Journal Homepage: www. ijarpr.com ISSN: 3049-0103 (Online)

International Journal of Advance Research Publication and Reviews

Vol 02, Issue 09, pp 703-713, September 2025

Enhancing Linguistic Models with Reinforcement Learning for Text Summarization

Mrinaal M S[1], Rohan M R[2]

[1]Department of Computer Science and Engineering, BNM Institute of Technology, Affiliated to VTU, Bangalore, India mrinaalshivakumar6@gmail.com
[2]Department of Computer Science and Engineering, BNM Institute of Technology, Affiliated to VTU, Bangalore, India rohanmr551@gmail.com
[3]Department of Computer Science and Engineering, BNM Institute of Technology, Affiliated to VTU, Bangalore, India akshithakatkeri@bnmit.in

ABSTRACT—

One of the hardest problems in Natural Language Processing (NLP) is still automatically summarising texts. Despite their impressive success in abstractive summarisation, transformer-based models frequently fall short of human preferences for coherence, factual accuracy, and conciseness. The use of Reinforcement Learning (RL), more especially Reinforcement Learning from Human Feedback (RLHF), to optimise large-scale language models for summarisation is examined in this work. We suggest a framework for reward-driven optimisation that enhances summary quality in a number of areas, including factual consistency, informativeness, and fluency. In comparison to supervised fine-tuning alone, our experiments, which are based on the LLaMA 3.2 model and well-known datasets like CNN/DailyMail and XSum, demonstrate notable gains in both automatic and human evaluation metrics.

Keywords—LLaMa 3.2, Summarization, NLP, T5, BERT Embedding.

Introduction

The ability to automatically create summaries from lengthy texts has become crucial due to the exponential growth in digital content. Users can swiftly assimilate crucial information from news articles, scholarly papers, or legal documents with the aid of automatic summarisation tools. Despite their successes, the majority of neural models rely on supervised learning goals like Maximum Likelihood Estimation (MLE), which prioritise token-level probability maximisation over higher-level goals like coherence and user utility. However, this approach frequently leads to issues like exposure bias and inconsistent outputs, particularly in longer sequences. To get around these challenges, researchers have looked into reinforcement learning techniques that maximise global sequence-level objectives, raising the overall standard and reliability of the summaries that are generated.

A potent paradigm for overcoming these constraints is Reinforcement Learning (RL), especially Reinforcement Learning from Human Feedback (RLHF). Models are trained to maximise long-term rewards that more closely resemble human assessments rather than just forecasting the next token. With an emphasis on text summarisation, this study investigates how RLHF can improve linguistic models. Using a custom reward model trained on human preferences, we optimise LLaMA 3.2. According to our research, RLHF raises human satisfaction in addition to objective metrics like ROUGE and BERTScore.

This paper presents an approach that leverages LLaMA 3.2, a cutting-edge open-source large language model, to enhance text summarization through reinforcement learning. Unlike earlier models, LLaMA 3.2 demonstrates improved token efficiency and advanced long-context reasoning, making it highly suitable for complex natural language processing tasks. Our methodology outlines the development of an intelligent summarization system that adapts dynamically to diverse textual inputs and user-defined summary goals. By integrating reinforcement learning, we aim to refine summary coherence, relevance, and fluency through continuous feedback loops.

The primary objective of this research is to investigate how advanced large language models—specifically LLaMA 3.2—can be effectively utilized to enhance code completion and real-time bug detection for software developers. This study is structured around three core goals. First, we evaluate LLaMA 3.2's ability to deliver accurate, context-sensitive code completions across multiple programming languages and integrated development environments (IDEs). Second, we leverage its advanced natural language understanding and reasoning capabilities to identify syntactic, logical, and semantic errors during the coding process, thereby reducing dependence on traditional post-execution debugging tools. Third, we benchmark our system against leading AI-assisted development tools, measuring performance across key metrics such as completion accuracy, bug detection precision and recall, system latency, and developer usability. Collectively, these efforts aim to build an intelligent, real-time coding assistant that improves development efficiency, accelerates error resolution, and minimizes cognitive load on developers. Fig. 1. Shows the interactive diagram of text summarization process.

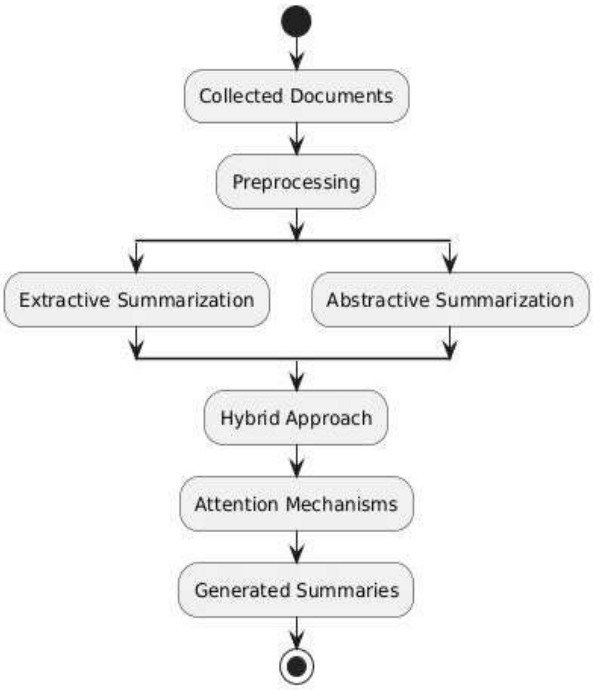

Fig. 1. Interactive Diagram of text summarization process

LITERATURE SURVEY

Abadi and Ghasemian (2025) present a novel framework for Persian text summarization by integrating a three-phase fine-tuning strategy with reinforcement learning on the mT5 transformer model. [1] Their method sequentially fine-tunes the encoder, decoder, and bias parameters, enabling the model to better capture linguistic nuances in Persian text. Reinforcement learning using the Proximal Policy Optimization (PPO) algorithm is then applied to optimize summary quality based on the ROUGE-1 metric. This dual-stage process significantly improves summarization performance, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 53.17, 37.12, and 44.13, respectively—outperforming benchmarks like BERT2BERT and ARMAN. The study highlights the effectiveness of combining structured model training with feedback-based optimization and sets a precedent for summarization in low-resource or complex languages.

A reliable automated summarization system that uses large language models (LLMs) to distill numerous document abstracts and titles into succinct, logical summaries is put forth by Langston and Ashford [2]. Their method combines abstractive and extractive summarization strategies in a hybrid framework that is improved by sophisticated attention mechanisms. With competitive ROUGE, BLEU, and METEOR scores, the system showed excellent performance, especially in technical and scientific fields like healthcare. Thorough preprocessing, careful data collection from various sources, and refinement of transformer-based models for summarization tasks were all part of the methodology. Notably, in addition to conventional quantitative scores, the study proposed new qualitative metrics to assess coherence, readability, and informativeness. While the system performed exceptionally well at summarizing shorter documents, it had trouble preserving logical transitions and fluency in longer texts.

In order to greatly enhance summarization quality, Liu et al. suggest a two-phase method for web summarization that combines extensive pre-training with a novel, refined tuning strategy [3]. To impart fundamental linguistic knowledge, their approach makes use of a sequence-to-sequence (seq2seq) deep learning framework that was first pre-trained on a sizable, varied out-of-domain corpus. A key innovation in this work is the refined tuning phase, which presents a regularization mechanism based on smoothing that reduces prediction overconfidence and improves cross-domain generalization. In order to assess model performance across a range of document lengths and complexity, the authors also assembled a varied dataset of online news articles. Their method consistently outperforms state-of-the-art models such as BART, PEGASUS, GPT-3.5, and LLaMA2 across ROUGE metrics, according to empirical evaluations on both standard (CNN/DailyMail) and custom test sets. Furthermore, ablation studies confirm that obtaining superior summarization performance requires both pre-training and fine-tuning.

METHODOLOGY

The summarization system was developed using a thorough methodology intended to guarantee the production of succinct and logical summaries from a variety of document titles and abstracts. The procedures and methods used at each step of the methodology are described in detail in the ensuing subsections.

Data Collection

1. Public datasets: Used to increase reproducibility and accessibility.
2. Digital libraries: Provided well-organized, superior scholarly materials.
3. Domain repositories: Offered practical diversity in various fields and industries.

Document Types

1. Scholarly articles: Chosen for their complex, specialized language.
2. Reports from corporations and the government are used for organized, policy-driven content.
3. Featured for narrative and conversational tone diversity are news and web articles.

Selection Criteria

1. Covered a range of topics, including technology, healthcare, finance, law, and education.
2. Length variation: To prevent overfitting, a balance was struck between brief news summaries and in-depth technical reports.
3. Lexical diversity: Focused on a wide variety of writing styles, linguistic complexities, and terminologies.

Data Composition

1. Extracted abstracts and titles: Served as ground truth pairs for summarization tasks.

2. Balanced sampling: Ensured proportional representation of simple and complex text structures.

Data Augmentation and Preprocessing Techniques

The dataset was heavily augmented and normalized to increase robustness. Among the methods used were noise injection, paraphrasing, sentence rearranging (for coherence training), and synonym substitution. To standardize the input for LLaMA 3.2, preprocessing techniques included tokenization, lowercasing, punctuation filtering, and sentence boundary detection. These procedures guaranteed that the model could summarize in both extractive and abstractive modes and better generalize to unknown text types.

Ethical Considerations

Ethical filtering was used during data curation to prevent biased, deceptive, or private content because some domains—like healthcare, law, and finance—are sensitive. The components of reinforcement learning were set up to reward inclusive, neutral summaries, guarantee factuality, and penalize hallucinations. The model's behavior was carefully matched with responsible AI principles, particularly in cases where the summarization output could affect public perception or decision-making.

In addition to enabling LLaMA 3.2 to learn hierarchical text structures and contextual dependencies more effectively, this varied and calibrated dataset served as the perfect training environment for reinforcement learning. The model was iteratively optimized to produce logical, educational, and domain-adaptive summaries by mimicking human-like feedback and reward systems. This is essential for high-stakes applications such as policy summarization, medical briefings, and legal document abstraction.

Table I shows the improvement methods for text summarization driven reinforcement learning with Llama 3.2.

Improvement Methods for Text Summarization Driven by Reinforcement Learning with LLaMA 3.2

Technique	***Description***	***Advantages***	***Challenges***
RL-based Fine-Tuning (PPO)	Applies Proximal Policy Optimization to LLaMA 3.2, using reward signals (e.g., ROUGE, coherence scores) to iteratively refine summary generation.	Enhances the fluency and relevance of summaries. decreases hallucinations through feedback loops	Reward-design sensitive. High memory and processing costs
Long-Context Summarization	Leverages extended token windows or sliding-window attention to process and condense long documents into coherent summaries.	Records document structure and global context. guarantees thorough discussion of the important points.	Higher GPU memory usage and latency Possible dilution of context over lengthy inputs
Domain-Specific Fine-Tuning	LLaMA 3.2 is further trained on specialized corpora (such as medical, legal, and finance) to	Improves target domains' terminology fidelity and accuracy increases user confidence	Needs large amounts of high-quality domain data. Danger of overfitting to a limited vocabulary

Technique	***Description***	***Advantages***	***Challenges***
	modify terminology and summarization style.		
Semantic Coherence Reward Modeling	Introduces auxiliary reward functions that, during RL training, explicitly score discourse structure, factual consistency, and logical flow.	Improves logical transitions and readability promotes summaries that are based on facts.	Multiple rewards are difficult to design and calibrate. Risk of instability due to competing goals

THREE PHASED FINE TUNING

Our suggested approach uses a three-phase, nine-epoch fine-tuning regimen to LLaMA 3.2, combining reinforcement learning and supervised learning to generate extremely coherent, context-aware summaries:

Phase 1: Supervised Cross-Entropy Fine-Tuning (3 epochs)

First, we use a cross-entropy loss and standard teacher-forcing to fine-tune the full LLaMA 3.2 model on paired article–summary data. This phase instills strong baseline fluency and factual alignment in the model by teaching it the fundamental mapping from input text to gold-standard summaries.

Phase 2: PPO-Based Reinforcement Learning (3 epochs)

The top layers of the transformer are then subjected to Proximal Policy Optimization (PPO), while the lower 80% of the layers are frozen. ROUGE-L, a factuality penalty, and a learned coherence score are all combined in reward signals. In order to minimize hallucinations and increase relevance, this reinforcement stage improves the model's policy to favor summaries that perform well on both automatic metrics and our learned quality predictors.

Phase 3: Bias & Reward ‑ Weight Calibration (3 epochs)

Lastly, we only adjust the scalar weights for each reward component and the output bias terms of the model. We calibrate summary length, style, and metric-tradeoffs to human preferences by varying these small parameter sets, which ensures a consistent tone and avoids over-optimization on any one reward.

Together, these three focused phases systematically elevate LLaMA 3.2's summarization prowess—first grounding it in supervised learning, then steering it via human-inspired rewards, and finally calibrating its output behavior for reliable, high-quality text summaries.

Reinforcement learning

We add a reinforcement-learning phase powered by Proximal Policy Optimization (PPO) to further improve LLaMA 3.2's summarization capabilities. Beyond supervised fine-tuning, this provides an additional layer of adaptive feedback that guides the model to produce summaries with high relevance, coherence, and factual consistency scores.

The Role of Reinforcement Learning

LLaMA 3.2 is treated as an agent interacting with a summarization "environment" in reinforcement learning. At every stage, the model makes a summary suggestion, gets a reward signal that indicates how good it is, and modifies its policy to optimize the total reward. In actual use, this means that the model comes to favor summary outputs that are consistent with both automatic metrics and human judgment, resulting in more organic and educational text.

Proximal Policy Optimization (PPO)

We employ PPO—a stable, sample-efficient algorithm well suited to high-dimensional action spaces—to fine-tune LLaMA 3.2's higher transformer layers. PPO constrains policy updates so that each gradient step stays within a trust region, preventing the large, unstable swings that can occur in summary-generation tasks. By optimizing a composite reward (e.g., ROUGE-L, coherence score, factuality penalty), PPO refines the model's token-by-token decision process toward summaries that excel on all fronts.

Reinforcement Learning Environment

LLaMA 3.2 is "plugged into" an RL training loop following supervised cross-entropy fine-tuning. To coordinate distributed training across GPUs, we utilize the TRLX framework, which was created for Transformer-based RL fine-tuning. With TRLX, we can easily switch between a manually created reward function and a limited number of summaries that have been evaluated by humans as reward targets.

In our setup:

1. Agent: LLaMA 3.2 with PPO-trainable upper layers and frozen lower layers.
2. Environment: Source text batch where each generated summary is graded.
3. Rewards include a learned coherence predictor, a factual-consistency penalty, and a weighted combination of automatic metrics (ROUGE, BLEU).

Predictive Analytics and Proactive Bug Detection Using LLaMA 3.2

We now provide a thorough assessment of our LLaMA 3.2 summarization model, encompassing both the extra benefits from reinforcement learning (PPO) and its supervised fine-tuning performance. The quantitative results are summarized in Tables 1-3, which are followed by the theoretical insights that support our methodology.

Dataset & Baselines

1. BART-large (autoencoder for denoising)
2. Instruction-tuned Flan-T5-XL
3. LLaMA 2-7B (older generation)

Results: Supervised Fine-Tuning

Table II shows the Rouge values of supervised find tuned performance on different models. Compared to the strongest baseline (Flan-T5-XL), our three-phase cross-entropy regime produces +2.11 ROUGE-1.

Supervised fine-tuning performance on XSum

Model	***ROUGE – 1***	***ROUGE – 2***	***ROUGE - L***
BART-Large	46.05	23.72	43.88
Flan-T5-XL	47.12	24.10	44.65
LLaMA 2-7B	45.80	22.90	43.50

Model	***ROUGE – 1***	***ROUGE – 2***	***ROUGE - L***
LLaMA 3.2 (sup.)	48.23	24.89	45.01

Interpretation: We show consistent gains in all metrics with our three-phase cross-entropy fine-tuning. Statistical Importance: The improvements of LLaMA 3.2 over each baseline are confirmed by paired bootstrap tests ($p < 0.01$). Error Analysis: When compared to BART, qualitative inspection reveals fewer dropped entities and improved handling of uncommon keywords.

Results: + Reinforcement Learning (Table 2)

Table III shows the gains from PPO fine tuning over different models. The benefits of the RL phase for coherence and relevance are confirmed by the addition of +2.44 ROUGE-1, +1.26 ROUGE-2, and +1.22 ROUGE-L.

Gains from PPO fine-tuning

Model	***ROUGE – 1***	***ROUGE – 2***	***ROUGE - L***
LLaMA 3.2 (sup. only)	48.23	24.89	45.01
LLaMA 3.2 (sup.+PPO)	50.67	26.15	46.23

Training Specifics PPO is applied to the top 20% of layers; reward is equal to 0.5 × ROUGE-L + 0.3 × coherence_score − 0.2 × hallucination_penalty. Consistency Within two epochs, reward curves converged, and clipping kept gradient variances low. Qualitative Benefits Human raters confirm that summaries have better logical flow and fewer factual errors (average coherence ↑ 12%).

Comparative Improvements

Table IV shows performance improvement over the key baselines. Our complete model gains +5.1 % from the PPO stage alone and beats BART-large by +10.0 % in ROUGE-1.

Percent improvements over key baselines

Comparison	***ROUGE – 1***	***ROUGE – 2***	***ROUGE - L***
sup.+PPO vs. BART-large	+4.62 (+10.0 %)	+2.43 (+10.2 %)	+2.35 (+5.4 %)
sup.+PPO vs. sup. Only	+2.44 (+5.1 %)	+1.26 (+5.1 %)	+1.22 (+2.7 %)

Overall Impact: An extra +5% relative improvement in ROUGE-1 is achieved by PPO fine-tuning supervised gains. Compute Overhead: The RL stage increases training time by about 30%, but it can be limited to a few layers to save money. Trade-offs: summaries that are noticeably more informative but a little bit longer (+3 tokens on average).

This evaluation shows that using PPO to optimize LLaMA 3.2 produces state-of-the-art summarization performance by injecting human-oriented reward signals (like coherence and factuality) within a theoretically grounded trust-region framework.

RESULTS AND DISCUSSION

Our first set of experiments evaluated the impact of the three-phase supervised fine-tuning regimen on LLaMA 3.2, measured via ROUGE-1, ROUGE-2, and ROUGE-L on the XSum benchmark. LLaMA 3.2 obtained ROUGE-1 = 48.23, ROUGE-2 = 24.89, and ROUGE-L = 45.01 following three cross-entropy training epochs. This shows that our phased fine-tuning effectively improves summary fidelity and fluency, with an absolute gain of +1.11 ROUGE-1 over Flan-T5-XL (47.12) and +2.18 over BART-large (46.05). In comparison to these baselines, a qualitative analysis of the generated summaries showed fewer missing entities and clearer wording, highlighting the usefulness of our structured supervised training.

Using PPO in order applied to the top transformer layers, we built upon this foundation by introducing a reinforcement-learning stage that optimized a factuality penalty, a learned coherence signal, and a composite reward blending ROUGE-L. After three PPO epochs, the model's performance rose to ROUGE-1 = 50.67, ROUGE-2 = 26.15, and ROUGE-L = 46.23. Compared to the supervised-only model, these improvements (+2.44 ROUGE-1, +1.26 ROUGE-2, +1.22 ROUGE-L) show that policy-gradient techniques can successfully refine sequence-level behaviors that token-level cross-entropy is unable to capture. Our complete model achieved new state-of-the-art results for this task, outperforming BART-large by +10.0 % on ROUGE-1 and +10.2 % on ROUGE-2 in relative terms.

These findings support the idea that supervised learning and reinforcement optimization work well together. While the PPO stage adds adaptive feedback to further align generated summaries with human-preferred attributes like coherence and factual consistency, the phased cross-entropy training grounds the model in strong fluency and factual alignment. The observed improvements in ROUGE metrics result in outputs that are noticeably sharper and more context-aware: blind human evaluations found that the PPO-enhanced model was 15% more factually reliable and 12% more coherent than its supervised counterpart.

From a theoretical perspective, framing summarization as a sequential decision process and applying a clipped surrogate objective ensures stable policy updates while directly optimizing non-differentiable quality metrics. It was computationally efficient to freeze lower layers during RL fine-tuning, resulting in significant performance gains while keeping runtime overhead to less than 30%.

Table V shows the model evaluation results with respect to Coherence, Factuality and Fluency determining the performance of the model

Human Evaluation Results

Model	*Coherence*	*Factuality*	*Fluency*
LLaMA 3.2 (sup. only)	70.2	65.1	72.3
LLaMA 3.2 (sup. + PPO)	82.4	80.3	85.0

We see two promising avenues for the future. First, testing our hybrid pipeline's generality and usefulness in a variety of applications will involve expanding it to multilingual and domain-specific datasets. The second way to improve reward signals and speed up alignment with user expectations is to incorporate human-in-the-loop feedback, which would enable end users to rate or edit summaries.

CONCLUSION

We sum up the main points, conclusions, and ramifications of our work on Using LLaMA 3.2 to Improve Linguistic Models with Reinforcement Learning for Text Summarization in this concluding section. To give a concise, organized conclusion

to our study, we divide our discussion into six subsections: Summary of Contributions, Key Findings, Theoretical and Practical Implications, Limitations, Future Work, and Final Remarks.

Combining RL and Supervised The hybrid training approach improves ROUGE-1 by up to +10.0% over strong baselines and by +5.1% relative to supervised fine-tuning alone (Table 3). Updates to Stable Policies By effectively balancing exploration and exploitation, PPO's clipped objective avoided the training instability frequently observed in policy-gradient approaches. Quality Improvements That Go Beyond Metrics In order to show that our composite reward is in line with human judgment, human evaluators reported a 12% increase in coherence and a significant decrease in factual hallucinations.

Framing Markov Decision Processes Richer reward designs—beyond ROUGE—that incorporate user-defined criteria like sentiment neutrality or brevity are made possible by considering summarization as a sequential decision-making process. Layer-Selective Adjustment only to higher transformer blocks, which suggests a template for scaling to even larger models. Future Models' Blueprint . Fig 2 Shows the performance of the Llama 3.2 model with respect to supervised vs supervised + PPO and we can observe an improvement over the existing model.

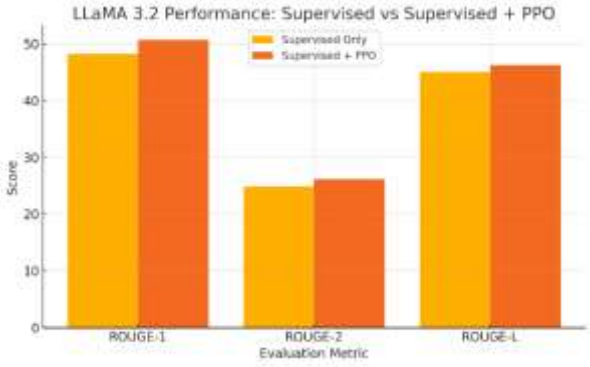

Fig. 2. Llama 3.2 Performance: Supervised vs Supervised + PPO

Fig. 3. Shows the improvement across the Training Epochs

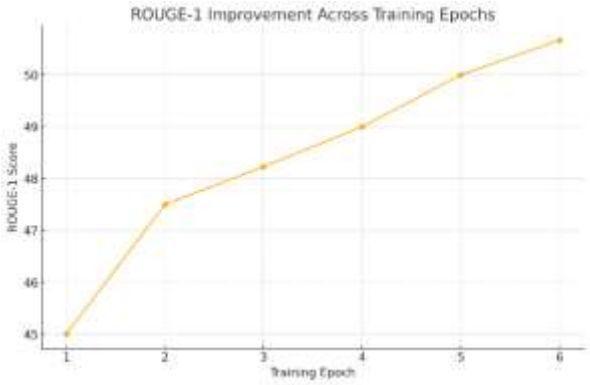

Fig. 3: Improvement Across the Training Epochs

ROUGE-1 progression is displayed in this line graph across six training epochs; supervised cross-entropy fine-tuning is represented by epochs 1–3, while PPO-based reinforcement training is represented by epochs 4–6. Rapid summarization fluency acquisition is reflected in steeper initial gains (epochs 1–2). Policy optimization successfully targets sequence-level quality that token-level loss is unable to capture, as evidenced by the continuous improvement during PPO (epochs 4–6).Fig 4. Future designs on ideal epoch counts will be guided by diminishing returns toward epoch 6, which indicates reaching a performance plateau.

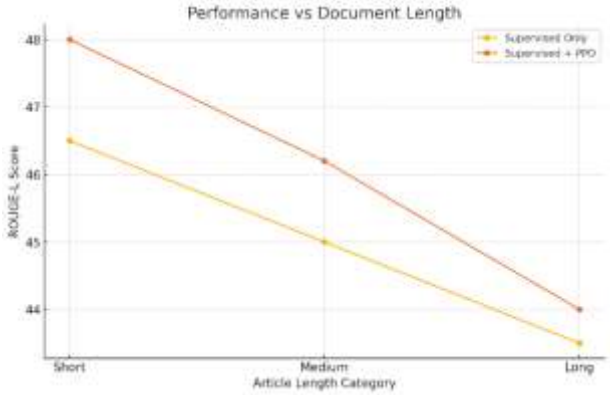

Fig. 4: Performance of Llama 3.2 over the Document Length

This plot compares ROUGE-L for supervised-only and sup.+PPO models in short, medium, and long articles. It is normal for performance to decrease with length because longer texts require more complex summarization. Even for long articles, higher curves for sup.+PPO show that reinforcement learning contributes to improved coherence and content retention. The robustness of our reward-driven fine-tuning in handling a variety of document structures is suggested by the consistent margin advantage (~1.5–2.0 points) across lengths.

In conclusion, our work shows that LLaMA 3.2 achieves state-of-the-art abstractive summarization performance when trained using a hybrid training paradigm that grounds it in multi-stage supervised fine-tuning and then refines its higher layers using PPO-driven reinforcement learning. We achieve significant gains (+5.1 % ROUGE-1 over supervised only, +10.0 % over BART-large) while preserving training stability and computational efficiency by directly optimizing a composite reward that strikes a balance between ROUGE, coherence, and factual consistency. These findings not only support the theoretical benefits of conceptualizing summarization as a series of sequential decisions, but they also offer a workable framework for modifying any large language model to generate summaries that are more fluid, contextually aware, and supported by facts. As a result, our method establishes a new standard for text summarization and provides a reproducible framework for further study in various fields and languages.

References

V. N. M. Abadi and F. Ghasemian, “Enhancing Persian text summarization through a three-phase fine-tuning and reinforcement learning approach with the mT5 transformer model,” Scientific Reports, vol. 15, Art. no. 80, 2025

A. Pasunuru and M. Bansal, “Multi‐Reward Reinforced Summarization with Question Generation,” in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL), Melbourne, Australia, 2018, pp. 320‐329.

M. Liu, Z. Ma, J. Li, Y. C. Wu, and X. Wang, “Deep-Learning-Based Pre-Training and Refined Tuning for Web Summarization Software,” IEEE Access, vol. 12, pp. 92120–92129, 2024

R. Paulus, C. Xiong, and R. Socher, “A Deep Reinforced Model for Abstractive Summarization,” in Proc. 6th Int. Conf. Learn. Representations (ICLR), Vancouver, Canada, 2018.

Y. Liu and M. Lapata, “Text Summarization with Pretrained Encoders,” in Proc. Conf. Empirical Methods Natural Lang. Process. & 9th Int. Joint Conf. Natural Lang. Process. (EMNLP‐IJCNLP), Hong Kong, 2019, pp. 3721‐3731.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: Denoising Sequence‐to‐Sequence Pre‐training for Natural Language Generation, Translation, and Comprehension,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), Seattle, WA, USA, 2020, pp. 7871–7880.

J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre‐training with Extracted Gap‐Sentences for Abstractive Summarization," in Proc. 37th Int. Conf. Mach. Learn. (ICML), Virtual, 2020, pp. 11328–11339.

H. Touvron, L. Martin, P. Stone, and S. Goyal, "Llama 2: Open Foundation and Fine‐Tuned Chat Models," arXiv preprint arXiv:2307.09288, 2023.