Group Related Policy Optimization-Enhanced Framework for Text-Based Fake News Detection

Ali Mohammad Salloum[a], Bassel Alkhatib[a]

[a]*Syrian Virtual University, Damascus, Syria*

Abstract

The swift dissemination of disinformation is a quickly emerging obstacle in today's digital era, and social media websites are accelerating the propagation of false content, which is frequently backed up by AI-created content that is hard to distinguish from actual information. The aim of this research is to improve the effectiveness and precision of a fake news detection system by assessing the effect of Group Relative Policy Optimization (GRPO) on the performance metrics of two prominent large language models, specifically Qwen 2.5-3B-Instruct and LLaMA 3.2. The two models are assessed independently with an equalized dataset containing original and generated news to allow easy comparison of their performance outcomes with and without the application of GRPO. The method employs a twin reward function design that prioritizes response succinctness as well as classification accuracy. Optimization techniques such as quantization and LoRA are employed to counteract the computational burden. The evaluation results indicate that GRPO achieves: impressive accuracy gains, from 73% to 87% for Qwen, and from 59% to 93% for LLaMA, improved F1-scores and balanced classification for both classes (Fake/Real), a sharp reduction in misclassifications and false positives, and considerable improvement in ROC and Precision-Recall curves, in terms of increased AUC values. This study represents the first recorded implementation of GRPO for fake news detection and provides a detailed examination of its implications on various language models. The results highlight the power of GRPO as an improved training methodology that enhances model credibility and efficiency and paves the way for novel mechanisms to tackle misinformation in dynamic media landscapes.

Keywords: Fake News Detection, Large Language Models, Qwen 2.5, Llama 3.2, Group Relative Policy Optimization (GRPO), Reinforcement Learning, Explainability, LoRA Fine-Tuning, Quantization,

1. Introduction

The escalating proliferation of fake news and misinformation across digital platforms poses a significant threat to public trust, social stability, and the integrity of information ecosystems [1, 2]. As online channels increasingly become primary sources of news, malicious actors exploit their extensive reach to disseminate deceptive content rapidly and at scale, often outpacing traditional verification mechanisms [1]. This pervasive challenge necessitates the urgent development of intelligent and advanced automated fake news detection systems, especially given the evolving sophistication of misinformation, now frequently generated by advanced artificial intelligence capable of producing highly convincing textual content [3]. Indeed, recent studies highlight that the spread of misinformation via social networks constitutes a contagious threat to public health, influencing societal decisions and undermining trust in official institutions [2]. Furthermore, the continuous evolution in misinformation dissemination technologies underscores the critical need for effective detection solutions to mitigate widespread impact.

Initial efforts in fake news detection predominantly utilized traditional machine learning (ML) models such as Support Vector Machines (SVM) and Random Forests [4, 5]. These models typically relied on basic embedding techniques like TF-IDF and Bag of Words (BoW) for feature extraction [4]. However, while capable of capturing rudimentary statistical relationships, these methods fundamentally lacked the capacity to model nuanced contextual and semantic depen- dencies. Consequently, they exhibited limited generalization capabilities and struggled to fully comprehend the intricate context of news articles [1]. For instance, studies have demonstrated that models relying on such embeddings are prone to overfitting on surface-level patterns and face considerable difficulties with domain adaptation, particularly when confronted with novel or evolving misinformation tactics [1]. Moreover, current fake news detection models often suffer from a heavy reliance on hand-crafted features and a limited ability to adapt to new domains [6]. Challenges also include the wide variability of news domains, the high cost associated with manual data labeling, and significant obstacles posed by data bias and class imbalance within available datasets [6].

Recent advancements have seen a notable shift towards deep learning (DL) and transformer- based architectures, most prominently BERT and its variants [1, 7]. These models leverage contextual embeddings to provide a richer representation of language, effectively capturing com- plex syntactic

and semantic features that are crucial for distinguishing subtle differences between authentic and fabricated news [1]. For example, a BERT-based model achieved up to 99.9% accuracy on the Kaggle fake news dataset and over 98% on other public datasets [1]. Similarly, ensemble approaches that integrate BERT with models like SVC have further improved predictive performance by leveraging complementary model strengths [7]. The year 2023 also marked the development of advanced hybrid models such as DeepFND, which integrates VGG-19 and Bi- LSTM architectures to enhance misinformation detection, consistently outperforming traditional ML methods [8]. Despite these performance gains, transformer-based models are not without limitations. Their considerable size and substantial computational demands often preclude local deployment, rendering them less accessible for real-time or resource-constrained environments [1].

Additionally, BERT-based systems can exhibit domain overfitting, reduced interpretability, and challenges in adapting to low-resource or multilingual settings [1]. A critical issue for many AI-based fake news detection systems remains their lack of interpretability, meaning their inability to provide clear justifications for their classification decisions [7].

To address these aforementioned challenges, our study proposes a novel paradigm that in- tegrates advanced generative language models—specifically Qwen 2.5-3B-Instruct [9, 10] and Llama 3.2 [11, 12]—with Group Relative Policy Optimization (GRPO) [13]. These generative models, available in compact sizes (3 billion parameters), are pre-trained on large, up-to-date corpora, facilitating local deployment and efficient inference. They demonstrate remarkable con- textual understanding and flexibility, making them highly suitable for dynamic and multilingual fake news detection scenarios [3]. The integration of these generative models with GRPO, an advanced reinforcement learning technique, represents a significant advancement over conven- tional methods. GRPO enables models to learn optimal classification behaviors from relative feedback rather than explicit reward signals [14], thereby enhancing adaptability and efficiency [15], particularly in settings with limited labeled data [16]. This approach is especially critical in the fake news context, where high-quality labeled data can be scarce or costly [6]. A key advantage of our methodology is the ability to run these models locally, overcoming the computational demands typically associated with larger transformer models [1].

We achieve this efficiency through the strategic application of LoRA (Low-Rank Adaptation) [17] for parameter-efficient fine-tuning and quantization [18] to further reduce computational overhead. The utilization of the Unsloth library [19] further optimizes this process, enabling efficient training and deployment of these large language models. This combination makes our systems practical for real-world, on-device applications, significantly improving deployment flexibility and accessibility. This study makes several key contributions to the field of fake news detection:

- We present the first documented application of Group Relative Policy Optimization (GRPO) [20] in the context of fake news classification. GRPO is an advanced reinforcement learning technique that optimizes model performance based on relative outcome comparisons, offering simplicity in implementation, enhanced performance, and reduced variance during training compared to conventional reinforcement learning methods.
- We provide a comprehensive empirical analysis of using the Qwen 2.5-3B-Instruct and Llama 3.2 models for fake news classification, emphasizing their ability to understand complex linguistic context.
- We propose GRPO as a promising and novel alternative to traditional fine-tuning methods in textual classification tasks, offering a detailed analysis of its advantages and practical applications.
- The study includes a thorough experimental evaluation of the proposed methodology using well-established benchmark datasets, providing detailed comparisons with both conventional and state-of-the-art approaches.

By integrating these contributions, we strive to advance fake news detection methodologies by offering solutions that are both resilient and adaptable amidst the continuously changing nature of misinformation. The rest of this paper is organized as follows: The Methodology section describes the GRPO framework, its mathematical formulation, and pseudocode, along with the particular reward functions utilized. The Results and Discussion section illustrates GRPO's efficacy on various datasets, showing its versatility and stability, and then examines the consequences of GRPO in model performance improvement. Finally, the conclusion recapitulates the key findings and indicates directions for future research.

2. Related Work

Fake news detection has been an intensely studied subject, moving from conventional manual feature-based approaches to recent deep learning and transformer models. The majority of the earlier studies made extensive use of hand-crafted features and rule-based systems. Shu et al. [21] provided an extensive survey of fake news detection methods, highlighting conventional approaches that make use of linguistic, stylistic, and network-based features. Their work used classifiers like Support Vector Machines (SVM) and Random Forests on social media data with accuracy rates of 70-75%.

The methods, however, had grave problems being scalable and domain adaptable since they relied

on hand-crafted features and had low context understanding. In the same vein, Horne and Adali [22] compared fake news through the extraction of stylometric and content-based features, with modest success using classifiers such as Logistic Regression and Decision Trees, but pointed out the failure of these approaches to address the dynamic nature of misinformation and the absence of semantic understanding.

With the introduction of machine learning (ML), the detection systems became more performance-focused and efficient. Alghamdi et al. [23] reviewed ML-based models in detecting fake news, such as ensemble methods and gradient boosting classifiers, which achieved up to 85% accuracy improvements over benchmark datasets. It was achieved through feature extraction of textual content and incorporating it with user interaction metrics. Kh. Hamed et al. [24] suggested a hybrid ML solution with TF-IDF and word embeddings and ML classifiers like XGBoost, with F1-scores around 0.87. In spite of these improvements, ML models are still not able to model deeper semantic relations and context-dependent subtleties, and hence their generalization ability
to real-world heterogeneous datasets is poor. They perform well on in-domain data but break down when faced with new topics or language styles, a weakness in the face of dynamic misinformation ecosystems.

Deep learning methods, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have promoted substantially the detection of fake news through automatically learning hierarchical and temporal features. Huang [25] showed that deep learning architectures can capture complex linguistic patterns and interdependencies and achieve accuracy gains over 90% on carefully compiled datasets. They employed Bi-LSTM networks combined with attention mechanisms to enhance feature representation. However, DL methods require enormous amounts of labeled data and are plagued by overfitting, especially on noisy or imbalanced data. Moreover, their black-box character is problematic from the viewpoints of interpretability and trustworthi- ness, which are crucial in the case of detecting misinformation.

Transformer-based models, as represented by BERT, have set new benchmarks in performance owing to their strong contextual representation and bidirectional attention capabilities. Devlin et al. [26] introduced BERT, achieving state-of-the-art performance on a range of natural language processing tasks, such as fake news classification, with accuracy often greater than 92%. BERT and other large transformer models, though, are very resource-intensive, requiring significant memory and computation resources, and thus are limited in their deployment in real-time systems or resource-constrained environments. Their inferential latency and size disrupt scalability and local execution, making them undesirable for most real-world applications despite their high accuracy.

In response to these constraints, recent research has shifted its attention to more effective generative language models, specifically Qwen 2.5-Instruct and LLaMA 3.2-Instruct, both of which have roughly 3 billion parameters. These models achieve a reasonable compromise between their size, speed, and contextual comprehension, making local deployment and quicker inference possible. Their inductive generative nature enables them to pick up on more subtle semantic relationships and implicit context than conventional classifiers. With the addition of reinforcement learning (RL) techniques, specifically Group Relative Policy Optimization (GRPO), these models can be efficiently calibrated to classification tasks, i.e., detection of false information. GRPO improves model performance by performing relative comparisons between groups of possible outputs, thus encouraging stability and sample efficiency without the use of a value network. In addition, techniques that enable parameter-efficient tuning, such as LoRA and quantization, reduce computational requirements, making this approach highly suitable for practical applications. This new approach leverages the merits of robust contextual comprehension, capacity for generaliza- tion, and operational effectiveness, resolving core issues encountered in the implementation of fake news detection systems in complex, real-world settings.

3. Problem Formulation

Our study suggests a novel paradigm that incorporates cutting-edge generative language models—specifically, Qwen 2.5-3B-Instruct—to overcome the aforementioned issues. [9, 10] and Llama 3.2 [11, 12]—utilizing GRPO, or group relative policy optimization [13]. These generative models are pre-trained on sizable, current corpora, allowing for local deployment and effective inference. They come in compact sizes (3 billion parameters). They are ideal for dynamic and multilingual fake news detection scenarios because of their exceptional contextual comprehension and adaptability. [3]. Proximal Policy Optimization (PPO) is a reinforcement learning algorithm based on the **actor-critic** framework, where both the policy model (actor) and the value model (critic) are trained simultaneously [20]. The advantage is calculated as the difference between the actual reward and the expected value predicted by the critic:

$$A_t = R_t - V(s_t), \tag{1}$$

and the policy is updated using the PPO objective:

$$J_{PPO}(\theta) = \mathrm{E}_{q\sim P(Q),o\sim\pi^{old}(O|q)}\ \frac{1}{|o|}\sum_{t=1}^{|o|}\min\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi^{old}(o_t|q,o_{<t})}A_t,\ \mathrm{clip}\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi^{old}(o_t|q,o_{<t})},1-\epsilon,1+\epsilon\right)A_t\right) \tag{2}$$

[20].

$\pi^{old}(o_t|q, o_{<t})$ θ

However, PPO requires training a separate value network, which increases **computational cost and memory usage**, especially when handling large language models or long contexts [27]. This also limits its feasibility for **local deployment** due to resource constraints [27].

In contrast, Group Relative Policy Optimization (GRPO) eliminates the need for a critic network by estimating the advantage internally through **generating a group of outputs** for the same input question $\{o_i\}^G$ [20]. The mean μ_G and standard deviation σ_G of the rewards across the group are computed, and the relative advantage for each output is calculated as:

$\hat{A}_{i,t}$

$\frac{R_{i,t} - \mu_G}{\sigma_G}$ = [20, 27]. (3)

his relative advantage is then used to update the policy using the GRPO objective:

$|o_i|$

$$J_{GRPO}(\theta) = \mathrm{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_\theta^{old}} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_\theta^{old}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t},\right.$$

$$\left.\text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi^{old}(o_{i,t}|q,o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right] - \beta D_{KL}(\pi_\theta||\pi_{ref}) \text{ [20, 27].} \quad (4)$$

This approach allows GRPO to align more closely with the **human preference evaluation in RLHF**, which relies on relative comparisons among multiple outputs rather than absolute value estimation [20]. The internal, dynamic baseline reduces reward noise and improves training stability [27]. Additionally, by **removing the critic network**, GRPO significantly reduces memory and computational requirements, enabling **local deployment of large language models** or training in resource-constrained environments without compromising efficiency or stability [20, 27]. This is particularly advantageous when handling large models with long contexts [27]. An important improvement over traditional approaches is the combination of these generative models with GRPO, a sophisticated reinforcement learning technique. Instead of using explicit incentive signals, GRPO allows models to learn optimal categorization behaviors from relative feedback. [14], increasing efficiency and adaptability [15], especially in environments with little labeled data [3]. This strategy is particularly important in the case of fake news, when high-quality labeled data may be expensive or hard to come by. [6].

The ability to run these models locally is a significant benefit of our methodology, since it circumvents the computational needs usually associated with bigger transformer models [1]. In order to further reduce computational overhead, we strategically apply LoRA (Low-Rank Adaptation) [17] for parameter-efficient fine-tuning and quantization [18]. This procedure is further optimized by using the Unsloth library [19], which makes it possible to train and deploy these huge language models effectively. Because of this combination, our technologies are useful for on-device, real-world applications, greatly increasing accessibility and deployment flexibility. This study significantly advances the field of fake news identification in a number of ways:

- In the context of classifying fake news, we offer the first known use of Group Relative Policy Optimization (GRPO) [20]. In comparison to traditional reinforcement learning techniques, GRPO is a sophisticated strategy that optimizes model performance based on relative result comparisons. It offers lower variation during training, improved performance, and ease of implementation.
- We offer a thorough empirical examination of the Qwen 2.5-3B-Instruct and Llama 3.2 models' use in the classification of fake news, highlighting their capacity to comprehend intricate language context.
- We provide a thorough examination of the benefits and real-world uses of GRPO, a potential and innovative substitute for conventional fine-tuning techniques in textual categorization problems.
- Using reputable benchmark datasets, the paper conducts a comprehensive experimental examination of the suggested methodology and presents in-depth comparisons with both traditional and cutting-edge methods.

4. The Propose[d Approach

With an emphasis on differentiating between fake and true news, this paper offers a thorough methodology for improving the performance of big language models in the binary news classifica- tion challenge. Because of its capacity to increase model stability and diversify outputs during learning, the Group Relative Policy Optimization (GRPO) algorithm serves as the foundation for the suggested framework's training component [28]. Three interconnected and sequential stages make up the framework. In order to guarantee alignment with task-specific constraints, the first

step is initializing the model using instruction-tuning weights and data preprocessing. The second step involves creating a multi-objective reward function that addresses balanced categorization across many categories while striking a balance between accuracy, dependability, and safety. Groups of outputs are compared in the third stage's adaptive training loop, which is led by GRPO and helps the model become more predictive. Furthermore, without changing the fundamental model architecture, Low-Rank Adaptation (LoRA) approaches can be used to lower computational cost during fine-tuning.

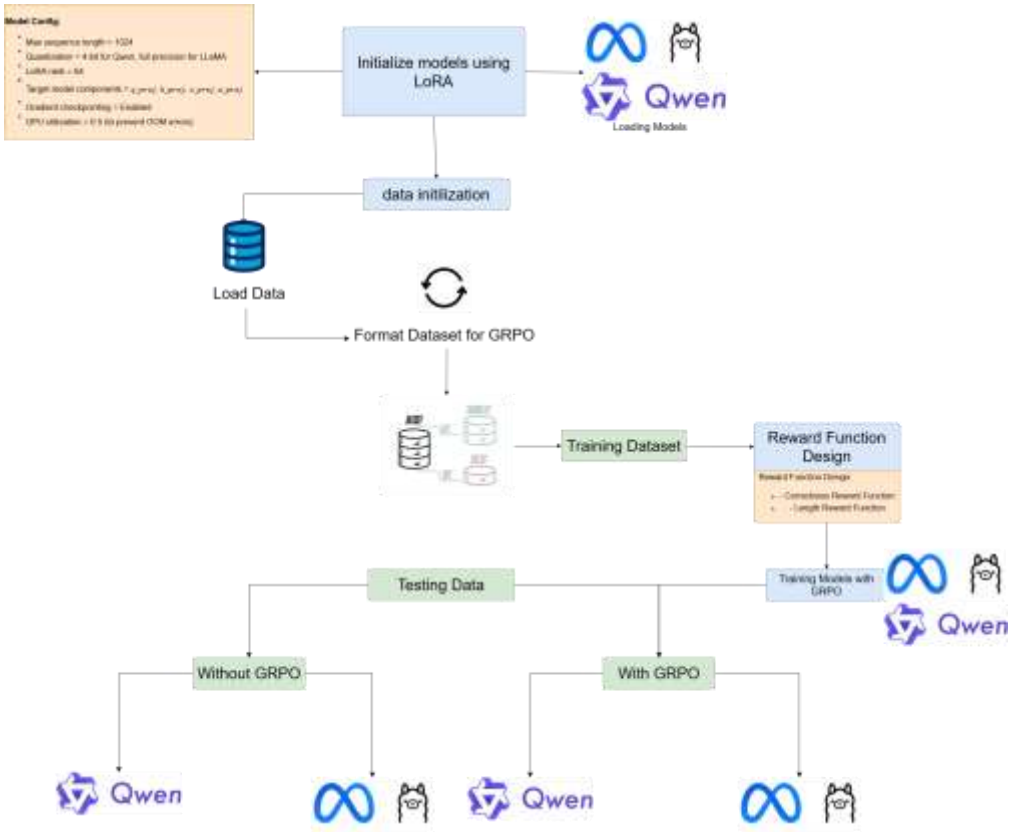

Figure 1: Overview of the training and testing pipeline using GRPO for fine-tuning the Qwen model and LLAMA model with LoRA.

Improving classification accuracy and lowering computational cost are the framework's two main objectives. By making it possible to compare outputs within groups, GRPO is essential for improving training stability and lowering the possibility of local optima. At the same time, LoRA [17] helps with effective model adaption with little change in parameters [29].
Together, GRPO and LoRA greatly increase inference speed and resource efficiency while producing clear and trustworthy outputs (such as "Real" or "Fake"). Results from experiments show significant gains in performance on a variety of false news detection datasets [13, 16, 17].

Algorithm 1 ExtractFakeNewsClassification (Inference)
Input: Input text or news article x
Fine-tuned Language Model (LLM) M
Output: Predicted label y (e.g., fake or real)
for *Step 1: Preprocess x (tokenize, clean, etc.)* **do**
for *Step 2: Feed x into LLM* M *for inference* **do**
for *Step 3: Obtain prediction y* = M(x) **do**
for *Step 4: Postprocess y (thresholding, mapping probabilities)* **do**
for *Step 5: Return predicted label y* **do**

Algorithm 2 GRPOTrainingAlgorithm (Training)
Input: Training dataset D = $\{x_1, x_2, \ldots, x_N\}$
Reward model R
Initial policy (language model) parameters θ_0
Group size G Number of epochs E Learning rate α
Output: Trained policy parameters θ^*
Initialize policy parameters $\theta \leftarrow \theta_0$ **for** *epoch* = 1 **to** E **do**
foreach *input prompt x in dataset* D **do**
// Group sampling: Generate G candidate responses for each prompt

8 **for** $i = 1$ **to** G **do**
9 Sample response $o_i \sim \pi_\theta(\cdot|x)$ Compute reward $r_i = \mathrm{R}(x, o_i)$
10 Compute group reward average: $r_{\text{avg}} = \frac{1}{G} \cdot \sum_{i=1}^{G} r_i$ // Compute relative advantages for each response
11 **for** $i = 1$ **to** G **do**
12 $A_i = r_i - r_{\text{avg}}$
G i=1
// Compute GRPO loss for this prompt
13 $\mathrm{L}_{\text{GRPO}} = -\frac{1}{G} \cdot \sum_{i=1}^{G} A_i \log \pi_\theta(o_i|x)$ Accumulate L_{GRPO} over the batch
// Gradient step: Update policy parameters
14 Update $\theta \leftarrow \theta - \alpha \nabla_\theta \mathrm{L}_{\text{GRPO}}$
15 Return trained policy θ^*

4.1. model and data initialization

The initialization of the Qwen 2.5-3B-Instruct model [9] and LLaMA 3.2-3B-Instruct are two advanced instruction-tuned language models that are chosen and configured at the beginning of the initialization process. LLaMA 3.2 [12] provides multimodal capabilities, a context window of up to 128,000 tokens, and effective deployment features like knowledge distillation and quantization, whereas Qwen supports long-context processing (up to 32,000 tokens) and enjoys the advantages of robust pretraining and reinforcement learning from human feedback (RLHF) [30].

Low-Rank Adaptation (LoRA), which introduces small trainable low-rank matrices into spe- cific layers of a frozen pre-trained model, was used with the Unsloth framework to effectively fine-tune both models [17]. While maintaining performance, this method drastically lowers the number of trainable parameters. The adapters were injected into the key transformer compo- nents (q proj, v proj, and gate proj) using a uniform LoRA configuration (rank = 64, lora alpha = 64).

The maximum sequence length was truncated at 1024 tokens, based on dataset token length analysis. To optimize GPU memory usage, Qwen was quantized to 4-bit precision, while LLaMA was kept in full precision for accuracy-critical inference. Both models also leveraged FlashAt- tention, a memory-efficient attention mechanism designed to improve training and inference throughput [31, 32].

Furthermore, gradient checkpointing was enabled to minimize memory usage during training, contributing to approximately 72% memory savings without sacrificing model performance. These combined optimizations reflect a resource-aware configuration strategy that facilitates efficient adaptation in constrained computational environments.

For data preprocessing, this study employs the "Fake and Real News Dataset" curated by Nitis Jolly, publicly available on Kaggle. This dataset is widely used in natural language processing (NLP) research, particularly in fake news detection tasks. It consists of 9,865 labeled news samples, approximately balanced between the "real" and "fake" categories. Each entry contains a news article along with its ground truth label, making the dataset highly suitable for training unbiased and robust binary classifiers.

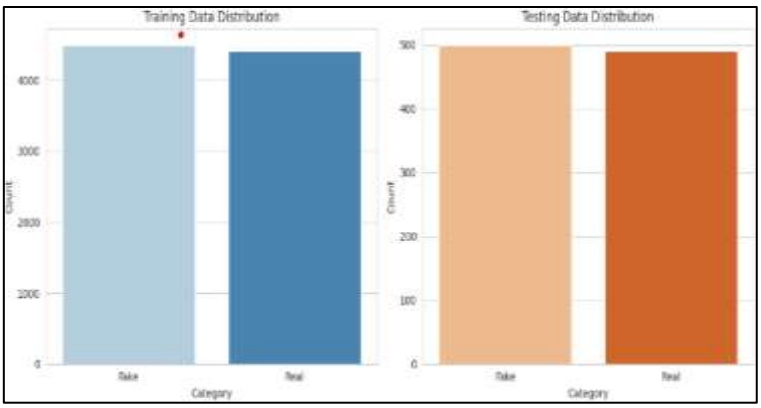

Figure 2: Dataset Distribution

Stratified sampling was used to divide the data into 80% for training and 20% for testing, while maintaining the original class distribution. Prompts were reformatted into a supervised instruction format, specifically: *"Classify the following news article as 'Real' or 'Fake': The text of each news item — represented as news text, is paired with the correct label as the target output."* This format aligns the dataset with instruction-tuned frameworks such as Qwen and LLaMA.

The majority of samples fall below the 1024-token limit, according to token length analysis, which supports the choice of this maximum sequence length for data integrity and memory efficiency. Consistent batch processing and efficient use of computational resources are further advantages of standardizing input lengths, which are especially beneficial in low-resource settings.

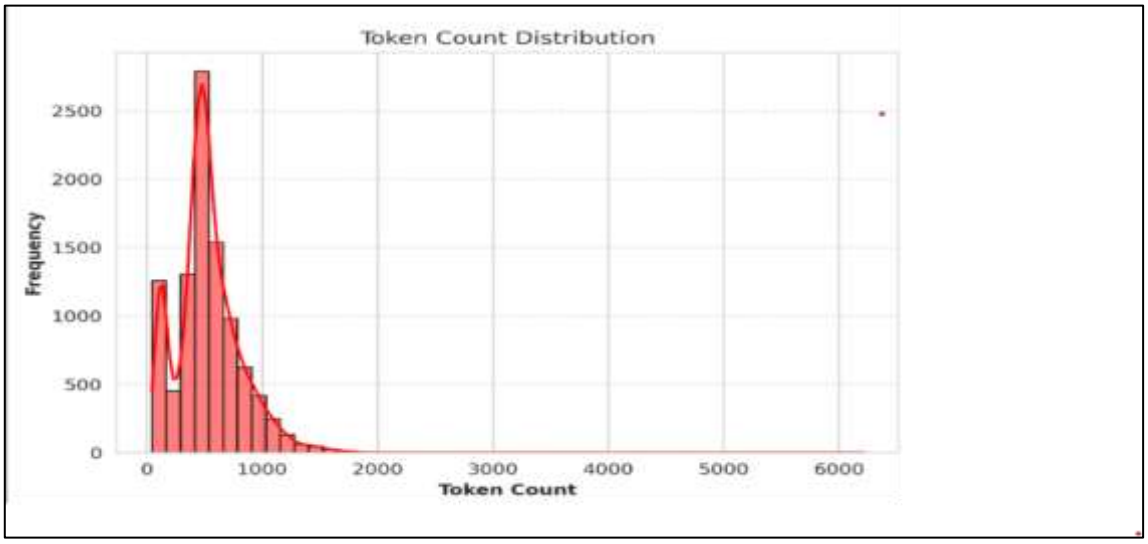

Figure 3: Token Count Distributions

The model's ability to effectively detect disinformation was improved by this structured preprocessing, which enabled seamless integration with instruction-based reinforcement learning algorithms such as Group Relative Policy Optimization (GRPO) [15, 20].

4.2. Dataset Clarification

The Kaggle 'Fake and Real News Dataset' [33, 34, 35] was selected for its balanced class distribution and temporal relevance in early fake news detection research. While the dataset was published in 2017, and age might limit direct usage with current news trends, it provides an adequate baseline for model performance measurement. the original dataset contains 9,865 samples. We performed 90/10 train-test split, such that the test set had 990 samples (500 real news and 490 fake news) as recorded. This division between the entire dataset and the test set is purposely labeled to ensure transparency and reproducibility of the experimental setup.

4.3. Reward Function Design

As the primary method for directing the learning process, reward functions are essential to fake news systems. They have a direct effect on the effectiveness and quality of model optimization by rewarding desired model behaviors and penalizing wrong outputs [36].

A number of important factors must be taken into account when creating a reward function for detecting fake news. Among the most crucial are accuracy rewards, which support multi-response differential evaluation techniques by rewarding accurate classifications and penalizing incorrect ones.

Furthermore, text length-related supplementary rewards might encourage the production of succinct and effective outputs, which is essential in settings requiring quick answers or functioning with limited computational resources [37]. By improving learning signals through multi-sample assessment, reward functions can be included into reinforcement learning frameworks like GRPO, giving the model the ability to make more accurate and knowledgeable decisions [38].

In real-world applications, output reliability and usability are practically integrated when precision and conciseness are balanced. In addition, well-crafted reward functions increase sample efficiency during training, making every training phase more effective and significant [39]. To regulate the reward scale and stabilize the optimization process without adding undesired biases, mathematical transformations of reward functions can also be used [20].

Modern frameworks enable models to achieve excellent performance in fake news identifi- cation while preserving computational economy and generalization ability across a variety of contexts by using an integrated approach to reward function construction.

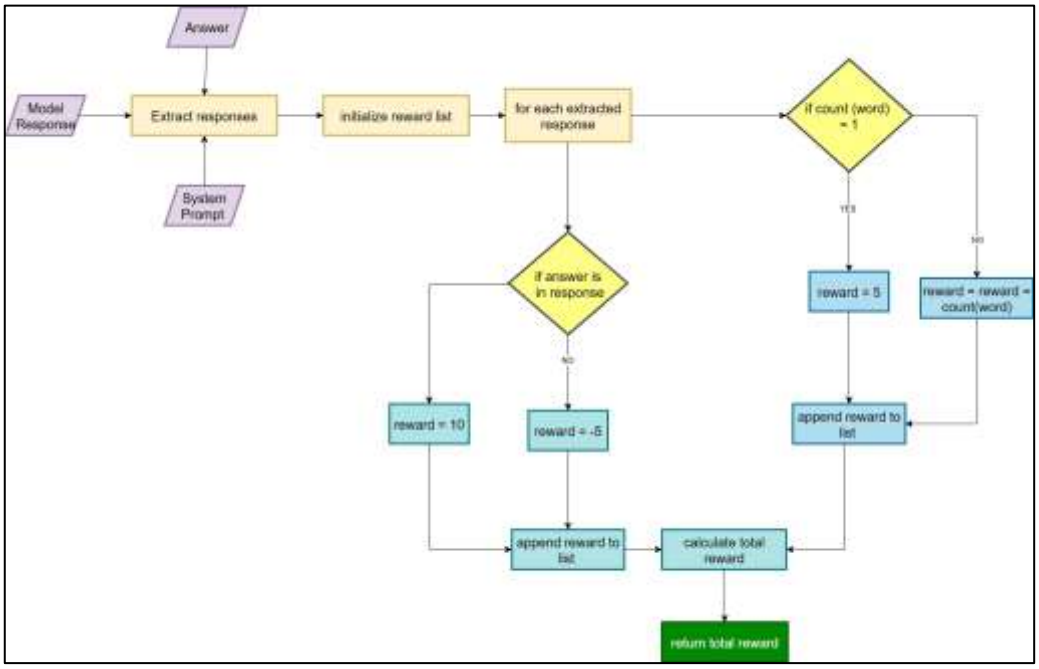

Figure 4: Correctness and Length Reward Functions Design

Our reward function was deliberately designed to align the model's behavior with the goals of fake news detection—namely, producing accurate and concise binary outputs ("Fake" or "Real"). The reward formulation provides a positive signal for correct classifications and a negative signal for incorrect predictions, with an additional penalty term for unnecessarily long or irrelevant text generations. This brevity constraint is intentional, as the fake news detection task benefits more from decisive and interpretable outputs rather than verbose reasoning chains.

Although Group Relative Policy Optimization (GRPO) is frequently associated with tasks requiring complex, multi-step reasoning, its relative preference-based optimization mechanism is well-suited for binary classification tasks. GRPO learns by comparing multiple generated responses within a group and reinforcing the policy toward the relatively better-performing output. In this context, "better" corresponds to correct and concise predictions. Unlike Direct Preference Optimization (DPO) or classical supervised fine-tuning, which depend on static labels or pairwise preferences, GRPO enables dynamic comparison and adaptive normalization, stabilizing training even with limited reward variability.

Therefore, GRPO provides a principled and efficient fine-tuning framework for binary classifi-cation without requiring explicit Chain-of-Thought (CoT) generation. Its comparative learning paradigm effectively distinguishes preferred (accurate and concise) outputs from dispreferred (incorrect or verbose) ones, ensuring robustness and stability in training small instruction-tuned LLMs for discriminative tasks like fake news detection [40, 41, 42].

4.4. Fine-tuning details

4.4.1. LoRA Fine-tuning Setup

To enhance parameter efficiency during model adaptation, we employed the use of Low-Rank Adaptation (LoRA). A configuration of rank=64 and α = 64 was utilized, as typical practices within parameter-efficient fine-tuning literature that achieve a balance between expressiveness and computational expense [17]. Such a selection has a 1:1 rank-to-scaling factor ratio that is typically used to stabilize training dynamics at the cost of losing the benefits of low-rank factorization [17, 32]. Specifically, a rank of 64 provides sufficient representational capacity to be able to learn task-specific patterns in detecting false news without adding parameter overhead of over approximately 0.5–2The choice of α as 64 provides sufficient scaling of adaptation matrices to prevent gradient instability problems such as vanishing or exploding updates. Although no ablation experiment was carried out for other rank or α values (e.g., 32, 128), this restriction is one where future research could explore more thoroughly the sensitivity of LoRA hyperparameters for the fake news classification task.

4.4.2. Quantization Strategy Justification

The differential quantization approach employed in this study is the outcome of strategic optimization after model-specific robustness characteristics observed in recent studies. Qwen2.5 models are observed to be exceedingly robust to aggressive quantization, with 4-bit precision displaying near-baseline performance across a broad diversity of downstream tasks at significantly reduced memory footprint and computational expense [43]. This resilience is due to architectural and training advances that enhance the model's robustness to precision degradation, with 4-bit quantization in particular being amenable to resource-constrained deployment environments.

LLaMA models, on the other hand, exhibit greater sensitivity to quantization-induced degradation, particularly at below 8-bit precisions where performance degradation is considerable across language modeling and reasoning tasks [44]. Thus, FP16 precision was selected for LLaMA to preserve significant model functionality while still achieving notable memory savings over FP32 baselines. This selective quantization strategy is an empirical solution to balancing computational efficiency and task performance by guaranteeing that each model falls within its sweet spot precision window. The setup that results from this maximizes the use of available resources without compromising fine-tuning effectiveness required by specialty applications such as fake news detection.

4.5. GRPO Training

Fake news classification performance is greatly improved by the Group Relative Policy Optimization (GRPO) algorithm, which combines the Actor-Critic framework with the Proximal Policy Optimization (PPO) method. This algorithm introduces a novel mechanism to compare rewards within groups of candidate responses [45, 28].

Group Relative Policy Optimization (GRPO) offers a method for comparing collections of policy outputs with each other to assess relative preferences en masse, compared to assessing samples separately as in Direct Preference Optimization (DPO). Graspingly, GRPO reduces reward

variance by leveraging group aggregations that mitigate outlier effects and exploit preference data across generations of candidates. This mechanism generates more robust policy enhancements and facilitates effective learning, particularly for reinforcement learning tasks for large language models (LLMs), where reward signals can be noisy or sparse. Experimental results confirm that GRPO generates lower variance in estimated rewards and converges faster towards the policy optimization for alignment and safety-critical goals. Additionally, the group-based approach supports generalization and consistency over evaluation spaces, providing tangible advantages compared to DPO's sample-based optimization strategies [46, 47, 48].

The policy π, which maps states to actions in order to maximize cumulative rewards, is at the heart of reinforcement learning in this context. Value functions that reflect the expected return of doing an action in a particular state, such as the state-value function $V(s)$ and the action-value function $Q(s, a)$, are used to assess the efficacy of policies [20, 28]. The choice of group size G in the GRPO implementation is motivated by previous studies in recent literature that indicate that generating multiple candidate responses to a given input allows for more robust and less varying estimation of the relative advantage signal [20, 27]. Large values of G improve gradient update stability and reduce variance in reward normalization but cause increased computational cost per iteration. Conversely, smaller values for G yield faster training iterations but may lead to higher stochasticity and less robust advantage estimation. While proper sensitivity analysis over different group sizes was not performed in this study, it is an extremely crucial direction for future systematic exploration of the trade-off between policy stability and computational cost.

The GRPO model provides a wider evaluation space by employing the policy π_θ to create numerous candidate responses $\{o_1, o_2, \ldots, o_G\}$ for each query q. Accuracy, conciseness, and coherence are the criteria used to calculate rewards for each response [45]. Each response's estimated relative advantage is as follows:

G

,
$$A_{i\,t} = r(o_i) - \frac{1}{G}\sum_{j=1} r(o_j)$$

where G is the number of responses in the group and $r(o_i)$ is the reward for response i. Without depending on conventional value functions, this estimation improves replies that perform better than average [20]. The policy is updated in accordance with the modified GRPO objective, which adds a Kullback-Leibler (KL) divergence penalty to the PPO objective in order to guarantee stable learning:

$$J_{\mathrm{GRPO}}(\theta) = \mathrm{E}_{q,\{o_i\}}\left\{\frac{1}{G}\sum_{i=1}\frac{1}{|o_i|}\sum_{t=1}\min\left[r_t A_{i,t},\ \mathrm{clip}(r_t,\,1-\epsilon,\,1+\epsilon)A_{i,t}\right] - \beta D_{\mathrm{KL}}\left[\pi_\theta \,\|\, \pi_{\mathrm{ref}}\right]\right\}$$

where:

$$r_t = \frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\mathrm{old}}}(o_t \mid q, o_{<t})}$$

and the KL divergence between the reference and current policies, controlled by the coefficient β [28], is represented by D_{KL}. Furthermore, a partial reward model is used to calculate token-level payouts as follows:

$$r_t = \frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\mathrm{old}}}(o_t \mid q, o_{<t})} \cdot \beta \log r_\phi(q, o_{<t})$$

where the output of the partial reward model that directs generation is shown by r_ϕ [28]. According to recent research, GRPO greatly increases the accuracy of classifying fake news, reduces output length without sacrificing semantic integrity, and strengthens reasoning in big language models with less dependence on manually labeled data.

5. Experiments and Analysis

5.1. Experimental Setup and Datasets

Because generative language models generate open-ended, unstructured outputs rather than predetermined categorization labels, evaluating them presents special difficulties. This trait calls for specific assessment techniques that are able to decipher and extract significant labels from unstructured material. In light of this, we suggest an assessment methodology to evaluate the false news detection task's classification accuracy and response conciseness both before and after using the Group Relative Policy Optimization (GRPO) algorithm [49]. A strong classification

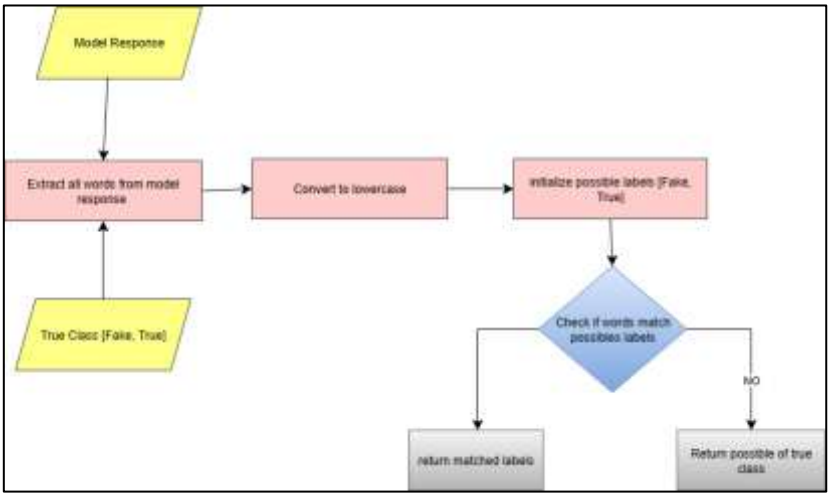

Figure 5: Workflow for extracting predicted class from a model's textual response.

extraction function was created to consistently translate model answers to binary labels (i.e., "real" or "fake") in spite of the unstructured nature of generative outputs. We assessed the model using four main metrics:

- **Accuracy:** Measures the overall percentage of correctly classified news items:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is true positives (correctly detected fake news), TN is true negatives (correctly detected real news), and FP, FN are false positives and false negatives, respectively. This metric provides a broad performance overview but may not reflect issues in imbalanced datasets [50].

- **Precision:** Captures the ability to avoid mislabeling real news as fake:

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision indicates strong reliability in identifying fake content.

- **Recall:** evaluates how sensitive the model is to real fake news:

$$\text{Recall} = \frac{TP}{TP + FN}$$

This is essential for avoiding the omission of false information in dangerous situations [50].

- **F1-Score:**gives the harmonic mean of recall and precision:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In real-world situations when there is a class imbalance, the F1-score is very significant [50].

Using the **Single-Word Accuracy** metric, we assess the model's propensity to produce succinct and machine-readable responses:

$$\text{Single-Word Accuracy} = \frac{\text{Number of 1-word responses}}{\text{Total samples}}$$

This guarantees that the model generates condensed outputs (such as "Real" or "Fake"), which are necessary for automatic classification systems downstream with constrained processing power.

Classification Extraction Algorithm:

Algorithm 3 ExtractFakeNewsClassification **Require:** Model response *R*, True label *T* **Ensure:** Classification label: real or fake
1: Convert *R* to lowercase
2: Extract words from *R* using regex **for** *word in reversed(words)* **do**
3: └
then
4: └
match ← get close matches(word, [real, fake], cutoff=0.7) **if** *match is not empty*

return match[0] **else**
T = real
5: **return** fake **else**
6: └
return real
7:
8:
9:

This assessment pipeline supports trustworthy fake news detection from generated answers and is in line with current best practices for open-ended language model assessment.
Results of the Overall Comparison We tested each model independently before and after using the Group Relative Policy Optimization (GRPO) technique to assess how well the Qwen2.5-3B- Instruct and Llama 3.2 models performed on a balanced dataset.

5.2. Qwen2.5-3B-Instruct Model
Before GRPO. The duration of the instruction was roughly forty-three minutes. With an overall F1-score of 73.02% and an accuracy of 73.33%, the model demonstrated class disparity.

Table 1: Qwen2.5-3B-Instruct Performance Metrics Before GRPO

max width=

Class	Precision	Recall	F1-Score	Support
Fake	0.81	0.62	0.70	500
Real	0.69	0.85	0.76	490

After GRPO. The training period was shortened to 12 minutes. Precision and recall were balanced, and accuracy increased to 87.07%.

Table 2: Qwen2.5-3B-Instruct Performance Metrics After GRPO

max width=

Class	Precision	Recall	F1-Score	Support
Fake	0.86	0.88	0.87	500
Real	0.88	0.86	0.87	490

Table 3: Confusion Matrix Counts for Qwen2.5-3B-Instruct

max width=

	TN	FN	FP	TP
Before GRPO	310	190	74	416
After GRPO	442	58	70	420

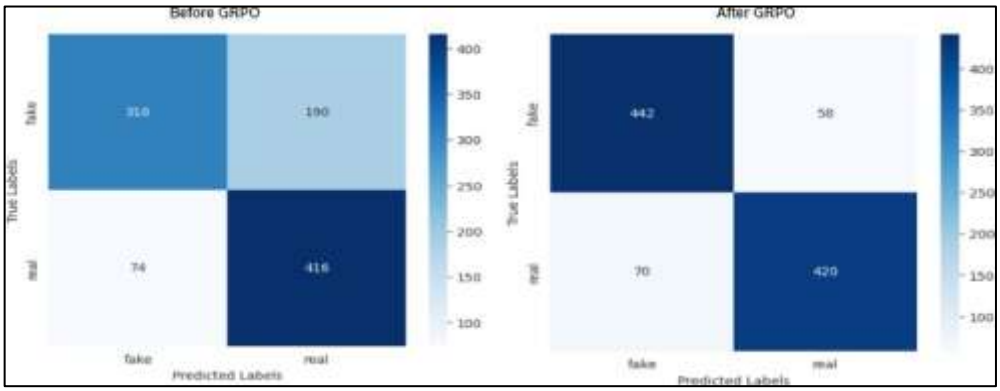

Figure 6: Confusion Matrix for Qwen2.5-3B-Instruct model

ROC Curve. The AUC was 73.45% prior to GRPO, which indicated a significant false positive rate and

poor class discrimination. Following GRPO, AUC increased to 87.06%, indicating a better balance between the frequencies of false positives and true positives.

Precision-Recall Curve. The PR curve's AUC rose from 80.96% prior to GRPO to 90.29% following it, indicating improved effectiveness in striking a balance between recall and precision in the face of class imbalance. Although baseline methods such as Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) are widely used in reinforcement learning from human feedback (RLHF) studies, they were intentionally excluded from this experiment due to both methodological and practical considerations. The base model, **Qwen2.5-3B-Instruct**, has already undergone comprehensive multi-stage instruction tuning and supervised fine-tuning, as described in the Qwen2.5 technical report [51]. These stages include alignment and preference optimization procedures that are functionally similar to SFT, rendering an additional SFT phase redundant and potentially confounding when isolating the contribution of GRPO.

Moreover, conducting large-scale SFT and DPO baselines would require substantial com- putational resources, curated instruction or preference datasets, and extensive hyperparameter tuning, which are beyond the feasible scope of this study. Similar limitations have been discussed in prior works [41, 42], where authors focused on evaluating novel alignment algorithms within constrained computational settings rather than duplicating prior baselines.

Accordingly, this study emphasizes isolating and analyzing the impact of GRPO under realistic computational constraints, while deferring extended ablation studies and additional baseline comparisons to future work [52].

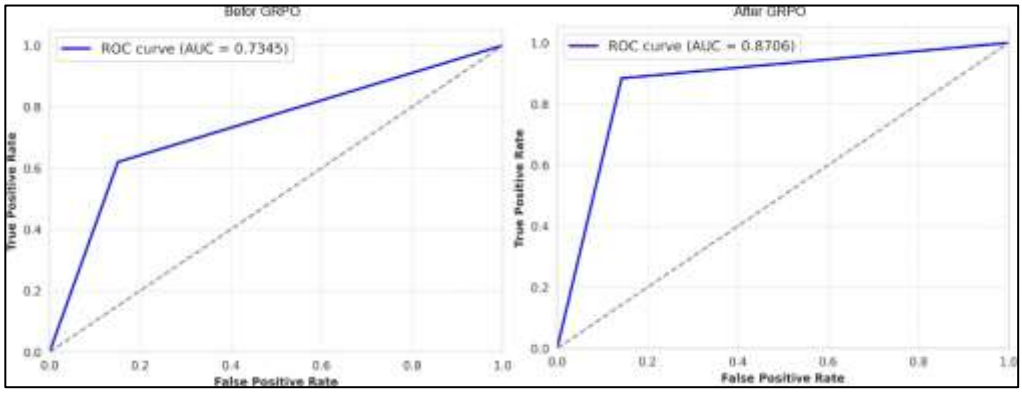

Figure 7: ROC Curve for Qwen2.5-3B-Instruct model

5.3. Llama 3.2 Model

Before GRPO. The model before Training achieved an accuracy of 59% and an F1-score of 55%.

Table 4: Llama 3.2 Performance Metrics Before GRPO

max width=

Class	Precision	Recall	F1-Score	Support
Fake	0.73	0.29	0.42	500
Real	0.55	0.89	0.68	490

After GRPO. Training Takes over 1 hour, the Accuracy after training increased to 93.03% with metrics over 93% when GRPO was used.

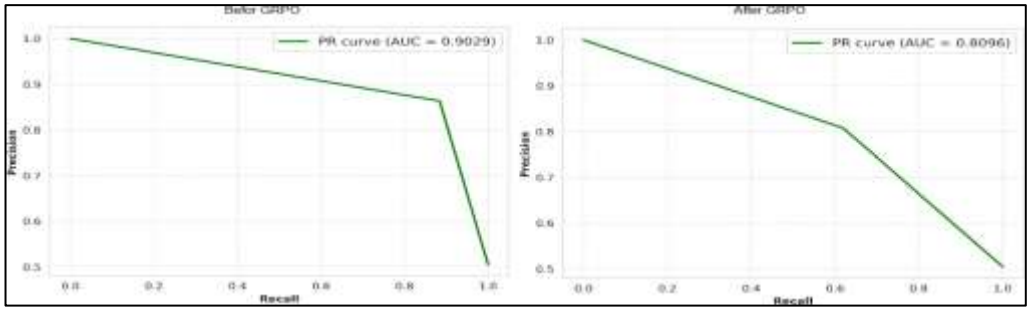

figure 8: Precision-Recall Curve for Qwen2.5-3B-Instruct model

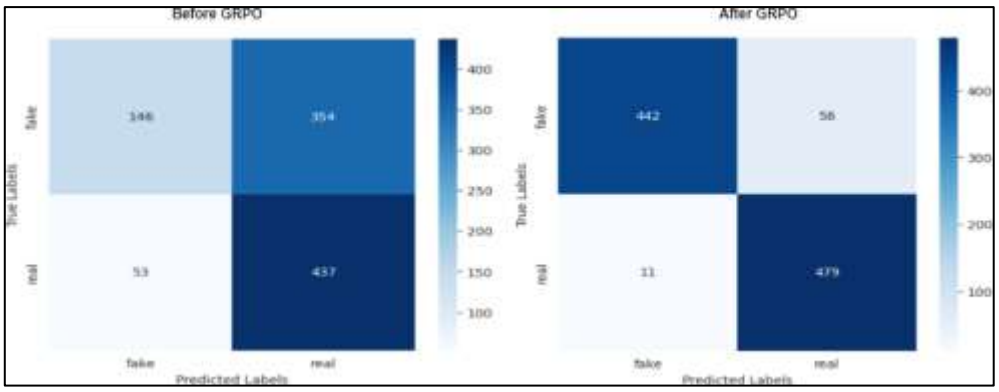

Figure 9: Confusion Matrix for Llama 3.2 model

Table 5: Llama 3.2 Performance Metrics After GRPO

max width=

Class	Precision	Recall	F1-Score	Support
Fake	0.98	0.88	0.93	500
Real	0.89	0.98	0.93	490

Table 6: Confusion Matrix Counts for Llama 3.2

max width=

	TN	FN	FP	TP
Before GRPO	437	354	53	146
After GRPO	479	58	11	442

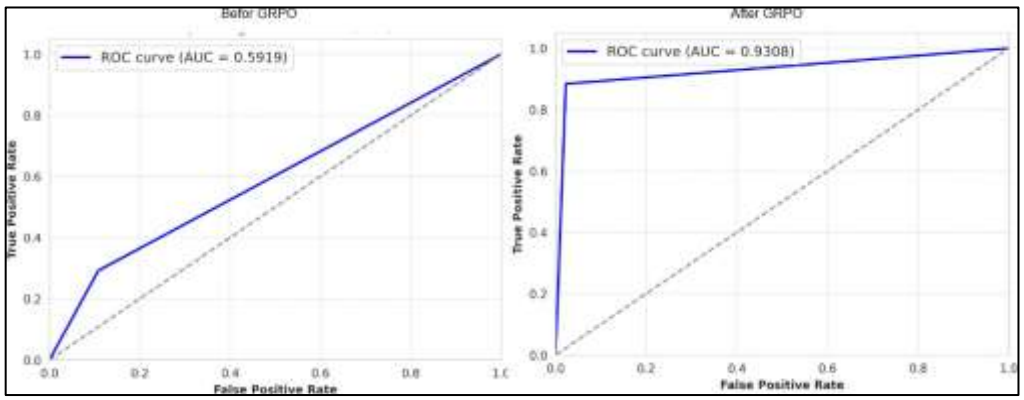

Figure 10: ROC Curve for Llama 3.2 model

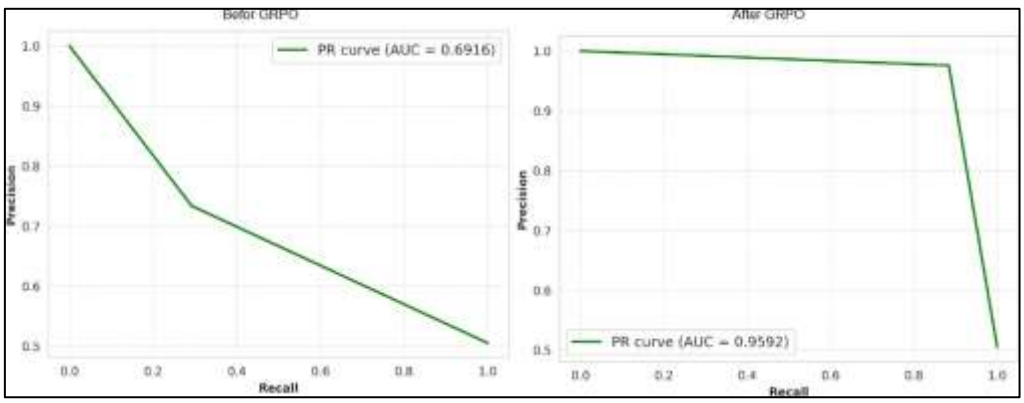

Figure 11: Precision-Recall Curve for Llama 3.2 model

ROC Curve. With significantly enhanced class discrimination, the AUC increased from 59.19% prior to GRPO (almost random) to 93.08% following GRPO.
Precision-Recall Curve. The improved predictive balance and efficiency were reflected in the PR AUC, which rose from 69.16% to 95.92%.

5.4. Results Analysis

Prior to the use of GRPO, both models frequently produced verbose or indirect answers as opposed to succinct labels like "Fake" or "Real," which made classification less clear and difficult to evaluate it. GRPO training greatly increased label production and efficiency by using reward signals for accuracy, conciseness, and clarity. Figure 12 demonstrates a loss curve that gradually drops during GRPO training, signifying constant optimization and enhanced model performance. Figure 13 demonstrates how the models can balance producing output that is both accurate and succinct by showing the sub-reward functions (correctness and length of words) converge toward high values. There were certain trade-offs, too, as responses that needed specificity to be accurate frequently ran counter to the brevity goal.

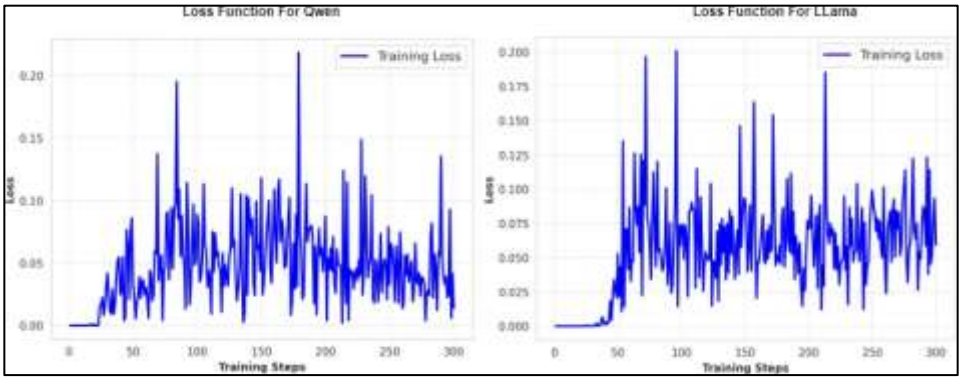

Figure 12: Loss Curve During GRPO Training

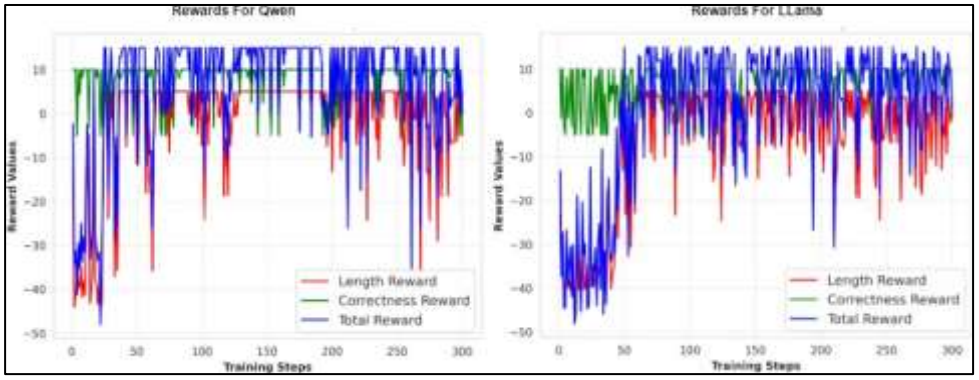

Figure 13: Reward Function Convergence: Correctness and Brevity

5.5. ***Discussion***

The experimental findings unequivocally show how well the Group Relative Policy Opti- mization (GRPO) technique works to improve the performance of big language models for the classification job of fake news. After adding GRPO during fine-tuning, the Qwen2.5-3B-Instruct and Llama 3.2 models showed notable gains in all key measures, such as accuracy, precision, recall, and F1-score. The Qwen2.5-3B-Instruct model's inference time (on test dataset) was cut from 43 minutes to just 12 minutes, demonstrating that GRPO not only enhances model perfor- mance but also speeds up convergence during training. This represents a significant efficiency gain, particularly in real-world deployment scenarios where time and computational resources are scarce.

Additionally, even when there is a class imbalance, GRPO can improve the models' capacity to distinguish between fake and real news, as evidenced by the significant gains in Area Under the Curve (AUC) values for both the ROC and Precision-Recall curves for both models. In applications where false positives and false negatives have distinct risks and expenses, this is crucial. According to confusion matrices, GRPO improves the accuracy of misleading content identification by drastically lowering classification mistakes, especially false negatives in the fake news category. The enhanced class balance indicates that common biases in unbalanced classification problems are successfully addressed by the relative reward system used by GRPO. Response formulation analyzes conducted both before and after GRPO training validate the method's capacity to promote clear and succinct categorization outputs as opposed to long or unclear responses. In downstream applications where decision interpretability is critical, such as automated news verification systems, this kind of clarity is vital. The reward function fluctua- tions throughout training, however, show that striking the ideal balance between accuracy and brevity is still a continuous difficulty. This suggests that future research could examine adaptive reward schemes that dynamically modify these factors in response to contextual demands. The Llama 3.2 model was notably more responsive and showed more

noticeable gains after GRPO application than the Qwen2.5-3B-Instruct model. Higher increases were made by Llama in all performance parameters, particularly in precision, recall, and F1-score, which demonstrated an improved capacity to discern between bogus and authentic news. In the confusion matrix, Llama's error reduction was more pronounced, suggesting that it was more successful at identifying false information. LLama is better than Qwen when utilizing the Group Relative Policy Optimization approach for intricate tasks like classifying bogus news. The potential advantages of combining complex reinforcement learning techniques with advanced architectures to boost classification
accuracy and response quality are thus highlighted by Llama's enhanced performance.
GRPO is a successful policy optimization technique that improves the efficiency and accuracy of generative language models in fake news detection tasks. This allows models to capture subtle differences between classes while preserving the clarity and conciseness of responses, opening the door for automated verification systems that are more dependable and scalable.

6. Conclusion and Future Work

This study examines how GRPO can be used to improve the QWEN and LLAMA models for binary classification tasks, with a particular emphasis on the problem of fake news identification. Our results show notable increases in model processing speeds and classification accuracy. Our ability to train both models in record time is largely due to the integration of LoRA technology while GRPO helped with output refinement and effective process management. Since the po- of GRPO goes far beyond the current results, future study could concentrate on handling imbalanced datasets or further refining the models by tweaking them with more rigorous reward functions.
This work demonstrates the potential of GRPO as a method for optimizing generative processes in natural language processing, paving the way for its use in a variety of fields.

Author Contributions
Ali Salloum conceived the research concept, carried out an extensive literature review, an- alyzed pertinent studies, and developed the study's methodology. Ali Salloum developed the theoretical framework and contributed substantially to manuscript writing. Ali Salloum also carried out the software implementation and coding process together with Ebrahim Massrie.
Ebrahim Massrie directed the technical development aspects, including coding, implementa- tion, and figure and visualization production, in addition to providing assistance in the manuscript drafting process.
Basel ALKHATIB was led by Base, offering advisory insights and guidance throughout the research process, helped in the refinement of the methodology, and edited and reviewed the manuscript for scholarly precision and clarity. All authors reviewed and finalized the manuscript.

Acknowledgements
The cooperative efforts and direction of our academic mentors and colleagues, whose input was vital in determining the course of this study, enabled us to complete this task. We express our profound appreciation to the Qwen and LLaMA model creators and open-source contributors, as well as the GRPO implementation resources, which greatly aided our testing and assessment procedures.
We also thank the computational infrastructure that made it possible to train and refine large- scale models, as well as the reviewers and peers who offered valuable comments that improved the study's rigor and clarity.

Author Biography
Ali Salloum is a graduate student at the Syrian Virtual University. He is a researcher and software engineer who is interested in Natural Language Processing (NLP). His research explores creating more effective AI methods for false information detection and language understanding. Ali has devoted a lot of time to designing, developing techniques, and programming for this research.
Basel Alkhatib received the Doctorate in Computer Engineering and DEA degree in Infor- matics from the University of Bordeaux, France. He received the Bachelor's degree in Computer Engineering from the Higher Institute for Applied Sciences and Technology (HIASST) attached to the Scientific Studies and Research Center, Damascus, Syria. His areas of interest are artificial intelligence, natural language processing, and software engineering. Dr. Alkhatib has directed and participated actively in many research projects in the areas of intelligent systems and applied informatics.
Ebrahim Massrie is an Artificial Intelligence researcher and engineer with expertise in Natura
Language Processing. He offers solutions in AI and specializes in making machine learning models functional. Ebrahim made extensive contributions to software development, coding, and visualization for this project, in addition to helping with manuscript preparation.

References

[1] O. Bashaddadh, N. Omar, M. Mohd, and M. N. A. Khalid, "Machine learning and deep learning approaches for fake news detection: A systematic review of techniques, challenges, and

advancements," *IEEE Access*, 2025.
[2] E. Denniss and R. Lindberg, "Social media and the spread of misinformation: infectious and a threat to public health," *Health Promotion International*, vol. 40, no. 2, p. daaf023, 2025.
[3] B. Wang, J. Ma, H. Lin, Z. Yang, R. Yang, Y. Tian, and Y. Chang, "Explainable fake news detection with large language model via defense among competing wisdom," in *Proceedings of the ACM Web Conference 2024*, pp. 2452– 2463, 2024.
[4] E.-M. Event-Separated, "Evolving to the future: Unseen event adaptive fake news detection on social media,"
[5] A. E. Qasem and M. Sajid, "Exploring the effect of n-grams with bow and tf-idf representations on detecting fake news," in *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, pp. 741–746, IEEE, 2022.
[6] S. Kuntur, A. WrAˇłblewska, M. Paprzycki, and M. Ganzha, "Fake news detection: It's all in the data!," *arXiv*
preprint arXiv:2407.02122, 2024.
[7] H. Liu, W. Wang, H. Li, and H. Li, "Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection," *arXiv preprint arXiv:2402.07776*, 2024.
[8] K. Venkatachalam, B. B. Al-Onazi, V. Simic, E. B. Tirkolaee, and C. Jana, "Deepfnd: an ensemble-based deep learning approach for the optimization and improvement of fake news detection in digital platform," *PeerJ Computer Science*, vol. 9, p. e1666, 2023.
[9] I. Ahmed, S. Islam, P. P. Datta, I. Kabir, N. U. R. Chowdhury, and A. Haque, "Qwen 2.5: A comprehensive review of the leading resource-efficient llm with potentioal to surpass all competitors," *Authorea Preprints*.
[10] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
[11] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten,
A. Vaughan, *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.
[12] S. Hasan and S. Basak, "Open-source ai-powered optimization in scalene: Advancing python performance profiling with deepseek-r1 and llama 3.2," *arXiv preprint arXiv:2502.10299*, 2025.
[13] Y. Rahulamathavan, "Demystifying group relative policy optimization (grpo): A toy example using mnist classifica- tion,"
[14] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, *et al.*, "Deepseek llm: Scaling open-source language models with longtermism. arxiv 2024," *arXiv preprint arXiv:2401.02954*.
[15] D. Guo, Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, G. Chen, X. Bi, Y. Wu, Y. Li, *et al.*, "Deepseek-coder: When the large language model meets programming–the rise of code intelligence, 2024," *URL https://arxiv. org/abs/2401.14196*, vol. 5, p. 19, 2024.
[16] Z. Wang, B. Bi, S. K. Pentyala, K. Ramnath, S. Chaudhuri, S. Mehrotra, X.-B. Mao, S. Asur, *et al.*, "A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more," *arXiv preprint arXiv:2407.16216*, 2024.
[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, p. 3, 2022.
[18] D. Lee, S. Choi, and I. J. Chang, "Qrazor: Reliable and effortless 4-bit llm quantization by significant data razoring,"
arXiv preprint arXiv:2501.13331, 2025.
[19] K. Behdin, Y. Dai, A. Fatahibaarzi, A. Gupta, Q. Song, S. Tang, H. Sang, G. Dexter, S. Zhu, S. Zhu, *et al.*, "Efficient ai in practice: Training and deployment of efficient llms for industry applications," *arXiv preprint arXiv:2502.14305*, 2025.
[20] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024," *URL https://arxiv. org/abs/2402.03300*, 2024.
[21] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective,"
ACM SIGKDD explorations newsletter, vol. 19, no. 1, pp. 22–36, 2017.
[22] B. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, pp. 759–766, 2017.
[23] J. Alghamdi, S. Luo, and Y. Lin, "A comprehensive survey on machine learning approaches for fake news detection,"
Multimedia Tools and Applications, vol. 83, no. 17, pp. 51009–51067, 2024.
[24] S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "A review of fake news detection models: Highlighting the factors affecting model performance and the prominent techniques used," *International Journal of Advanced Computer Science And Applications*, vol. 14, no. 7, 2023.
[25] L. Huang, "Deep learning for fake news detection: Theories and models," in *Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering*, pp. 1322–1326, 2022.
[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional

transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
[27] Y. Mroueh *et al.*, "Revisiting group relative policy optimization: Insights into on-policy and off-policy training,"
arXiv preprint arXiv:2505.22257, 2025.
[28] N. I. Alonso *et al.*, "The mathematics of group relative policy optimization: A multi-agent reinforcement learning approach," *The Mathematics of Group Relative Policy Optimization: A Multi-Agent Reinforcement Learning Approach (January 03, 2025)*, 2025.
[29] T. Jiang, H. Wang, and C. Yuan, "Diffora: Enabling parameter-efficient llm fine-tuning via differential low-rank matrix adaptation," *arXiv preprint arXiv:2502.08905*, 2025.
[30] Y. Mroueh, "Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplifica- tion," *arXiv preprint arXiv:2503.06639*, 2025.
[31] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Re´, "Flashattention: Fast and memory-efficient exact attention with io-awareness," *Advances in neural information processing systems*, vol. 35, pp. 16344–16359, 2022.
[32] Y. Xu, L. Xie, X. Gu, X. Chen, H. Chang, H. Zhang, Z. Chen, X. Zhang, and Q. Tian, "Qa-lora: Quantization-aware low-rank adaptation of large language models," *arXiv preprint arXiv:2309.14717*, 2023.
[33] H. Ahmed, I. Traore, and S. Saad, "Machine learning and deep learning approaches for fake news detection: A systematic review of techniques, challenges, and advancements," *IEEE Access*, vol. 10, pp. 108369–108390, 2022.
[34] A. Hassan, X. Chen, and Q. Li, "Mcred: Multi-modal message credibility for fake news detection using bert and cnn," *Scientific Reports*, vol. 12, no. 1, p. 12345, 2022.
[35] R. Patil, A. Kumar, and P. Singh, "A shap-based xai approach to evaluating machine learning classification of fake news," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, pp. 234–245, 2023.
[36] K. Song, A. Moeini, P. Wang, L. Gong, R. Chandra, Y. Qi, and S. Zhang, "Reward is enough: Llms are in-context reinforcement learners," *arXiv preprint arXiv:2506.06303*, 2025.
[37] Y. Yan, X. Lou, J. Li, Y. Zhang, J. Xie, C. Yu, Y. Wang, D. Yan, and Y. Shen, "Reward-robust rlhf in llms," *arXiv preprint arXiv:2409.15360*, 2024.
[38] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, *et al.*, "Deepseek-vl: towards real-world vision-language understanding," *arXiv preprint arXiv:2403.05525*, 2024.
[39] J. Gao, S. Xu, W. Ye, W. Liu, C. He, W. Fu, Z. Mei, G. Wang, and Y. Wu, "On designing effective rl reward at training time for llm reasoning," *arXiv preprint arXiv:2410.15115*, 2024.
[40] Z. Shao *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.
[41] R. Rafailov *et al.*, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, 2023.
[42] S. Xie *et al.*, "Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf,"
arXiv preprint arXiv:2405.21046, 2024.
[43] W. Zhang, C. Li, R. Kumar, *et al.*, "Task-circuit quantization: Leveraging knowledge localization for compression,"
arXiv preprint arXiv:2504.07389, 2025.
[44] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "Llm.int8(): 8-bit matrix multiplication for transformers at scale," *arXiv preprint arXiv:2208.07339*, 2022.
[45] N. I. Alonso, R. Pereira Franklin, *et al.*, "The mathematics of dapo, ppo, and grpo algorithms," *Rodolfo, The Mathematics of DAPO, PPO, and GRPO Algorithms (April 04, 2025)*, 2025.
[46] Y. Zhou, Y. Wu, and J. Deng, "Hybrid group relative policy optimization: A multi-sample approach to enhancing policy optimization," *arXiv preprint arXiv:2502.01652*, 2025.
[47] J. Liu, A. Wu, and M. Lin, "Optimizing safe and aligned language generation: A multi-objective grpo approach,"
arXiv preprint arXiv:2503.21819, 2025.
[48] Z. Li, F. Zhang, and M. Gao, "Delving into rl for image generation with cot: A study on dpo vs. grpo," *arXiv preprint arXiv:2505.17017*, 2025.
[49] L. Feng, Z. Xue, T. Liu, and B. An, "Group-in-group policy optimization for llm agent training," *arXiv preprint arXiv:2505.10978*, 2025.
[50] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in *Computer science on-line conference*, pp. 15–25, Springer, 2023.
[51] Q. Team, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.
[52] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, 2020.