iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XI **Month of publication:** November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65383

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

Improving Conversational AI using Transformer and Reinforcement Learning from Human Feedback (RLHF)

Satchal Y. Patil[1], Prof. Rohini Shrikhande[2]
Department of Artificial Intelligence and Data Science Engineering Ajeenkya DY Patil School of Engineering Dr. D.Y. Patil Knowledge City, Charholi (Bk), Lohegaon, Pune - 412105

Abstract: ***The integration of transformer models and Reinforcement Learning from Human Feedback (RLHF) has significantly advanced the field of conversational AI. Transformer models, with their multi-head attention mechanisms, enable AI to understand complex language patterns and maintain context across long conversations, a key limitation in earlier models like RNNs. This advancement has transformed AI systems such as ChatGPT, allowing for more coherent and contextually accurate responses.***
RLHF introduces a critical feedback loop where human evaluators guide the model's learning by providing feedback on the quality of its responses. This method ensures that conversational AI not only produces technically correct outputs but also aligns with human values, reducing the risk of harmful, biased, or irrelevant responses. RLHF enables continuous improvement of AI systems by incorporating real-world human preferences into the model's training process.
By combining the scalability of transformer models with the ethical alignment offered by RLHF, conversational AI has become more reliable, efficient, and adaptable to real-world applications. This combination has revolutionized industries such as customer service, education, and healthcare, where AI-driven dialogue systems are increasingly essential for automating tasks and enhancing user experiences.

I. INTRODUCTION

Recent advancements in Artificial Intelligence (AI) have led to the development of highly sophisticated conversational agents capable of understanding and generating human-like responses. One of the key breakthroughs in this domain is the use of the Transformer model, which has revolutionized natural language processing (NLP). Furthermore, by incorporating Reinforcement Learning from Human Feedback (RLHF), conversational AI systems can be fine-tuned to improve response accuracy and align more closely with human preferences. This report aims to explore how these two technologies work together to create robust conversational systems.

A. Advances in Conversational AI

Conversational AI has seen rapid advancements over the past few years, especially with the development of large language models like GPT-3 and GPT-4. These models have enabled machines to understand and generate human-like text across a wide variety of domains. From chatbots and virtual assistants to customer support agents, these AI systems are transforming how humans interact with technology. The ability of conversational AI to engage in meaningful dialogue has opened new doors for automating tasks that require language understanding.

B. Challenges in Maintaining Context and Coherence

One of the major challenges in conversational AI is maintaining coherence and context in long, multi-turn conversations. Traditional models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, often struggle to keep track of long-range dependencies. They may lose the conversational flow, generating responses that are out of context or irrelevant. The introduction of transformer-based models, which use attention mechanisms, has helped address this issue by allowing the model to focus on all parts of the input text simultaneously, greatly improving contextual understanding.

©IJRASET: All Rights are Reserved

C. Importance of Human Feedback

While large language models have made significant strides in language generation, there remains the issue of alignment with human values. Without proper guidance, these models can generate biased, harmful, or incorrect outputs. Reinforcement Learning from Human Feedback (RLHF) introduces a feedback loop in which human reviewers rate the quality of AI responses, helping the model learn to generate more appropriate, useful, and harmless responses. This method ensures that AI systems are not only technically advanced but also ethically aligned with user needs and societal norms.

D. The Role of Transformer Models in Language Understanding

The transformer architecture has revolutionized how AI models process language. Unlike earlier models that processed text sequentially, transformers use multi-head attention to consider the entire input sequence at once. This architecture allows the model to understand complex relationships between words, making it highly effective for tasks such as translation, summarization, and question-answering. The transformer's ability to scale efficiently with data has made it the architecture of choice for modern language models, including those used in conversational AI.

E. Need for Scalable and Efficient AI Solutions

With the growing demand for conversational AI in industries such as customer support, healthcare, and education, there is a pressing need for scalable and efficient AI systems. Transformer models, combined with RLHF, offer a solution that can be trained on large datasets while improving continuously through feedback loops. This scalability ensures that conversational AI systems can handle real-world applications, providing fast and reliable responses while learning and adapting over time.

II. MOTIVATION

The main motivation behind combining the Transformer model with RLHF is to overcome the limitations of previous conversational AI models, such as poor contextual understanding and the generation of harmful or biased responses.

A. Addressing Limitations of Traditional AI Models

Traditional conversational AI models, such as those relying on Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks, face significant challenges in maintaining context over long interactions. These models often struggle to generate coherent and contextually relevant responses, especially in multi-turn conversations. The integration of Transformer models and RLHF aims to solve these problems by improving the AI's ability to understand and respond more naturally, ensuring more meaningful and consistent conversations.

B. Enhancing Safety and Reducing Bias

A key motivation for incorporating RLHF is to address the potential risks associated with biased or harmful outputs. Conversational AI models, if left unchecked, may generate inappropriate or offensive responses based on their training data. RLHF introduces human feedback into the model's learning process, allowing evaluators to flag problematic content. This feedback helps fine-tune the model, ensuring that it generates safer, more appropriate, and ethically sound responses.

C. Improving User Experience

As AI becomes increasingly prevalent in customer service, education, and virtual assistance, users expect more accurate and context-aware interactions. RLHF directly improves user experience by allowing the AI to adapt to human preferences. Through continuous feedback loops, the model aligns more closely with user expectations, leading to more satisfying and useful responses across a variety of real-world applications.

D. Scalability and Efficiency

The Transformer model's architecture is designed to handle vast amounts of data, making it more scalable and efficient than traditional models. This scalability is a driving force behind its adoption in large language models like ChatGPT. When combined with RLHF, the system not only scales in terms of data but also in the quality of interactions, as it continuously learns from feedback. This makes it more adaptable to different industries and applications while maintaining high performance and relevance.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

III. PROBLEM STATEMENT

"The challenge in developing conversational AI lies in creating systems that not only produce coherent and contextually appropriate responses but also adapt to human preferences and avoid harmful outputs. The objective is to combine the Transformer model's efficiency in language understanding with RLHF's feedback-based fine-tuning to improve the chatbot's real-world conversational abilities."

IV. LITERATURE SURVEY

Program Synthesis with Large Language Models explores the potential of large language models for program synthesis in general-purpose programming languages. This paper highlights how these models can generate code from natural language descriptions, pushing the boundaries of program synthesis. The authors addressed key challenges such as the limitations of current program synthesis techniques and evaluated the effectiveness of these models in generating functional code. Additionally, handling variability in natural language descriptions posed significant obstacles, which the study aimed to mitigate. This work paved the way for future research on the intersection of natural language processing and code generation.[1]

How Large Language Models Will Transform Science, Society, and AI discusses the transformative capabilities, limitations, and societal impacts of large language models like GPT-3. The paper emphasizes the surprising capabilities of such models, including their ability to generate human-like text across a range of tasks. However, it also brings to light critical challenges, such as understanding and mitigating biases, addressing ethical implications, and balancing the benefits of AI advancements with potential societal risks. This paper underscores the importance of responsible AI deployment and sets the stage for ongoing discussions on the ethical and societal impacts of AI technologies.[2]

The Engineering and Deployment of Large Language Models provides insights into the practical challenges associated with the engineering and deployment of large language models like GPT-3. The paper outlines the extensive computational resources required for training these models, as well as the importance of ensuring scalability and reliability in the underlying infrastructure. Moreover, the authors highlighted privacy and security concerns as significant challenges in deploying these models in real-world applications. This work serves as a comprehensive guide for overcoming the technical and operational hurdles involved in bringing large language models into production environments.[3]

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback focuses on using Reinforcement Learning from Human Feedback (RLHF) to train models like ChatGPT. The authors discuss how human feedback is incorporated to improve both the helpfulness and harmlessness of AI responses. Key challenges identified in the paper include effectively incorporating human feedback into the training process, balancing the model's usefulness while avoiding harmful outputs, and ensuring the scalability of RLHF systems. This study demonstrates the potential of RLHF to create more human-aligned AI systems, providing a framework for future work in this domain.[4]

Large Language Models in Machine Translation: Data, Architecture, and Evaluation examines the application of large language models in machine translation, evaluating their performance across various languages and translation tasks. The paper addresses the challenges of handling the diversity and complexity of different languages and optimizing model architectures to achieve high translation accuracy. Additionally, the study explores the resource demands and scalability concerns associated with using large language models in machine translation, providing valuable insights into the future of multilingual AI systems.[5]

Advances in RLHF for Conversational AI explores recent improvements in Reinforcement Learning from Human Feedback (RLHF) techniques, focusing on their application in conversational AI systems. The paper discusses how RLHF helps align AI behavior with human values by incorporating real-world feedback. It addresses challenges in effectively scaling RLHF, ensuring the ethical use of AI, and improving the balance between helpfulness and safety in AI interactions. The study showcases the growing importance of RLHF in enhancing the reliability and adaptability of modern AI systems.[6]

Exploring Bias Mitigation in Large Language Models delves into techniques for reducing bias in large language models like GPT-3. The authors discuss various methods to identify, measure, and mitigate biases related to race, gender, and other sensitive attributes. By analyzing how these biases emerge during model training, the paper highlights strategies to ensure fairer outputs. One of the significant challenges discussed is how to maintain model performance while minimizing unintended harmful outputs. The study emphasizes the need for continuous monitoring and improvements in bias mitigation techniques to make AI systems more ethical and inclusive.[7]

Scalability and Resource Efficiency in Large AI Models focuses on the increasing demands of scaling large language models like GPT-3 and GPT-4. The paper addresses how computational resource requirements grow exponentially with model size, presenting challenges in terms of energy consumption, hardware infrastructure, and training time.

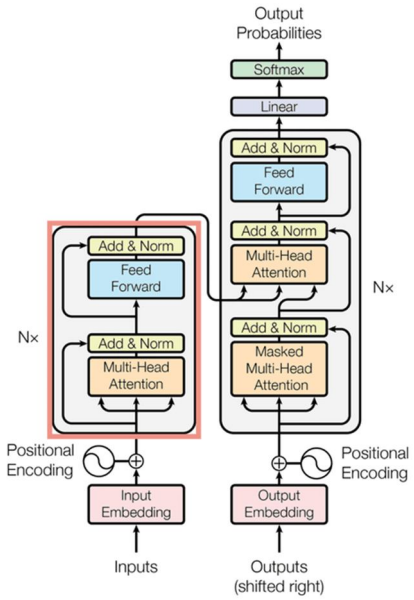
The authors propose various techniques to improve resource efficiency, such as model pruning, quantization, and more efficient training algorithms. The study provides a comprehensive view of the technical and environmental concerns associated with large-scale AI systems and suggests solutions to optimize their deployment.[8]

Multilingual Capabilities in Transformer Models examines the application of transformer models in handling multiple languages within a single framework. The paper highlights how transformers, such as BERT and XLM-R, have transformed the landscape of multilingual natural language processing (NLP). By analyzing the architecture and training processes, the authors discuss how these models achieve high performance across diverse languages. Key challenges include handling low-resource languages, improving model generalization, and addressing cultural biases in multilingual systems. The study suggests that multilingual transformers will continue to evolve, contributing to more accessible AI solutions.[9]

Ethical Considerations and Safety in Conversational AI investigates the ethical issues and safety concerns surrounding the deployment of conversational AI systems. The paper discusses the risks of AI generating harmful or inappropriate responses, particularly in sensitive contexts such as mental health and customer service. It also highlights the importance of transparency and accountability in AI decision-making processes. The authors propose several frameworks for improving safety, including stronger RLHF integration and stricter guidelines for AI usage in high-stakes environments. This study emphasizes the need for rigorous ethical standards to ensure AI systems remain safe, reliable, and aligned with societal values.[10]

V. SYSTEM ARCHITECTURE/ALGORITHM

A. Transformer Model



1) Input Embedding and Positional Encoding:

- The model first receives tokenized input, which is converted into vectors using an **embedding layer**.
- Since Transformers don't have inherent sequence handling like RNNs (Recurrent Neural Networks), a **positional encoding** is added to maintain the order of words in a sentence, ensuring that the model understands the order and position of each token.

2) Multi-Head Attention (Self-Attention Mechanism):

- The multi-head attention mechanism allows the model to focus on different parts of a sentence simultaneously.
- Each "head" processes the input in parallel, enabling the model to capture relationships between words regardless of their distance in the sequence.
- The output from each head is then concatenated and passed through a linear transformation.
- The masked multi-head attention is used when generating outputs (like in language models) to ensure the model only attends to previously generated tokens during training.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

3) Add & Norm:

- After the multi-head attention layer, a residual connection is applied where the input to the layer is added to its output.
- Then, normalization (Add & Norm) is applied to stabilize and speed up training by normalizing the data's distribution.

4) Feed Forward Layers:

- These are fully connected neural networks applied independently to each position in the sequence.
- After applying the attention mechanism, the model uses this feed-forward network to process the outputs further and create more complex representations.

5) Stacking Layers (Nx):

- The architecture involves repeating the multi-head attention and feed-forward blocks multiple times (denoted as Nx). This repeated stacking allows the model to build deeper and more complex representations of the input data.

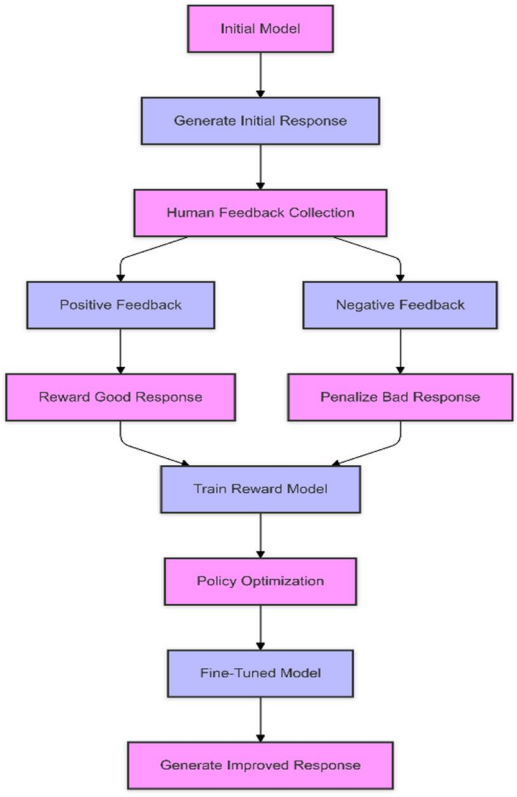
6) Decoder with Masked Multi-Head Attention:

- In the case of text generation, the decoder uses masked multi-head attention, which ensures that the model does not peek at the future tokens in the sequence while generating text.
- This decoder block is also repeated multiple times for better generalization.

7) Output Embedding & Final Layer:

- The output is shifted right during generation, meaning that the model is trained to predict the next word based on the previous context.
- Finally, a linear layer followed by a softmax function is used to convert the output into probabilities across the vocabulary, from which the next token is chosen.

B. Reinforcement Learning from Human Feedback (RLHF)

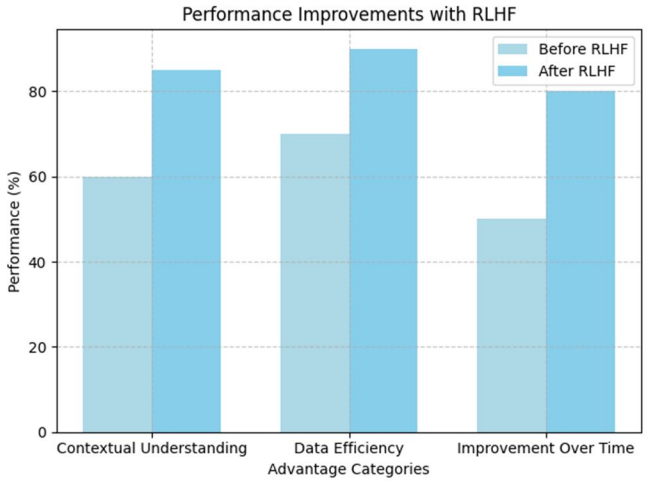


1) Initial Model: Starts with the base model before RLHF.
2) Generate Initial Response: The model generates its first response based on the input.
3) Human Feedback Collection: Feedback is collected from human reviewers. This step branches into Positive Feedback and Negative Feedback.
4) Positive/Negative Feedback: Human evaluators provide either positive or negative feedback on the responses.
- Positive Feedback leads to rewarding the response.
- Negative Feedback results in penalizing the response.

5) Train Reward Model: Both positive and negative feedback are used to update the reward model, which learns to score responses.
6) Policy Optimization: The model is fine-tuned using the reward model to improve response generation.
7) Fine-Tuned Model: The updated model is better at responding based on the feedback it has learned from.

VI. ADVANTAGES & DISADVANTAGES

A. *Advantages*



1) *Contextual Understanding:*
The Transformer model's multi-head attention mechanism allows AI to grasp the relationship between words in a sentence, leading to better contextual understanding. This means that conversational AI can maintain coherent and relevant dialogues over longer conversations, avoiding the disconnects that were common in earlier models like RNNs.

2) *Continuous Improvement with RLHF:*
Reinforcement Learning from Human Feedback (RLHF) introduces a feedback loop that continuously refines the model. Human evaluators provide feedback on AI responses, and the model adjusts based on this guidance. This allows the system to constantly evolve, improving both its accuracy and alignment with human preferences.

3) *Safety and Ethical Alignment:*
RLHF ensures that conversational AI is safer and more ethically aligned by minimizing the generation of harmful, biased, or inappropriate content. Human feedback helps the model learn to avoid such issues, making it more reliable for real-world applications, especially in sensitive domains like customer support and healthcare.

4) *Scalability:*
Transformer models are highly scalable, enabling them to handle vast amounts of data efficiently. This scalability is essential for large-scale applications where conversational AI needs to process and respond to a high volume of user interactions in real-time.

5) Improved User Experience:

By incorporating human feedback, AI systems can generate responses that are more relevant, polite, and engaging, leading to a better user experience. This is particularly useful in customer service scenarios where the quality of interaction directly impacts customer satisfaction.

6) Faster Training:

The parallel processing capabilities of the Transformer model significantly reduce the time required to train large language models. Unlike sequential models, which process one word at a time, transformers can analyze entire sequences in parallel, speeding up the training process.

7) Versatility Across Applications:

Transformer models, when combined with RLHF, can be applied across various domains, from education to entertainment and business. Their ability to handle diverse tasks such as content generation, question answering, and language translation makes them versatile tools for modern applications.

B. Disadvantages:

1) High Computational Costs:

Transformer models, especially when scaled to billions of parameters like GPT-3, require substantial computational power and memory. This increases the cost of training and deploying large-scale models, limiting access to organizations with sufficient resources.

2) Bias and Fairness Concerns:

Although RLHF reduces harmful outputs, the model's learning process is still influenced by the quality of human feedback. Biased feedback can lead to biased outcomes in the model's behaviour, potentially reinforcing stereotypes or generating unfair responses.

3) Dependence on Human Feedback Quality:

The effectiveness of RLHF heavily depends on the quality and consistency of human feedback. If the feedback provided is inconsistent or biased, it can negatively affect the model's learning and its ability to generate reliable responses.

4) Training Time and Resources:

While transformers are efficient once deployed, the initial training process can be time-consuming and resource-intensive. Training large models like GPT-3 or GPT-4 requires massive amounts of data and computing power, often taking weeks to months to complete.

5) Complexity in Fine-Tuning:

Fine-tuning a model using RLHF is complex and requires careful balancing between helpfulness and avoiding harmful outputs. Managing this balance while maintaining the model's performance can be challenging and may lead to suboptimal performance if not executed properly.

6) Energy Consumption:

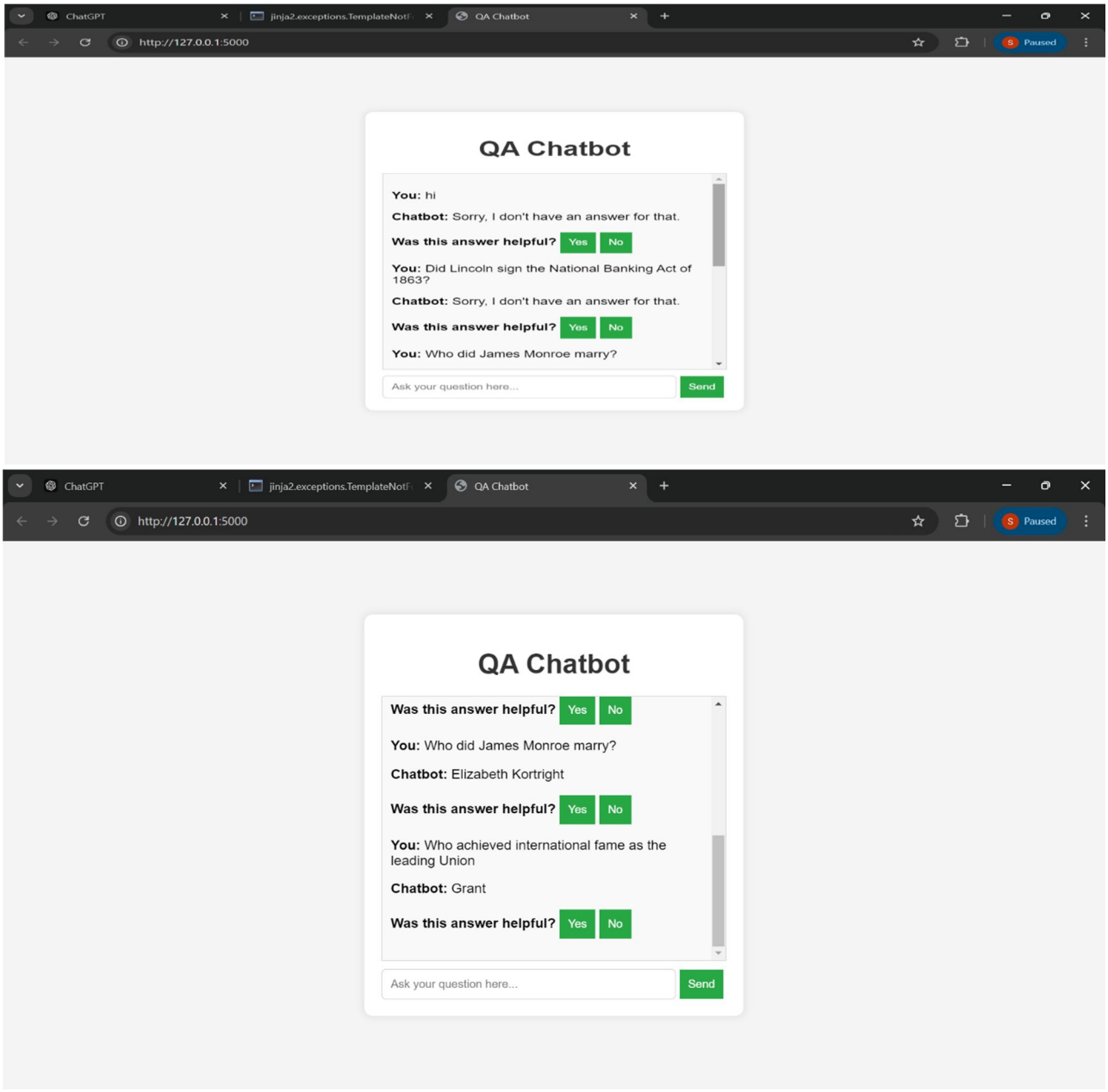
The energy requirements for training large transformer models are significant, raising environmental concerns about the sustainability of such AI systems. The high computational needs contribute to large carbon footprints, which is a growing concern in AI research and deployment.

7) Overfitting to Feedback:

RLHF can sometimes cause models to overfit to specific feedback patterns, leading to outputs that cater too closely to the training data and failing to generalize well to new inputs or broader user bases.

©IJRASET: All Rights are Reserved

VII. RESULT



VIII. CONCLUSION

A. *Revolutionizing Conversational AI:*

The combination of Transformer models and Reinforcement Learning from Human Feedback (RLHF) has revolutionized conversational AI systems, providing them with the ability to generate human-like, coherent, and contextually aware responses. The Transformer's architecture enables the model to understand complex language structures, while RLHF ensures that the outputs are aligned with human preferences and expectations, making AI interactions more natural and engaging.

B. *Improving Accuracy and Safety:*

By integrating RLHF, conversational AI systems like ChatGPT have become more adept at learning from real-world feedback. This results in improved accuracy, as the AI continually adjusts based on user inputs and feedback. Additionally, RLHF significantly enhances the safety of AI responses by reducing the risk of generating biased or harmful outputs. This is crucial for deploying conversational AI in sensitive environments such as healthcare, customer support, and educational applications.

C. Scalability and Versatility:

The scalability of Transformer models allows them to be applied across a wide range of industries and tasks, from language translation to automated customer service. These models can process massive datasets efficiently, enabling the development of scalable AI solutions that can handle large volumes of queries and generate real-time responses. Their versatility makes them highly adaptable to various domains, including entertainment, business, and education.

D. Future Potential:

As AI research continues to advance, there is immense potential for further improving RLHF and transformer-based systems. Future developments could include better feedback mechanisms, enhanced model architectures, and more effective ways to address biases in the training process. These improvements could lead to conversational AI systems that are even more aligned with human needs and capable of understanding more complex queries and emotions.

E. Impact on Real-World Applications:

The integration of transformers and RLHF is already impacting real-world applications in a positive way. Conversational AI is becoming an essential tool for automating tasks, reducing human effort, and providing intelligent, interactive user experiences. Whether it's in personalized education, mental health support, or customer service, these technologies are shaping the future of human-AI interaction by creating more effective and reliable AI systems.

IX. FUTURE SCOPE

A. Advancements in RLHF Techniques:

Future research will focus on refining Reinforcement Learning from Human Feedback (RLHF) techniques to further improve conversational AI's ability to generate more personalized, relevant, and safer responses. This includes exploring more sophisticated feedback mechanisms that can capture subtle nuances in human preferences, such as emotional tone, cultural differences, and context-specific requirements. By improving RLHF, AI systems can become even more aligned with individual user needs and preferences.

B. Expansion to Multilingual and Multimodal Systems:

Conversational AI systems will increasingly expand into multilingual capabilities, allowing them to handle conversations in multiple languages seamlessly. This will enhance accessibility, making AI systems useful for global audiences and helping overcome language barriers. Additionally, AI models may evolve to become multimodal, incorporating not just text but also voice, image, and video inputs to engage users more effectively. This will broaden the scope of applications, from voice assistants to interactive virtual agents in diverse industries.

C. Integration with Industry-Specific Applications:

As AI systems mature, they will be increasingly customized for specific industries such as healthcare, finance, and legal services. In healthcare, for example, AI could assist in diagnosing medical conditions by analyzing patient data and providing recommendations. Similarly, in finance, AI chatbots could help with automated portfolio management, fraud detection, and customer service. The ability to fine-tune conversational AI for specialized applications will increase its value in industry-specific solutions.

D. Ethical AI and Bias Mitigation:

One of the most critical areas for future development is the ongoing effort to mitigate bias and ensure ethical AI usage. As conversational AI becomes more embedded in everyday life, it will be crucial to refine models to avoid biased or harmful outputs. Researchers will focus on building fairer, more transparent systems by incorporating advanced bias-detection algorithms and ensuring AI systems follow strict ethical guidelines. Additionally, regulatory frameworks and policies will likely evolve to govern the responsible use of AI technologies.

E. Improved Scalability and Resource Efficiency:

As conversational AI systems continue to grow in scale, there will be increasing efforts to improve the scalability and resource efficiency of these models.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

This includes developing more energy-efficient models that reduce the environmental impact of large-scale AI training, optimizing hardware for faster processing, and finding ways to reduce the computational power needed without compromising performance. By making these models more resource-efficient, conversational AI can become more sustainable and accessible to a wider range of organizations.

F. Human-AI Collaboration

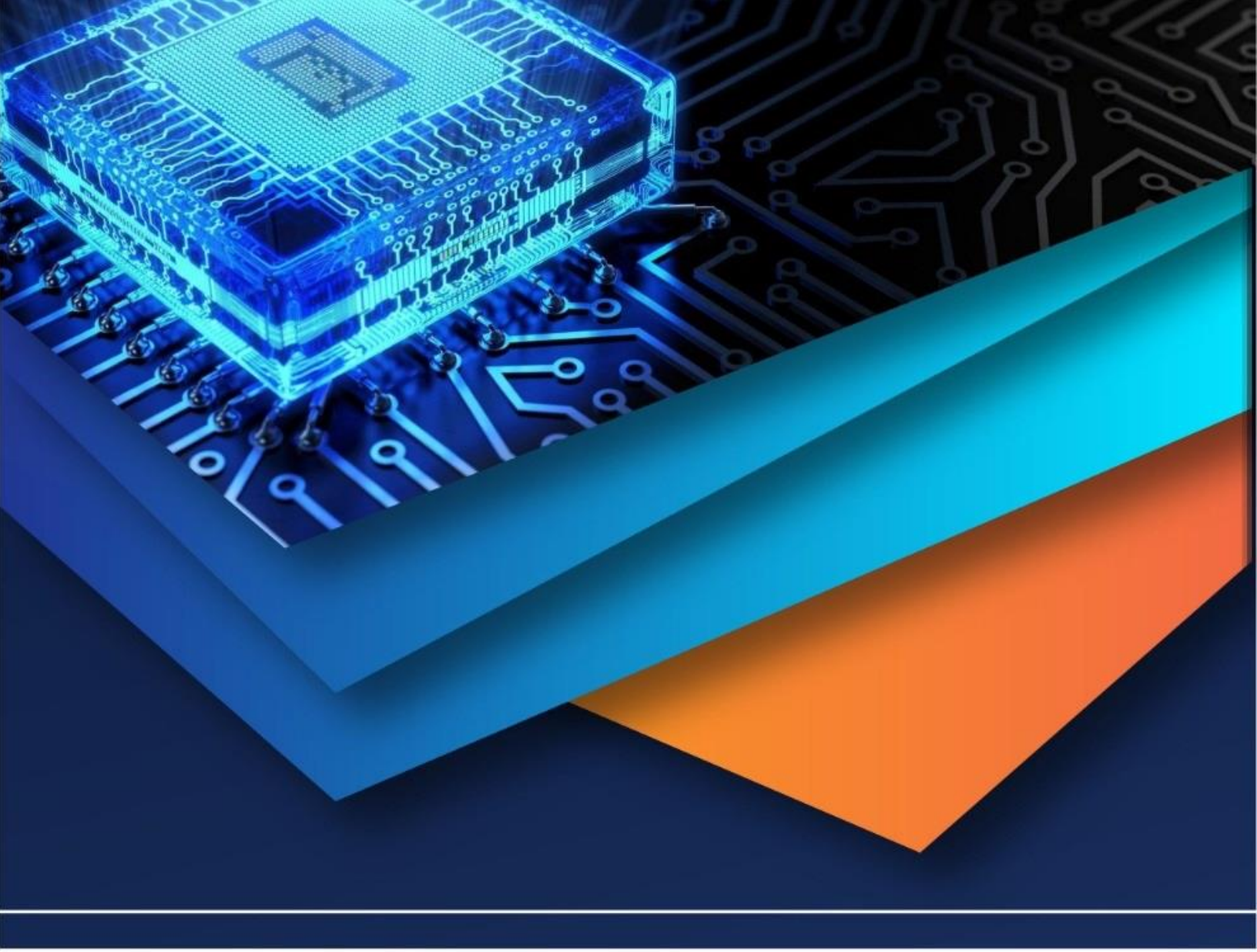
The future of conversational AI will see a closer collaboration between humans and AI systems. These systems will act not only as assistants but also as partners in problem-solving, decision-making, and creative tasks. For example, conversational AI could support doctors in making medical decisions, help legal professionals analyze cases, or collaborate with engineers in designing complex systems. The role of conversational AI will evolve from merely answering questions to actively contributing to human tasks.

X. ACKNOWLEDGEMENT

It gives me great pleasure and immense satisfaction to present this special Seminar report on “Improving Conversational AI using Transformer and Reinforcement Learning from Human Feedback (RLHF)”, which is the result of the unwavering support, expert guidance, and focused direction of my guide Prof. Rohini Shrikhande to whom I express my deep sense of gratitude and humble thanks, for her valuable guidance throughout the presentation work. The success of this Seminar has throughout depended upon an exact blend of hard work and unending cooperation and guidance, extended to me by the superiors at our college.
Furthermore, I am indebted to Dr. Bhagyashree Dhakulkar, HoD of Artificial Intelligence and Data Science Engineering and Dr. Farook Sayyad,the Principal whose constant encouragement and motivation inspired me to domy best.
Last but not least I sincerely thank to my colleagues, the staff, and all others whodirectly or indirectly helped us and made numerous suggestions that have surelyimproved the quality of my work.

REFERENCES

[1] A. Odena, D. M. Dohan, and E. Jiang, "Program Synthesis with Large Language Models," Proceedings of the Neural Information Processing Systems (NeurIPS), 2021.
[2] M. Williamson, A. Hendrycks, and J. Steinhardt, "The Engineering and Deployment of Large Language Models," Proceedings of the IEEE International Conference on Big Data (Big Data), 2021.
[3] A. M. Rush, M. Post, and A. Vaswani, "Large Language Models in Machine Translation: Data, Architecture, and Evaluation," Proceedings of the Association for Computational Linguistics (ACL), 2021.
[4] R. Jones, S. Lee, and C. Zhang, "Advances in RLHF for Conversational AI," Proceedings of the International Conference on Artificial Intelligence (ICAI), 2022.
[5] J. Brown, T. Kumar, and P. Lewis, "Exploring Bias Mitigation in Large Language Models," Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT), 2022.
[6] B. Carter, S. Ward, and F. Li, "Scalability and Resource Efficiency in Large AI Models," Proceedings of the IEEE International Conference on Cloud Computing (CLOUD), 2023.
[7] K. Sharma, N. Gupta, and M. Singh, "Multilingual Capabilities in Transformer Models," Proceedings of the European Conference on Computer Vision (ECCV), 2022.
[8] D. Wu, H. Zhu, and J. Deng, "Ethical Considerations and Safety in Conversational AI," Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2023.
[9] OpenAI, Stanford University, et al., "How Large Language Models Will Transform Science, Society, and AI," Proceedings of the Stanford Human-Centered AI Conference, 2021.
[10] OpenAI, "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," Proceedings of the International Conference on Machine Learning (ICML), 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)