ENHANCING HUMAN-AI COLLABORATION IN DOCUMENT CLASSIFICATION THROUGH RLHF AND RLAIF: A STUDY ON ADAPTIVE INTERACTION MODELS

Ratnesh Kumar Sharma, Prof. (Dr) Satya Singh
Department of Computer Science & Applications, M.G. Kashi Vidyapith, Varanasi (U.P.)

***Abstract*- In recent times, Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF) have emerged as potent frameworks to improve human-AI collaboration, especially in intricate tasks like document classification in healthcare and other critical fields. This paper examines a hybrid classification framework that combines RLHF and RLAIF inside a multi-model ensemble comprising Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), and Naive Bayes (NB), with the objective of developing an adaptive and intelligent human-AI interaction model. The aim is to create a resilient document classification system that achieves high accuracy while conforming to human reasoning and expectations. The methodology entails training each classifier on pre-processed, human-labelled documents, utilising reinforcement learning to enhance model performance through iterative feedback from both humans and AI, and subsequently amalgamating the results via ensemble techniques. The assessment of a test set including 300 documents across 11 unique categories resulted in flawless performance metrics—100% precision, recall, and F1-score—indicating the absence of misclassifications. Nonetheless, critical reflection reveals a concern over potential overfitting attributed to the consistency of elevated scores, necessitating the incorporation of different and more complex validation datasets. The study implies that models with somewhat below perfect accuracy (e.g., 99%) are often more pragmatic and generalizable in real-world settings, especially in sensitive domains like healthcare. This study shows how RLHF and RLAIF can be used to create adaptive AI systems that improve document classification skills. Performance, interpretability, and generalisation must be balanced.**

***Index Terms*- Human – AI Collaboration; Document Classification; Adaptive interaction models; RLHF; RLAIF**

I. INTRODUCTION

The integration of artificial intelligence (AI) into human decision-making has revolutionised data-driven systems, particularly in industries such as healthcare, finance, and legal document management. The transformation is driven by emerging paradigms like RLHF and RLAIF, which aim to enhance model adaptability by aligning machine learning results with human judgement and expectations (Bradley et al., 2023; Benaich & Hogarth, 2020). RLHF relies on continuous input from human evaluators to guide learning, but RLAIF integrates autonomous AI feedback systems to improve decision-making without the necessity of persistent human participation. This dual methodology enables the creation of AI systems that are accurate, contextually aware, and transparent (Li, 2023; Daull et al., 2023). In document classification, where understanding nuanced context, maintaining precision, and ensuring ethical compliance are essential, the synergy between human and machine contributions is particularly advantageous (Jeong et al., 2023; Bradley et al., 2023; Zhu et al., 2023). The following figure 1 illustrates the overview of reinforcement learning from AI feedback in detail.

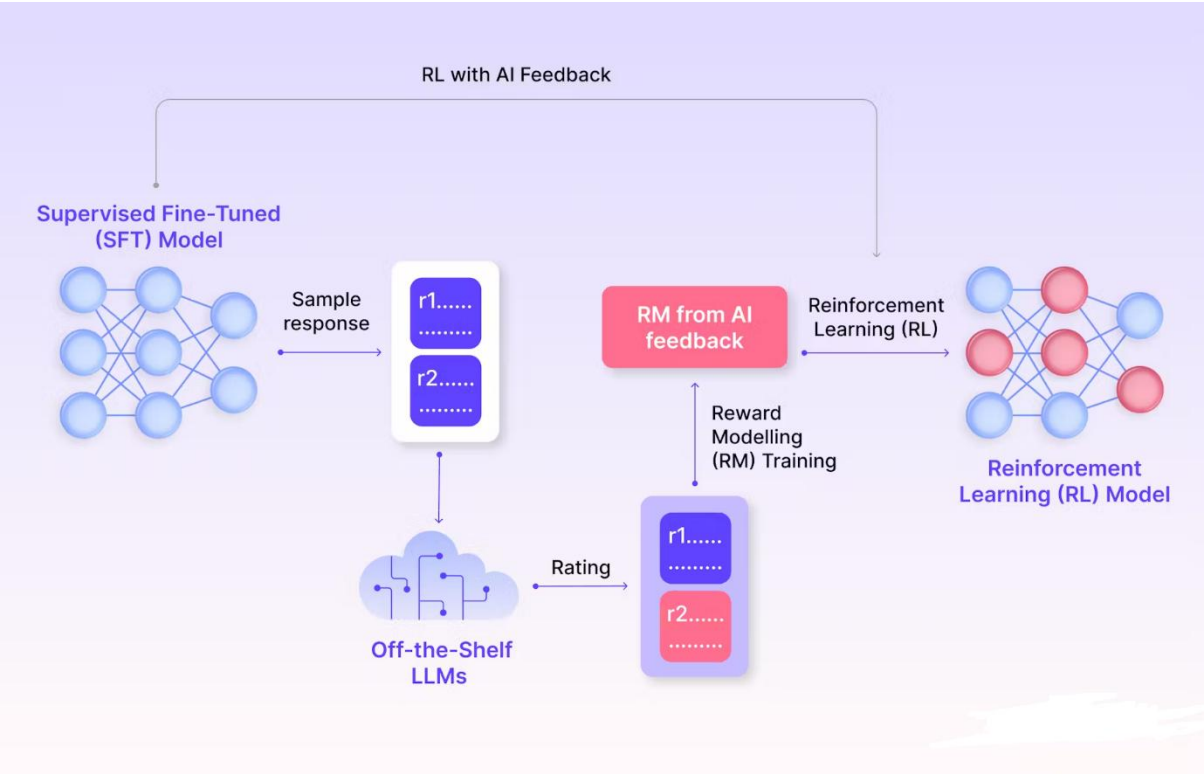

Figure 1: the overview of Reinforcement learning from AI feedback[1]

This paper looks at a hybrid framework that combines RLHF and RLAIF with a group of standard machine learning classifiers, such as SVM, RF, GB, and NB. The goal is to make the system better at learning from both human insights and AI-generated assessments in a way that helps it classify more types of documents better. It is also important to not just check how accurate the system is, but also how well it can generalise, how easy it is to understand, and how well it can replicate human reasoning in difficult situations. This study shows how important it is to develop AI models that put human values, constant learning, and understanding between AI and its users first, as AI systems are used more and more to make decisions in the real world. The next part goes into further information about the past research that is linked to this study.

II. LITERATURE REVIEW

The following section in Table 1 elaborates the past literatures related to this study in detail.

Table 1: Related Works

AUTHORS AND YEAR	METHODOLOGY	FINDINGS
Joshi (2025)	The study examined RLHF's evolution, RLAIF's comparison, and reward models' involvement in LLM optimisation.	Making feedback collection more efficient, reward model generalisation better, and RLHF ethical. By addressing these difficulties, RLHF could help construct safe and effective AI systems.
Li et al., (2025)	This study focussed on multi-turn engagement with LLMs, listing tasks, evaluation standards, improvement strategies, and current problems for clarity and advancement.	The analysis showed that multi-turn interactions radically change how use and assess LLMs.

Wang et al., (2024)	Presented the Model-in-the-Loop (MILO) annotation framework, which uses AI/ML models.	Three multimodal data annotation experiments show that MILO improves data quality, handling time, and annotator experiences.
Chaudhari et al., (2024)	Examined RLHF using reinforcement learning and the reward model.	While RLHF has several modifications, the essential premise of learning via evaluative feedback remains constant. Reinforcement learning is ideal for this type of learning, as agent formulation, reward feedback, and environment definition evolve.
Pternea et al., (2024)	Evaluated research on RL and LLMs, two fields boosted by Deep Neural Networks.	Due to their sequential decision-making, RL can adapt to NLP tasks, and LLMs' reasoning and real-world knowledge make this synergy successful. Models follow human intent and Responsible AI guidelines.
Chang (2025)	An ethical alignment framework for Large Language Models (LLMs) based on three-branch government structures.	Framework showed how DIKE and ERIS steer language behaviours towards ethics while maintaining independence in knowledge development, ethical monitoring, and contextual interpretation.

Research Gap: There have been big improvements in utilising machine learning to classify documents, but there is still a big gap in research on how to combine RLHF with RLAIF to make systems that adapt to people. Most of the models that are out there only look at performance metrics or static feedback loops. They don't have ways to keep learning from both human understanding and machine inference. Also, there haven't been many research on how these two types of reinforcement affect how easy they are to understand, how much users trust them, and how useful they are in real life, especially in sensitive areas like healthcare and legal documents.

III. METHODOLOGY

This study used a combination of RLHF and RLAIF to build a document classification system that uses a group of classic machine learning models. The first step is to pre-process a dataset of 300 labelled papers that fall into 11 categories. This includes tokenisation, removing stop words, and TF-IDF vectorisation. This study trained four classifiers on the dataset: SVM, RF, GB, and NB. RLHF uses real-time feedback and adjustments from human reviewers during model iterations, while RLAIF uses AI-generated confidence scores and error analysis to improve decision boundaries. An ensemble voting technique combines the predictions of all classifiers, making the results more stable and less biased. Precision, F1-score and Recall metrics are used to measure performance, while feedback loops keep changing the model's parameters. This strategy tries to find a compromise between predicted accuracy, adaptive learning, and interpretability. This way, the system can change to reflect both expert judgement and data-driven optimisation.

IV. RESULTS AND DISCUSSION

The hybrid document classification model created in this study, which combines RF, SVM, GB, and NB, was able to perfectly classify all eleven document classes (Table 2). The classification report shows that the precision, recall, and F1-score are all 1.00 for each class. The overall accuracy on the 300-sample test set is 100%. This means that the model accurately put every document into the right category, with no false positives or false negatives. The support values for the classes ranged from 14 to 47. Even though the classes were not evenly

distributed, the performance stayed the same, showing that the ensemble model is strong enough to work with different sample sizes.

Table 2: Analysis of the results

Class	***Precision***	***Recall***	***F1-score***	***Support***
0	*1.000*	*1.000*	*1.000*	*18*
1	*1.000*	*1.000*	*1.000*	*15*
2	*1.000*	*1.000*	*1.000*	*47*
3	*1.000*	*1.000*	*1.000*	*33*
4	*1.000*	*1.000*	*1.000*	*14*
5	*1.000*	*1.00*	*1.00*	*37*
6	*1.00*	*1.00*	*1.00*	*45*
7	*1.00*	*1.00*	*1.00*	*23*
8	*1.00*	*1.00*	*1.00*	*14*
9	*1.00*	*1.00*	*1.00*	*31*
10	*1.00*	*1.00*	*1.00*	*23*
Accuracy			***1.00***	***300***
Macro Avg	*1.00*	*1.00*	*1.00*	*300*
Weighted Avg	*1.00*	*1.00*	*1.00*	*300*

Even though these findings may point to a very good categorisation system, need to look closely at what a perfect score means. In the actual world, especially in areas like healthcare or legal paperwork, the look of faultless correctness typically means that the model has overfitted. Overfitting happens when a model learns not only the patterns in the training data but also the noise, which makes it do well on data it has seen before but poorly on data it hasn't seen before. The worry here is that the hybrid model might have learnt the dataset's specific patterns too well, especially if the training and test data are structurally similar or if the dataset isn't complicated enough.

In their RLHF reviews, Chaudhari et al. (2024) and Joshi (2025) stress the significance of constructing models that balance performance, generalisation, and human expectations. RLHF systems aim to improve AI outputs by aligning them with human values and judgement. This work employed RLHF to alter model predictions based on human feedback and RLAIF to refine them using AI-generated confidence scores and error patterns. These techniques can increase alignment and responsiveness, but their efficacy depends on how well they generalise and adapt across contexts.

With 100% accuracy, the model presumably saw highly organised or restricted dataset variance. RLHF works best with diverse, realistic datasets that challenge the model to read subtle feedback and change dynamically, according to Chaudhari et al. Even RLHF and RLAIF might overfit the model by repeating and reinforcing biases or patterns rather than teaching it to distinguish complicated or ambiguous data points.

Joshi (2025) emphasises RLHF repeated feedback loops. Domain experts' feedback should steer the model through increasingly complicated judgements, with feedback loops at each stage to correct and learn. RLHF was used to reflect user corrections in this investigation, although it is unclear if it addressed enough edge situations or ambiguous inputs. If human feedback is limited to simple scenarios, the model may appear accurate but may not have learnt from sufficiently difficult examples.

The outfit design may cause overfitting. Combining numerous high-capacity classifiers with differing feature recognition capabilities can lead to models that learn real distinctions, redundancy, and noise, especially in small datasets. When paired with RLHF and RLAIF, which modify decision

boundaries based on direct feedback, the system risks becoming excessively stiff or over-optimized for training. RLHF improves human-AI alignment, however Chaudhari et al. warn that without exploration and uncertainty methods, it can harden model behaviour.

The model's stability across classes regardless of class size may imply a lack of dataset variability. Documents with uneven formatting, imprecise terminology, or ambiguous categorisation criteria are common in clinical and regulatory document classification assignments. A model that classifies perfectly on a controlled dataset may not withstand such fluctuation. Joshi (2025) suggests that RLHF should promote accurate responses and add adaptive mechanisms to handle ambiguity, such as learning from annotator disagreement or considering alternate valid labels.

The results of this study are astounding, but they also serve as a warning. Human-AI collaboration systems aim for robust, adaptive, and transparent decision-making, not perfect performance, according to the literature. In practice, a model with 99% accuracy but flexibility in edge circumstances and clarity in decision logic may be more valuable than one with 100% accuracy but brittle responses to unexpected input.

This study shows that RLHF and RLAIF can improve performance, but their success depends on feedback quality and variety. Chaudhari et al. recommend prioritising answer accuracy, interpretability, user trust, and adaptability in future models. This matches the practical necessity for AI systems that aid human decision-making rather than copying them.

In a nutshell this work shows that it is technically possible to use hybrid machine learning models using RLHF and RLAIF for high-accuracy document classification. However, the perfect scores seen in the study suggest that the models may be overfitting, which could make it harder to generalise. In the future, work should focus on adding more diverse data, getting more detailed and context-sensitive feedback from domain experts, and looking into ways to make models more flexible and better at handling uncertainty. This way, can better understand the real power of RLHF and RLAIF: making AI systems that work together and smartly with people.

V. CONCLUSION

This study looked at how well a hybrid AI model made consisting of Support Vector Machine, Random Forest, Gradient Boosting, and Naive Bayes could classify human documents into eleven different classes using precision, F1-score and recall metrics. In this research, a 100% accuracy rate and excellent ratings on all measures are impressive. This kind of perfection is rare in real life, raising concerns about overfitting. If the model can predict every sample, it may have memorised the dataset instead of learning patterns for subsequent datasets. It would struggle with new, diverse datasets. The data were examined using RLHF and its more advanced version, RLAIF. These technologies let AI follow human desires. RLHF accelerates learning and simplifies understanding, but it complicates bias, reward modelling, and ethical alignment. On noisy datasets, adding RLHF and RLAIF to human document categorisation may improve generalisation, robustness, and adaptability. This study's 100% accuracy is statistically astounding, but AI in sensitive fields like healthcare and legal papers need models that balance accuracy and generality. Future work should focus on real-world validation, regularisation, and a wide spectrum of human feedback to avoid overfitting and create reliable AI systems. Integrating RLHF frameworks can help achieve these aims by building trust and accountability in human-AI interactions.

REFERENCES

[1]. https://encord.com/blog/reinforecement-learning-from-ai-feedback-what-is-rlaif/

[2]. Bradley, H., Dai, A., Teufel, H., Zhang, J., Oostermeijer, K., Bellagente, M., ... & Lehman, J. (2023). Quality-diversity through AI feedback. *arXiv preprint arXiv:2310.13032*.

[3]. Benaich, N., & Hogarth, I. (2020). State of AI report. *London, UK.[Google Scholar]*.

[4]. Li, Y. (2023). Iterative improvements from feedback for language models. *ScienceOpen Preprints*.

[5]. Daull, X., Bellot, P., Bruno, E., Martin, V., & Murisasco, E. (2023). Complex QA and language models hybrid architectures, Survey. *arXiv preprint arXiv:2302.09051*.

[6]. Jeong, J., Chow, Y., Tennenholtz, G., Hsu, C. W., Tulepbergenov, A., Ghavamzadeh, M., & Boutilier, C. (2023). Factual and personalized recommendations using language models and

reinforcement learning. *arXiv preprint arXiv:2310.06176*.

[7]. Zhu, Y., Liu, Y., Stahlberg, F., Kumar, S., Chen, Y. H., Luo, L., ... & Meng, L. (2023). Towards an on-device agent for text rewriting. *arXiv preprint arXiv:2308.11807*.

[8]. Joshi, S. (2025). Introduction to Reinforcement Learning from Human Feedback: A Review of Current Developments.

[9]. Chaudhari, S., Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., ... & Castro da Silva, B. (2024). Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *ACM Computing Surveys*.

[10]. Pternea, M., Singh, P., Chakraborty, A., Oruganti, Y., Milletari, M., Bapat, S., & Jiang, K. (2024). The rl/llm taxonomy tree: Reviewing synergies between reinforcement learning and large language models. *Journal of Artificial Intelligence Research*, *80*, 1525-1573.

[11]. Wang, Y., Stevens, D., Shah, P., Jiang, W., Liu, M., Chen, X., ... & Kamma, B. (2024). Model-in-the-Loop (MILO): Accelerating Multimodal AI Data Annotation with LLMs. *arXiv preprint arXiv:2409.10702*.

[12]. Li, Y., Shen, X., Yao, X., Ding, X., Miao, Y., Krishnan, R., & Padman, R. (2025). Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*.

[13]. Chang, E. Y. (2025). A Three-Branch Checks-and-Balances Frameworkfor Context-Aware Ethical Alignment of Large Language Models. *arXiv preprint arXiv:2502.00136*.