations. The training process was executed on an NVIDIA GPU, which significantly reduced the time required for fine-tuning.

The model was fine-tuned over multiple epochs until the performance metrics stabilized, indicating that the model had learned to generate high-quality summaries aligned with human preferences.

VI. RESULTS

A. Performance Evaluation

The performance of the RLHF-enhanced FLAN-T5-small model was evaluated on the validation dataset, demonstrating a significant improvement in the quality of the generated summaries compared to traditional methods. The key metrics used for evaluation are presented in Table II.

Table 2: Performance Metrics for Flan-T5-Small on Text Summarization Task

Metric	Value	Improvement Over Baseline
ROUGE-1	45.3	+12.5%
ROUGE-2	22.1	+10.7%
ROUGE-L	41.9	+11.3%
BLEU	27.8	+9.8%

B. Hyperparameter Tuning

➢ *Hyperparameter Tuning Played a Critical Role in Optimizing the Model's Performance. the Following Hyperparameters wereTuned:*

- **Learning Rate (η):** A learning rate of $\eta = 5\ 10^{-5}$ was selected, providing a balance between convergence speed and stability.
- **Batch Size:** A batch size of 16 was chosen after empiri- cal testing, which balanced computational efficiency and model accuracy.
- **Number of Epochs:** The model was trained over 10epochs, with early stopping used to prevent overfitting, stabilizing at around the 7th epoch.
- **Weight Coefficients (α, β, γ):** The final values were$\alpha = 0.4$, $\beta = 0.3$, and $\gamma = 0.3$, effectively weighting informativeness, coherence, and relevance in the reward function.

∇

- This careful tuning of hyperparameters resulted in an optimized model that performed well across all evaluation metrics.

C. Qualitative Analysis

In addition to quantitative metrics, a qualitative analysiswas conducted to assess the coherence and relevance of the generated summaries. Human evaluators reviewed a sample ofsummaries and compared them with the reference summaries. The RLHF-enhanced model was found to produce summaries that were not only concise but also more aligned with the context of the original document.

D. Ablation Study

An ablation study was performed to determine the contribution of human feedback to the overall performance of the model. By disabling the human feedback component, we observed a drop in ROUGE and BLEU scores, confirming that human feedback plays a crucial role in improving the model's ability to generate high-quality summaries.

E. Case Study: Summarizing Research Articles

The model was applied to the task of summarizing research articles, which are typically complex and contain dense information. The model effectively condensed these articles into concise summaries while retaining the essential information.

F. Limitations and Future Work

While the model performed well in most cases, there were instances where it struggled with highly technical content or ambiguous language. These limitations suggest areas for future research, such as improving the model's ability to handle domain-specific terminology or incorporating additional forms of feedback, such as expert reviews.

G. Discussion

The results indicate that the integration of RLHF significantly enhances the quality of text summarization. The model not only achieves higher ROUGE and BLEU scores but also produces summaries that are more coherent and aligned with human expectations. The hyperparameter tuning process played a pivotal role in achieving these results, ensuring that the model was both effective and efficient.

VII. CONCLUSION AND FUTURE SCOPE

In this research, we explored the potential of Reinforcement Learning with Human Feedback (RLHF) for enhancing text summarization tasks. Our findings demonstrate that RLHFnot only improves the coherence and relevance of generated summaries but also aligns them more closely with humanexpectations. This approach addresses the limitations of tradi- tional summarization methods, providing a robust framework for producing high-quality summaries in various applications, including news aggregation, legal document analysis, and scientific research.Future work could explore the integration ofRLHF with more complex NLP tasks, such as multi-documentsummarization or real-time summarization in dynamic envi- ronments. Additionally, investigating the scalability of this approach with larger and more diverse datasets could further validate its effectiveness. Expanding the scope to include multi-modal summarization, where text is combined with visual or auditory data, presents another promising avenue for research. These future directions could significantly contribute to the advancement of intelligent summarization systems.

REFERENCES

[1]. S. Ryang and T. Abekawa, "Framework of automatic text summariza- tion using reinforcement learning," in *Proceedings of the 2012 JointConference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 256–265, 2012.

[2]. T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, "Abrief overview of chatgpt: The history, status quo and potential future development," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023.

[3]. Alomari, N. Idris, A. Q. M. Sabri, and I. Alsmadi, "Deep reinforcementand transfer learning for abstractive text summarization: A review," *Computer Speech & Language*, vol. 71, p. 101276, 2022.

[4]. Y. Sun and J. Platosˇ, "Abstractive text summarization model combining a hierarchical attention mechanism and multiobjective reinforcement learning," *Expert Systems with Applications*, vol. 248, p. 123356, 2024.

[5]. K. Yao, L. Zhang, T. Luo, and Y. Wu, "Deep reinforcement learning for extractive document summarization," *Neurocomputing*, vol. 284, pp. 52– 62, 2018.

[6]. Y. K. Atri, V. Goyal, and T. Chakraborty, "Multi-document summarization using selective attention span and reinforcement learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023

[7]. B. Shukla, S. Gupta, A. K. Yadav, and D. Yadav, "Text summarizationof legal documents using reinforcement learning: A study," in *Intelligent Sustainable Systems: Proceedings of ICISS 2022*, pp. 403–414, Springer, 2022.

[8]. S. Lamsiyah, A. El Mahdaouy, A. Nourbakhsh, and C. Schommer, "Fine-tuning a large language model with reinforcement learning for educational question generation," in *International Conference on Artificial Intelligence in Education*, pp. 424–438, Springer, 2024.