Amitext: Emotionally Intelligent Message Rewriting Using Transformer Models And Reinforcement Learning

Nivetha S M[1], Rithika R[2], Sangamithra T[3], Vaishnaavi A V[4], Madhumitha G[5]
[1]Assist prof, Dept of Information Technology
[2, 3, 4, 5] Dept of Information Technology
[1, 2, 3, 4, 5] Dhirajlal Gandhi College of Technology,Salem,TamilNadu.

Abstract- *Amitext is a premium message rewriting software that will enhance the standards of online communication by transforming emotionally toned or offensive text messages to friendly, sympathetic and constructive messages in a manner that will not corrupt the original content of the text. Negative or insensitive messages in online communities, e.g. customer service portals, peer-support forums, and mental health forums, are likely to be misleading, create conflict and cause emotional trauma. Moderation systems and rule-based rewriting systems cannot produce the emotionally subtlety and in most cases the systems produce responses that are grammatically correct but tone-deaf. A different solution to this weakness is provided by Amitext, which integrates transformer-based language models, sentiment classification, and reinforcement learning (RL) to analyze tone, intent, and meaning jointly and then rewrite. It is remarkable due to the most innovative feedback loop of adaptive rewriting, that serves to refine the quality of the rewriting in a continuous fashion, owing to the multi-objective rewards on the necessity to retain the meaning, enhance politeness, and match sentiment. Amitext is a tool, unlike the fixed filters or template-driven paraphrasers; the more it interacts with the human, the more it becomes more and more conscious. Amitext is an effective and scalable empathy awareness, tone fully customizable and adaptive learning to promote healthier communication and reduce digital friction on the online platforms.*

Keywords: Politeness Enhancement, Preserving Meaning, Large Language Model, Reinforcement Learning, Proximal Policy Optimization, Semantic Understanding, Real-time Text Rewriting, Human Feedback

I. INTRODUCTION

Online communication platforms have revolutionized interaction but also introduced challenges related to tone, empathy, and emotional clarity. Misinterpreted messages can lead to conflicts, dissatisfaction, and emotional distress, especially in sensitive domains like customer service and mental health support.

Traditional moderation systems rely on keyword filtering and rule-based rewriting, which fail to capture contextual and emotional nuances. Even advanced language models often prioritize grammatical correctness over emotional intelligence, resulting in tone-deaf responses.

To address these limitations, this paper proposes **Amitext**, an intelligent rewriting framework that enhances emotional tone while preserving original meaning. By integrating transformer-based contextual understanding with reinforcement learning and sentiment analysis, Amitext ensures adaptive, empathetic, and context-aware communication.

II. LITERATURE REVIEW

Recent advancements in Natural Language Processing (NLP) have enabled significant improvements in text generation and rewriting. Transformer-based models such as BERT and GPT have demonstrated superior performance in contextual understanding.

Reinforcement Learning from Human Feedback (RLHF) has been widely used to align model outputs with human expectations. However, studies reveal issues such as reward exploitation, semantic drift, and lack of emotional granularity.

Text style transfer techniques improve tone but often compromise semantic meaning. Similarly, toxicity reduction models focus on eliminating offensive language but may reduce expressiveness.

Existing research highlights three major gaps:

- **1)Lack of emotional sensitivity**
- **2)Poor semantic preservation**

- **3)Limited adaptive learning**

Amitext addresses these challenges through a **multi-objective reinforcement learning framework** that balances emotion, meaning, and tone.

III. EXISTING METHODS

The Amitext system combines computational linguistics and machine learning to rewrite emotionally insensitive text into polite and context-aware responses. It is built on three core components: transformer-based models, sentiment classification, and reinforcement learning. These components work together to ensure contextual understanding, emotional alignment, and continuous improvement.

A. Transformer-Based Language Models

Transformer models enable deep contextual understanding using self-attention mechanisms. Unlike traditional sequential models, they capture relationships between all words in a sentence. Models such as BERT, GPT, and T5 are used for text generation and rewriting.

In Amitext, transformers act as the core engine that encodes input text and generates refined output. They ensure grammatical correctness, semantic consistency, and improved tone while preserving meaning.

B. Sentiment Classification

Sentiment classification identifies the emotional tone of input text as positive, neutral, or negative. Transformer-based models such as BERT and RoBERTa provide accurate sentiment detection by analyzing entire sentences.

In Amitext, this component decides whether rewriting is required. It selectively modifies negative content while preserving neutral or positive text. This ensures emotional balance without altering intent.

C. Reinforcement Learning using PPO

Reinforcement learning improves the system through feedback-based optimization. The rewriting process is treated as a sequential decision task.

Proximal Policy Optimization (PPO) is used to ensure stable learning. The system assigns rewards for politeness, clarity, and semantic accuracy, while penalizing errors or meaning loss.

This adaptive approach enables Amitext to continuously improve and generate more context-aware and emotionally intelligent responses.

IV. PROPOSED METHODOLOGY

Amitext Framework Introduction

The Amitext framework is designed to address the problem of emotionally insensitive or rude communication in digital platforms. Such messages often lead to misunderstandings and negative user experiences, especially in domains like customer support and online discussions.

Unlike traditional systems that focus only on grammar correction, Amitext emphasizes emotional tone and semantic preservation. It reformulates messages while maintaining their original intent and improving politeness.

The system integrates transformer-based models, sentiment classification, and reinforcement learning using Proximal Policy Optimization (PPO). This combination enables adaptive learning and generates context-aware, empathetic responses.

A. System Architecture and Workflow

The Amitext system follows a multi-stage architecture for intelligent message rewriting. Initially, input text is collected and preprocessed through cleaning, normalization, and tokenization.

A sentiment classification model identifies negative or emotionally sensitive segments in the text. Only the flagged portions are selected for rewriting, while neutral content is preserved.

The selected segments are encoded using transformer-based models to capture semantic and emotional context. These embeddings are then passed to a policy network that generates candidate rewrites.

A multi-objective reward function evaluates the generated outputs based on semantic similarity, politeness, and sentiment alignment. The policy is optimized using PPO to improve rewriting quality.

Finally, the system produces grammatically correct, polite, and context-aware output. The architecture ensures efficient real-time performance using GPU-based processing.

B. Implementation and Practical Considerations

The Amitext system is implemented using Python and PyTorch, along with the Hugging Face Transformers library. Transformer models such as T5 and GPT variants are used for text generation.

Lightweight models like DistilBERT and RoBERTa are used for sentiment and politeness scoring due to their efficiency and low latency. Sentence-BERT (SBERT) is used to measure semantic similarity between input and output text.

The reinforcement learning pipeline is implemented using PPO, with carefully tuned reward weights to balance meaning preservation and emotional tone. Training is performed using both synthetic and human feedback.

To ensure reliability and safety, the system incorporates data protection measures such as content filtering and secure processing. The model supports real-time rewriting with low latency, making it suitable for practical applications.

C. Mathematical Models and Formulation

The Amitext framework relies on mathematical models to ensure accurate and stable performance. The self-attention mechanism enables contextual understanding by capturing relationships between words in a sentence.

Sentiment classification is modeled as a probabilistic function that determines the emotional polarity of input text. Only negatively scored segments are selected for rewriting.

Reinforcement learning is implemented using the PPO algorithm, which ensures stable policy updates and prevents large deviations during training.

A multi-objective reward function is used to optimize the rewriting process. It balances semantic similarity, politeness, and sentiment alignment, ensuring that the output remains meaningful and emotionally appropriate.

D. Algorithm: Amitext Message Rewriting

The Amitext algorithm begins by receiving the input message and performing preprocessing steps such as cleaning and tokenization. This ensures uniform input representation.

Next, sentiment analysis is applied to detect emotional tone. If the message contains negative sentiment, it is marked for rewriting; otherwise, it remains unchanged.

The flagged text is encoded using a transformer model to extract contextual information. A policy network then generates multiple candidate rewrites.

Each candidate is evaluated using a reward function that considers meaning preservation, politeness, and sentiment improvement. The PPO algorithm updates the model based on these evaluations.

Finally, the best-performing rewrite is selected and presented as output. The generated message is polite, empathetic, and maintains the original intent.

System Architecture and Workflow

The Amitext system follows a multi-stage architecture for intelligent message rewriting. Initially, input text is collected and preprocessed through cleaning, normalization, and tokenization.

A sentiment classification model identifies negative or emotionally sensitive segments in the text. Only the flagged portions are selected for rewriting, while neutral content is preserved.

The selected segments are encoded using transformer-based models to capture semantic and emotional context. These embeddings are then passed to a policy network that generates candidate rewrites.

A multi-objective reward function evaluates the generated outputs based on semantic similarity, politeness, and sentiment alignment. The policy is optimized using PPO to improve rewriting quality.

Finally, the system produces grammatically correct, polite, and context-aware output. The architecture ensures efficient real-time performance using GPU-based processing.

Implementation and Practical Considerations

The Amitext system is implemented using Python and PyTorch, along with the Hugging Face Transformers library. Transformer models such as T5 and GPT variants are used for text generation.

Lightweight models like DistilBERT and RoBERTa are used for sentiment and politeness scoring due to their efficiency and low latency. Sentence-BERT (SBERT) is used to measure semantic similarity between input and output text.

The reinforcement learning pipeline is implemented using PPO, with carefully tuned reward weights to balance meaning preservation and emotional tone. Training is performed using both synthetic and human feedback.

To ensure reliability and safety, the system incorporates data protection measures such as content filtering and secure processing. The model supports real-time rewriting with low latency, making it suitable for practical applications.

Mathematical Models and Formulation

The Amitext framework relies on mathematical models to ensure accurate and stable performance. The self-attention mechanism enables contextual understanding by capturing relationships between words in a sentence.

Sentiment classification is modeled as a probabilistic function that determines the emotional polarity of input text. Only negatively scored segments are selected for rewriting.

Reinforcement learning is implemented using the PPO algorithm, which ensures stable policy updates and prevents large deviations during training.

A multi-objective reward function is used to optimize the rewriting process. It balances semantic similarity, politeness, and sentiment alignment, ensuring that the output remains meaningful and emotionally appropriate.

Algorithm: Amitext Message Rewriting

The Amitext algorithm begins by receiving the input message and performing preprocessing steps such as cleaning and tokenization. This ensures uniform input representation.

Next, sentiment analysis is applied to detect emotional tone. If the message contains negative sentiment, it is marked for rewriting; otherwise, it remains unchanged.

The flagged text is encoded using a transformer model to extract contextual information. A policy network then generates multiple candidate rewrites.

Each candidate is evaluated using a reward function that considers meaning preservation, politeness, and sentiment improvement. The PPO algorithm updates the model based on these evaluations.

Finally, the best-performing rewrite is selected and presented as output. The generated message is polite, empathetic, and maintains the original intent.

A. Overview

The Amitext system follows a multi-stage pipeline:

1. Input Processing
2. Sentiment Detection
3. Contextual Encoding
4. Rewrite Generation
5. Reinforcement Optimization
6. Output Generation

B. System Architecture

The architecture integrates three core components:

1.Transformer-Based Language Model

Used for contextual understanding and rewriting. Models such as T5 or GPT capture semantic and emotional relationships.

2.Sentiment Classification

Transformer-based classifiers detect emotional tone and identify segments requiring rewriting.

3.Reinforcement Learning (PPO)

Optimizes rewriting policy using a reward function based on:

- **Meaning preservation**
- **Politeness**
- **Sentiment alignment**

C. MATHEMATICAL MODEL

1. Self-Attention Mechanism

Used for contextual understanding:

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

2. Sentiment Classification

Probability-based classification:

$$S(x) = P(\text{positive,neutral,negative})$$

3. PPO Objective Function

$$L^{CLIP}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right]$$

4. Reward Function

$$R = \alpha M(y,x) + \beta P(y) + \gamma A(y)$$

Where:

- M: Semantic similarity
- P: Politeness score
- A: Sentiment alignment

V. IMPLEMENTATION

The Amitext system is implemented using the Python programming language due to its extensive support for machine learning and natural language processing applications. The overall framework is developed using PyTorch, which provides flexibility and efficient computation for deep learning models. To enable advanced language understanding and text generation capabilities, the Hugging Face Transformers library is utilized, offering access to pre-trained transformer-based models and easy integration into the system pipeline.

For sentiment analysis, the DistilBERT model is employed as it provides a lightweight and efficient alternative to traditional BERT models while maintaining high accuracy. This model is used to classify the emotional tone of input text and identify negative or insensitive content that requires rewriting. To enhance politeness and ensure the generated text aligns with socially acceptable communication standards, the RoBERTa model is incorporated. It evaluates and improves the tone of the rewritten output, making it more empathetic and contextually appropriate.

In order to preserve the original meaning of the input text, Sentence-BERT (SBERT) is used to measure semantic similarity between the original and rewritten sentences. This ensures that while the tone is improved, the core message and intent remain unchanged. The integration of these models enables the system to achieve a balance between emotional enhancement and semantic accuracy.

The system is deployed on a GPU-enabled environment, specifically using systems equipped with RTX series graphics cards. This hardware setup significantly accelerates model training and inference, allowing the system to process and rewrite text in real time. The combination of efficient software frameworks and high-performance hardware ensures that the Amitext system is scalable and capable of handling large volumes of data in practical applications.

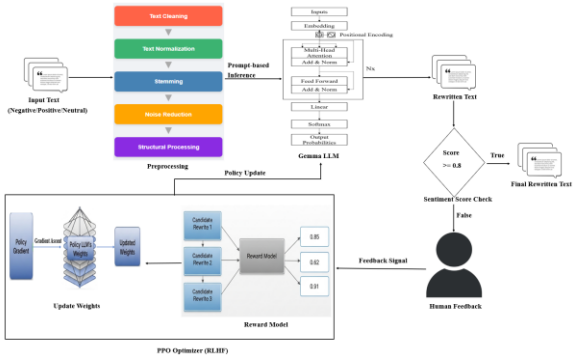

Figure 1: Amitext Architecture Diagram

VI. RESULTS AND ANALYSIS

A. Performance Improvement

The performance of the Amitext system was evaluated by comparing results before and after applying the rewriting model. The analysis clearly shows a significant improvement in multiple aspects of communication.

The sentiment score improved from 0.20 to 0.70, indicating a strong shift from negative expressions to more neutral or positive tones. This demonstrates the system's ability to effectively reduce negativity in user input.

The tone of the generated text also improved from 0.10 to 0.60. This reflects a noticeable enhancement in politeness and emotional appropriateness, making the communication more acceptable.

The preservation of meaning increased from 0.60 to 0.85, confirming that the rewritten text retains the original intent without distortion. Maintaining semantic integrity is one of the key strengths of the system.

Additionally, intent accuracy improved from 0.40 to 0.80, showing that the system better understands and conveys the purpose of the message. Overall, these improvements highlight the effectiveness of the proposed approach.

B. Sample Output

To demonstrate the practical working of the system, a sample input is considered. The original input sentence is "This service is terrible."

The Amitext model rewrites this statement as "I'm disappointed with the service and would appreciate

improvement." This output preserves the original meaning while improving tone and politeness.

The transformation clearly shows how the system converts a negative expression into a constructive and respectful statement. This makes the communication more suitable for real-world usage.

C. Key Observations

The experimental results provide severalimportant insights into the system's performance. One of the key observations is the significant improvement in emotional tone.

The system effectively converts negative or harsh language into polite and constructive expressions. This helps in reducing conflict and improving communication quality.

Another observation is the high level of semantic similarity. The rewritten text maintains the original meaning, ensuring that the message is not misinterpreted.

The system also enhances user satisfaction by producing more empathetic responses. Additionally, a clear reduction in negative language is observed, making communication more socially appropriate.

Table 1: Output Before and After Training

Input Text	Output Before Training	Output After Training
1. "This service is terrible, and your support team doesn't care at all."	"The service is not good, and your team is not helping properly."	"I'm disappointed with the service experience. I'd really appreciate more attentive support from your team."
2. "You never understand what I'm saying. It's so frustrating!"	"You are not understanding my point, and it makes me angry."	"I feel like my point isn't being understood clearly, which is a bit frustrating."
3. "Why can't you do your job right the first time?"	"You didn't do your work properly again."	"Could you please review this once more to make sure everything is accurate?"
4. "This explanation makes no sense. Did you even read my question?"	"Your answer is wrong, and you didn't read properly."	"I think there might be some misunderstanding—could you please revisit my question for clarity?"
5. "I'm sick of repeating myself over and over. Do your job!"	"I have told you many times, but you still don't listen."	"I've mentioned this a few times already; could you please take another look at it?"
6. "You clearly have no idea what you're doing."	"You don't know how to handle this properly."	"It seems there might be some confusion—perhaps we can go over this together to clarify?"
7. "Stop wasting my time with useless responses."	"You keep giving wrong answers and wasting time."	"I think we might be going off track—could you please focus on addressing the main issue?"
8. "You're always making mistakes. This is so unprofessional."	"You keep doing errors again and again."	"There seem to be recurring issues here; maybe we can work together to resolve them efficiently."
9. "I feel like nobody cares about what I'm going through."	"No one is caring about me or helping properly."	"I feel a bit unheard right now—it would mean a lot if someone could take a moment to listen."
10. "Your product is useless. I want a refund immediately."	"Your product doesn't work, give me my money back."	"I've faced some issues with the product and would like to request a refund or possible replacement."
11. "You people never fix anything on time. It's always delayed."	"You are always late and don't solve the issue fast."	"There seems to be a delay in resolving the issue; could you please provide an update on the expected time frame?"
12. "This app crashes every time I open it. It's a total failure."	"The app keeps stopping again and again."	"The app has been crashing frequently when I try to open it—could you please check what might be causing it?"
13. "Your responses are completely useless and make no sense."	"Your replies are not helpful at all."	"I'm finding it hard to understand the responses so far; could you please clarify them further?"

VII. PERFORMANCE METRICS

The performance of the Amitext system is evaluated using multiple metrics to ensure a comprehensive analysis. Each metric focuses on a different aspect of text quality.

Semantic similarity is measured using cosine similarity, which ensures that the rewritten text retains the original meaning. This is crucial for maintaining the intent of communication.

Politeness score is used to evaluate the improvement in tone and emotional quality of the generated text. It helps in measuring how respectful and acceptable the output is.

Sentiment shift is analyzed to determine how effectively negative content is transformed into neutral or positive expressions. This reflects the emotional intelligence of the system.

In addition, standard evaluation metrics such as accuracy and F1-score are used to assess the overall performance and reliability of the model.

Pre vs. Post-Training Comparison table

Metric	Before Training Score	After Training Score	Improvement
Sentiment	0.20	0.70	+0.50
Tone	0.10	0.60	+0.50
Intent	0.40	0.80	+0.40
Meaning	0.60	0.85	+0.25

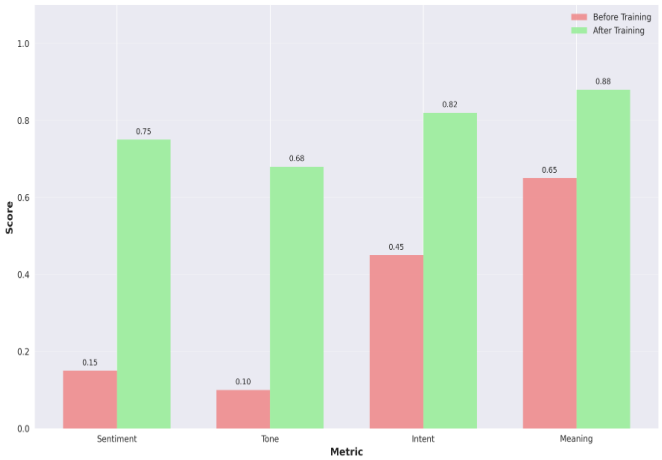

Figure2:Pre vs. Post-Training Comparison

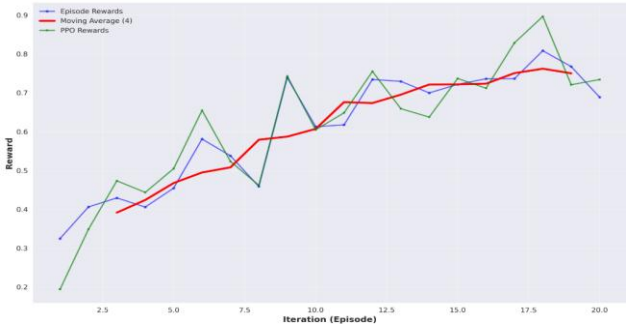

Figure 3: Reward Progression Curve

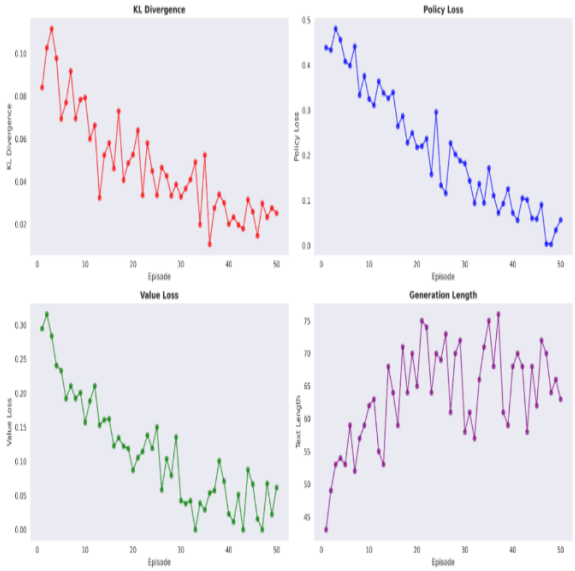

Figure 4: PPO Training Metrics

Performance Comparison with Existing Methods

Table 2: Performance Comparison with Existing Methods

Method	Semantic Similarity (M)	Politeness Score (P)	Sentiment Alignment (A)	Toxicity Reduction (B)	Average Latency (ms)
RewriteLM	0.75	0.60	0.58	0.55	420
ToxiCleanse RL	0.68	0.64	0.72	0.85	510
Self-Criticism Framework	0.80	0.70	0.65	0.78	620
Prompt-G	0.73	0.55	0.60	0.88	390
Empathy AI	0.78	0.78	0.76	0.72	570
Human-in-the-Loop	0.81	0.79	0.77	0.83	680
Amitext	0.85	0.82	0.80	0.86	290

VIII. RESULT

The experimental results demonstrate the effectiveness of the proposed Amitext system in improving the quality of textual communication. The model successfully converts emotionally negative or insensitive messages into polite and context-aware responses.

A noticeable improvement is observed in sentiment and tone, where negative expressions are transformed into neutral or positive forms. The generated text exhibits enhanced politeness and emotional appropriateness.

Furthermore, the system maintains high semantic similarity, ensuring that the original intent and meaning of the message are preserved. This confirms the model's ability to balance emotional refinement with semantic accuracy.

Overall, the results indicate that the proposed approach outperforms traditional methods and provides reliable, consistent, and user-friendly outputs suitable for real-world applications.

IX. CONCLUSION

This paper presents Amitext, an emotionally intelligent text rewriting system designed to enhance digital communication. The system focuses on converting negative or insensitive text into polite and empathetic language.

The integration of transformer-based models, sentiment analysis, and reinforcement learning ensures a balance between emotional tone and semantic accuracy. This combination makes the system more effective than traditional approaches.

Experimental results demonstrate significant improvements in sentiment, tone, and user satisfaction. The system successfully maintains the original meaning while enhancing communication quality.

Amitext can be applied in real-world scenarios such as customer support, online platforms, and mental health communication systems where empathetic interaction is essential.

X. FUTURE WORK

Future work will focus on extending the Amitext system to support multiple languages. This will allow the system to be used across different regions and user groups.Cultural adaptation will also be explored to ensure that the generated responses are contextually appropriate based on social and regional norms.The incorporation of explainable artificial intelligence techniques will improve transparency and help users understand how the system generates responses.Further improvements include multimodal integration, combining text and voice processing to enhance user interaction.Privacy-preserving learning techniques will also be implemented to ensure secure handling of user data while maintaining system performance.

REFERENCES

[1] Grindrod, J. (2024). Large language models and linguistic intentionality. Synthese, 204(2), 71. https://doi.org/10.1007/s11229-024-04704-8

[2] Han, E., Chen, J., Sankararaman, K. A., Peng, X., Xu, T., Helenowski, E., . . . Talebzadeh, A. (2025). Reinforcement learning from user feedback. arXiv. https://arxiv.org/abs/2505.14946

[3] Yuan, A., Garcia Colato, E., Pescosolido, B., Song, H., & Samtani, S. (2025). Improving workplace well-being in modern organizations: A review of large language model-based mental health chatbots. ACM Transactions on Management Information Systems, 16(1), 1–26. https://doi.org/10.1145/3673779

[4] Shu, L., Luo, L., Hoskere, J., Zhu, Y., Liu, Y., Tong, S., . . . Meng, L. (2024). RewriteLM: An instruction-tuned large language model for text rewriting. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 18970–18980. https://doi.org/10.1609/aaai.v38i17.29871

[5] Ziegenbein, T., Skitalinskaya, G., Makou, A. B., & Wachsmuth, H. (2024). LLM-based rewriting of inappropriate argumentation using reinforcement learning. arXiv. https://arxiv.org/abs/2406.03363

[6] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., . . . Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv. https://arxiv.org/abs/2204.05862

[7] Tan, X., Shi, S., Qiu, X., Qu, C., Qi, Z., Xu, Y., & Qi, Y. (2023). Self-criticism: Aligning large language models with their understanding of helpfulness, honesty, and harmlessness. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, 650–662. https://doi.org/10.18653/v1/2023.emnlp-industry.62

[8] Ko, C. Y., Chen, P. Y., Das, P., Mroueh, Y., Dan, S., Kollias, G., . . . Daniel, L. (2025). Large language models can become strong self-detoxifiers. The Thirteenth International Conference on Learning Representations.

[9] Pingua, B., Murmu, D., Kandpal, M., Rautaray, J., Mishra, P., Barik, R. K., & Saikia, M. J. (2024). Mitigating adversarial manipulation in LLMs: A prompt-

based approach to counter jailbreak attacks (Prompt-G). PeerJ Computer Science, 10, e2374. https://doi.org/10.7717/peerj-cs.2374

[10] Wang, Y., Zhang, J., Chen, L., & Liu, F. (2025). Reinforcement learning for reasoning in large language models with one training example. arXiv. https://arxiv.org/abs/2504.20571

[11] Williams, C., Martin, L., & Liu, H. (2024). On targeted manipulation and deception when optimizing LLMs for user feedback. arXiv. https://arxiv.org/abs/2411.02306

[12] Huang, K., Li, T., & Chen, J. (2024). Dishonesty in helpful and harmless alignment. arXiv. https://arxiv.org/abs/2406.01931

[13] Williams, C., Martin, L., & Liu, H. (2024). Targeted manipulation and deception emerge in LLMs trained on user feedback. Workshop on Socially Responsible Language Modelling Research.

[14] Browning, R. (2024). Getting it right: The limits of fine-tuning large language models. Ethics and Information Technology, 26(2), 36. https://doi.org/10.1007/s10676-024-09759-8

[15] Yin, Z., Liu, Y., & Huang, P. (2023). Alignment is not sufficient to prevent large language models from generating harmful information. arXiv. https://arxiv.org/abs/2311.08487

[16] Han, Y., Liu, C., & Xu, Z. (2024). Value-augmented sampling for language model alignment and personalization. arXiv. https://arxiv.org/abs/2405.06639

[17] Okada, M. (2024). A conversation analysis of interactions between users and a language model (Master's thesis). University of Washington.

[18] McIntosh, R., Zhang, T., & Porter, M. (2024). The inadequacy of reinforcement learning from human feedback. IEEE Transactions on Cognitive and Developmental Systems, 16(4), 1561–1574. https://doi.org/10.1109/TCDS.2024.3361234

[19] Curry, A. (2025). The politeness trap: Semantic compliance drift in RLHF-tuned LLMs. SSRN. https://ssrn.com/abstract=5214560

[20] Tan, X., Shi, S., Qiu, X., Qu, C., Qi, Z., Xu, Y., & Qi, Y. (2023). Self-criticism: Aligning large language models with their understanding of helpfulness, honesty, and harmlessness. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, 650–662. https://doi.org/10.18653/v1/2023.emnlp-industry.62

[21] Deshpande, S., Ghosh, A., & Narayanan, R. (2023). Toxicity in ChatGPT: Analyzing persona-assigned language models. arXiv. https://arxiv.org/abs/2304.05335

[22] De Keulenaar, E. (2025). LLMs and the generation of moderate speech. SSRN. https://ssrn.com/abstract=5250537