Automatic Sinhala Headline Generation for Newspaper Articles

Student: B.P.A. Cooray(209319K)
Supervisors: Dr. Surangika Ranatunge
Dr. Nisansa de Silva

Overview

Introduction

Research Problem

Research Objectives

Related Work

Dataset

Methodology

Experiments

Results

Conclusion

Introduction

- First Impression → Drives engagement
- High Impact → Useful in education, law, media [3][4]
- Faster Comprehension → Aids understanding
- Manual Effort → Slow, not scalable
- Automation → Fast, efficient, boosts reach
- Powered by NLP → Abstractive methods & PLMs [1–3]

[1] B. Baykara and T. Güngör, “Turkish abstractive text summarization using pretrained sequence-to-sequence models,” Natural Language Engineering, vol. 29, no. 5, pp. 1275–1304, 2023.
[2] A. Vaswani, “Attention is all you need,” Advances in Neural Information Processing Systems, 2017.
[3] R. Jayakody and G. Dias, “Performance of recent large language models for a low-resourced language,” in 2024 International Conference on Asian Language Processing (IALP). IEEE, 2024, pp. 162–167.
[4] B. Rathnayake, K. Manathunga, and D. Kasthurirathna, “” talking books”: A sinhala abstractive text summarization approach for sinhala textbooks,” in 2023 IEEE 8th International Conference for Convergence in Technology (I2CT). IEEE, 2023, pp. 1–6

Motivation

- Focus → Tailored for low-resource Sinhala
- Low-Resource → Data scarcity limits progress
- Tool Mismatch → Existing NLP tools not suited [1]
- PLMs Work → Proven in low-resource settings [2]

[1] M. Jahan and K. Wijesekara, “Automated text summarization of sinhala online articles,” Journal of Science-FASSEUSL, vol. 4, no. 01, pp. 01–15, 2023.
[2] R. Jayakody and G. Dias, “Performance of recent large language models for a low-resourced language,” in 2024 International Conference on Asian Language Processing (IALP). IEEE, 2024, pp. 162–167.

Motivation: Sinhala-Specific Gaps

For Sinhala Language,

- Mostly Extractive → Lacks coherence & nuance [1–3]
- Abstractive via Translation → Context loss [4]
- Direct Abstractive Models → Underexplored [5][6]
- Translation-Based Output → Reduces quality

[1]H. Jayawardane, “Automatic sinhala text summarization for government gazettes using abstractive and extractive methods,” Ph.D. dissertation, 2022.
[2] O. Wimalasuriya, “Automatic text summarization for sinhala,” Ph.D. dissertation, 2021.
[3]W. Welgama, “Automatic text summarization for sinhala,” Ph.D. dissertation, 2012.
[4] B. Rathnayake, K. Manathunga, and D. Kasthurirathna, “” talking books”: A sinhala abstractive text summarization approach for sinhala textbooks,” in 2023 IEEE 8th International Conference for Convergence in Technology (I2CT). IEEE, 2023, pp. 1–6
[5] R. Jayakody and G. Dias, “Performance of recent large language models for a low-resourced language,” in 2024 International Conference on Asian Language Processing (IALP). IEEE, 2024, pp. 162–167.
[6]K. Hewapathirana, N. de Silva, and C. Athuraliya, “M2ds: Multilingual dataset for multi-document summarisation,” in International Conference on Computational Collective Intelligence. Springer, 2024, pp. 219–231.

Research Problem

How can we leverage PLM-based **direct abstractive** approaches to generate coherent, and meaningful Sinhala headlines—without relying on translation-based methods?

Research Objectives

- Advance Beyond Extractive Methods in Sinhala
  - Explore effective abstractive techniques for low-resource Sinhala.
- Adapt Models for Sinhala
  - Fine-tune and customize for better performance.
- Evaluate PLM + PEFT
  - Assess impact on headline quality.
- Evaluate output quality rigorously
  - Automatic metrics like ROUGE

PEFT: https://huggingface.co/docs/peft

Related Work: Abstractive Text Summarization using PLMs

- Abstractive Summarization
  - Human-like Output → Predicts next token for natural flow
  - Better than Extractive → More flexible & concise
  - Challenges → Limited data, high compute, long training [1]
  - Solutions → Fine-tuning, prompting, distillation [1]

[1]] H. Jin, Y. Zhang, D. Meng, J. Wang, and J. Tan, “A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods,” 2024. [Online]. Available: https://arxiv.org/abs/2403.02901

Related Work: Pretrained Models for Sinhala

- Monolingual
  - SinBerto[1] & SinhalaBERTo[2] → Underused, small corpora
  - SinBERT[1] (RoBERTa-based) → Strong for classification [1]
  - Gap → No models for summarization/generation
- Multilingual
  - XLM-R[2] → Excels in Sinhala classification, beats LaBSE/LASER [1]
  - mBART[3] → Effective for multilingual generation & summarization [4]
  - mT5 → Top-tier text-to-text model, strong across tasks [5][6]

1. https://huggingface.co/Kalindu/SinBerto
2. https://huggingface.co/keshan/SinhalaBERTo
[1] V. Dhananjaya, P. Demotte, S. Ranathunga, and S. Jayasena, “Bertifying sinhala–a comprehensive analysis of pre-trained language models for sinhala text classification,” arXiv preprint arXiv:2208.07864, 2022.
[2] Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." arXiv preprint arXiv:1911.02116 (2019).
[3] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, “Multilingual denoising pre-training for neural machine translation,” 2020. [Online]. Available: https://arxiv.org/abs/2001.08210
[4] P. Dhakal and D. S. Baral, “Abstractive summarization of low resourced nepali language using multilingual transformers,” arXiv preprint arXiv:2409.19566, 2024.
[5] L. Xue, “mt5: A massively multilingual pre-trained text-to-text transformer,” arXiv preprint arXiv:2010.11934, 2020.
[6] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, “Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation,” in International Conference on Machine Learning. PMLR, 2020, pp. 4411–4421.

Related Work: Recent Sinhala Literature

Ref	Method	Key Focus	Results
[1]	Abstractive	LLMs (Llama 3, Mistral, GPT-4 & Claude), fine-tuning	GPT-4 & Claude excel with prompting; fine-tuning boosts Llama 3 & Mistral
[2]	Extractive/ Statistical	Statistical & pretrained models	Llama 2 7B leads; PRIMERA fine-tuning improves results
[3]	Extractive/ Abstractive	RNN & Transformer models	mT5 scored 39% for Sinhala
[4]	Extractive	TF-IDF & TextRank	TF-IDF better for medium-length articles

[1] R. Jayakody and G. Dias, “Performance of recent large language models for a low-resourced language,” in 2024 International Conference on Asian Language Processing (IALP). IEEE, 2024, pp. 162–167
[2] K. Hewapathirana, N. de Silva, and C. Athuraliya, “M2ds: Multilingual dataset for multi-document summarisation,” in International Conference on Computational Collective Intelligence. Springer, 2024, pp. 219–231.
[3] Y. Verma, A. Jangra, R. Kumar, and S. Saha, “Large scale multi-lingual multi-modal summarization dataset,” arXiv preprint arXiv:2302.06560, 2023
[4] M. Jahan and K. Wijesekara, “Automated text summarization of sinhala online articles,” Journal of Science-FASSEUSL, vol. 4, no. 01, pp. 01–15, 2023

Related Work: Recent Sinhala Literature

Ref	Method	Key Focus	Results
[5]	Abstractive	Translate → Summarize → Back-translate	Abstractive outperforms extractive (ROUGE + human eval)
[6]	Abstractive /Extractive	Extractive uses keywords from abstractive summaries	F-scores: 37%–71%
[7]	Extractive	Linguistic + statistical + TextRank	ROUGE scores reported

[5] B. Rathnayake, K. Manathunga, and D. Kasthurirathna, “” talking books”: A sinhala abstractive text summarization approach for sinhala textbooks,” in 2023 IEEE 8th International Conference for Convergence in Technology (I2CT). IEEE, 2023, pp. 1–6.
[6] H. Jayawardane, “Automatic sinhala text summarization for government gazettes using abstractive and extractive methods,” Ph.D. dissertation, 2022.
[7] O. Wimalasuriya, “Automatic text summarization for sinhala,” Ph.D. dissertation, 2021.

Related Work: Low-Resource Other Languages

Ref	Method	Key Focus	Results
[9]	Abstractive	Turkish summarization	mBART outperforms mT5 & PEGASUS
[10]	Abstractive	Nepali summarization	mBART beats mT5, especially with quantization

[9] B. Baykara and T. Güngör, “Turkish abstractive text summarization using pretrained sequence-to-sequence models,” Natural Language Engineering, vol. 29, no. 5, pp. 1275–1304, 2023.
[10] P. Dhakal and D. S. Baral, “Abstractive summarization of low resourced nepali language using multilingual transformers,” arXiv preprint arXiv:2409.19566, 2024.

Related Work: PEFT Frameworks

- Goal: Efficient Multi-Task Learning with Fewer Parameters
- Unified Framework [1]:
  - Scaled parallel adapters boost FFN layers
  - Larger models benefit more
  - Prefix tuning (~0.1% params) yields strong results
  - MAM Adapter
    - Prefix tuning on attention (bottleneck l=30)
    - Scaled parallel adapters (r=512) for FFN
- UNIPELT Framework [2]:
  - Gating to select best submodules (Adapter, PrefixTuning, LoRA) per task
  - Outperforms single PEFT methods, excels in low-resource cases

[1] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, “Towards a unified view of parameter-efficient transfer learning,” arXiv preprint arXiv:2110.04366, 2021.
[2] Y. Mao, L. Mathias, R. Hou, A. Almahairi, H. Ma, J. Han, W.-t. Yih, and M. Khabsa, “Unipelt: A unified framework for parameter-efficient language model tuning,” arXiv preprint arXiv:2110.07577, 2021.

Dataset

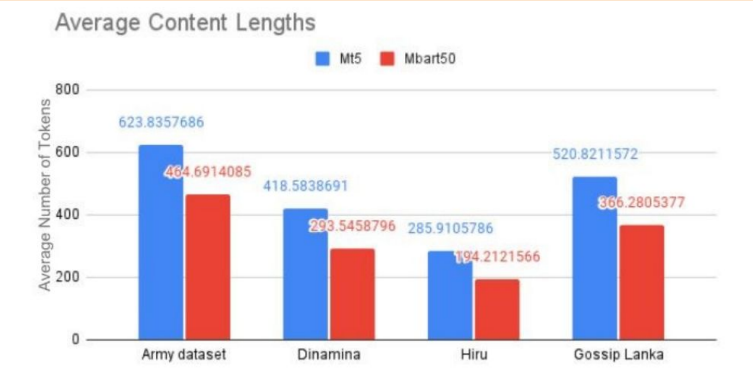
- Source: Sinhala newspaper repositories[1]
- Content: Articles + headlines
- Sources: Army, Dinamina, Hiru, Gossip Lanka
- Sizes:
  - Small: 1700 / 150 / 139
  - Large: 5100 / 450 / 417
- Reason: Diverse writing styles for richer summarization

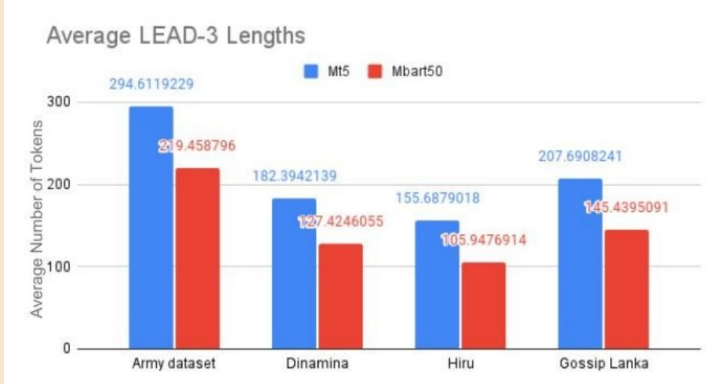
1. https://dms.uom.lk/s/6n5mWrQoaJgjEWF?path=%2FSinhala

Dataset: Qualitative Analysis

Dataset	Army	Dinamina	Hiru	Gossip Lanka
Samples	නිවාඩු නොමැතිව සේවයට නොපැමිණ සිටින ත්‍රිවිධ හමුදා සාමාජිකයින් සඳහා මසක පොදු සමා කාලයක් ප්‍රකාශයට පත් කෙරේ	පටු දේශපාලන භේද පසෙකලා එකතු වෙමු - අගමැති එජාප මැතිවරණ ව්‍යාපාරය ගැන අගමැති කියයි උ. කොරියාව 'ගුවාම්' දූපතට පහර දෙයි ද?	ලක්ෂ්මන් වෙඬරුවගේ නිවසට වෙඩි ප්‍රහාරයක් 'කුෂ්' සමග ඉරාන ජාතික කාන්තාවක් දැලේ හෙලිකොප්ටරයක් කඩා වැටී 25ක් මරුට	දෙස්තර නෝනාට ලෙඩ පෙන්වන්න ඇවිත් අවි පෙන්නා රන්බඩු කොල්ලකාලා ගුවන් යානයක් ගුවනේදී හදිසියේ ගිනි ගත් හැටි
Key Characteristics	Structured, long, formal headlines.	Less structured, formal with some spoken language.	Mix of formal and spoken style, broader patterns than Dinamina.	Informal, spoken style with modern idioms.

Dataset: Quantitative Analysis





Avg. Token Count (mT5 & mBART50)

- Measured on: Headlines & article content
- Tokenizers: mT5, mBART50

Methodology: Model Selection

- Encoder-Decoder
  - Used in summarization & headline generation.
  - Encodes input, decodes output with attention for long-range context.
- Decoder-Only
  - Autoregressive, predicts tokens from left context only.
  - Good for fluent, coherent text but lacks bidirectional context.

Methodology: Experiments Flows

Dataset Strategy

Small: Full fine-tuning for early experiments & fast prototyping

Large: PEFT-based training for efficiency with fewer parameters

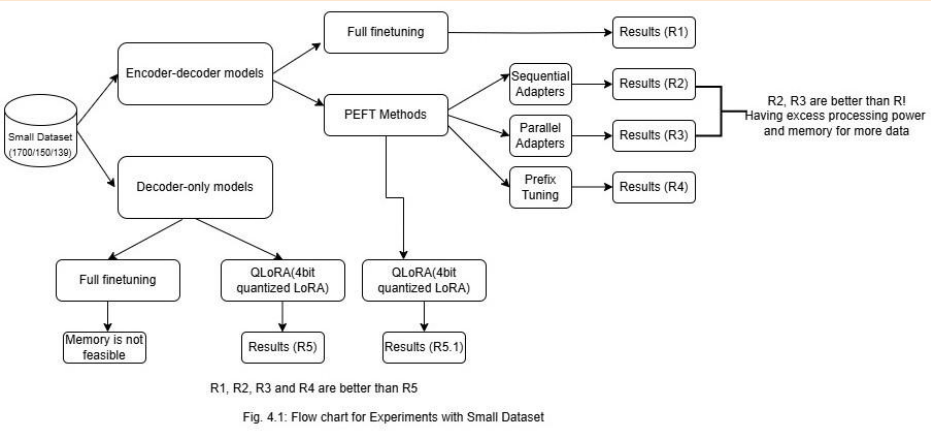


Fig. 4.1: Flow chart for Experiments with Small Dataset

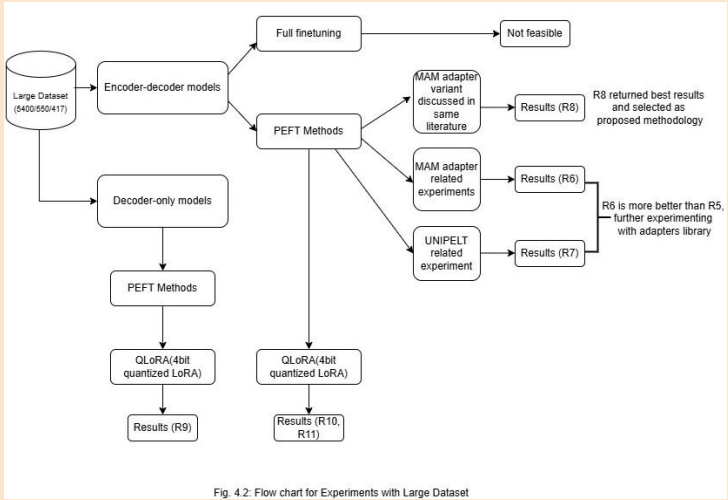


Fig. 4.2: Flow chart for Experiments with Large Dataset

Proposed Methodology

- Focus: Apply in FFN & Attention for better headlines
- Tuning:
  - High reduction → Attention
  - Low reduction → FFN (see Fig. 4.3)
- Adaptive Scaling: Dynamic, task-aware adjustment
- Benefits:
  - Efficient & performant
  - More coherent headlines

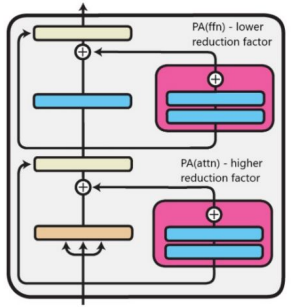


Fig. 4.3: Proposed methodology - architecture diagram: the original diagram depicted from adapters library and modified.

Experiments

Transformer based Encoder-Decoder Models for the Target task

Approach: Fine-tune pre-trained Seq2Seq models for Sinhala headlines (based on Baykara & Güngör [1])

- Adaptations: Handled Sinhala-specific tokens & sentence splits
- Evaluation: Improved ROUGE accuracy
- Optimization: Standard fine-tuning setup (Adam + LR scheduler)

1. https://www.nltk.org/_modules/nltk/tokenize/punkt.html

[1] B. Baykara and T. Güngör, “Turkish abstractive text summarization using pretrained sequence-to-sequence models,” Natural Language Engineering, vol. 29, no. 5, pp. 1275–1304, 2023.

Transformer based Encoder-Decoder Models for the Target task

Approach: Adapter-based fine-tuning (Adapters vs Adapter Frameworks (MAM & UNIPELT))

Dataset: Sinhala Dinamina (semi-formal style)

Tools: Hugging Face + Adapters library

Benefits: Modular, task-specific, easy to integrate

1. https://docs.adapterhub.ml/

Transformer based Encoder-Decoder Models for the Target task

- Approaches Explored:
  - Full tuning, Prefix, Sequential (SA), Parallel (PA), UNIPELT, MAM
- MAM Adapter (Mix-and-Match) [1]
  - Combines adapter styles
  - Best performance with Parallel Adapters in FFN
  - Prefix tuning effective with minimal parameters (with ~0.1% parameters)
- UNIPELT [2]
  - Combines multiple PEFT methods with dynamic gating
  - Underperformed vs. MAM, on Dinamina dataset

[1] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, “Towards a unified view of parameter-efficient transfer learning,” arXiv preprint arXiv:2110.04366, 2021.
[2] Y. Mao, L. Mathias, R. Hou, A. Almahairi, H. Ma, J. Han, W.-t. Yih, and M. Khabsa, “Unipelt: A unified framework for parameter-efficient language model tuning,” arXiv preprint arXiv:2110.07577, 2021.

Experiments: Transformer based Decoder only Models for the Target task

- Why LLaMA 3[1]
  - Open-source, low-resource friendly, clean tokenization
- QLoRA
  - Objective: Instruction tuning under limited resources
  - Model Used: Llama-3.2-3B (4-bit, ~2.24 GB)
  - Inspired by MAM Adapter findings
    - Adjusted attention and FFN matrices:
- Tools: Unsloth for fast, memory-efficient training
- Evaluation: ROUGE scores

1. https://unsloth.ai
[1] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan et al., "The llama 3 herd of models," arXiv preprint arXiv:2407.21783, 2024.
[2] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," Advances in Neural Information Processing Systems, vol. 36, 2024.

Results

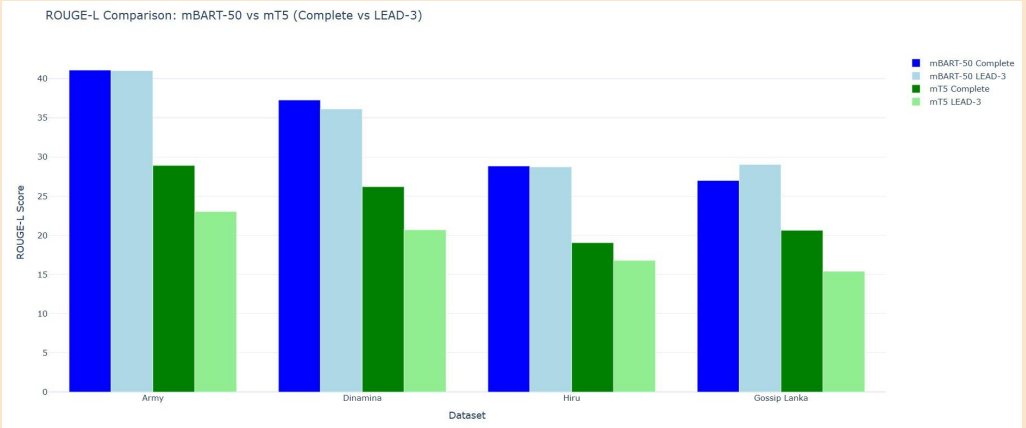
Transformer based Encoder-Decoder Models Experimental Results and Discussion

Experiment 1: Vanilla fine-tuning

Models: mBART vs. mT5

Finding: mBART outperformed mT5 ROUGE scores

Note: mT5 generated longer, more redundant headlines



Transformer based Encoder-Decoder Models Experimental Results and Discussion

Model Strengths

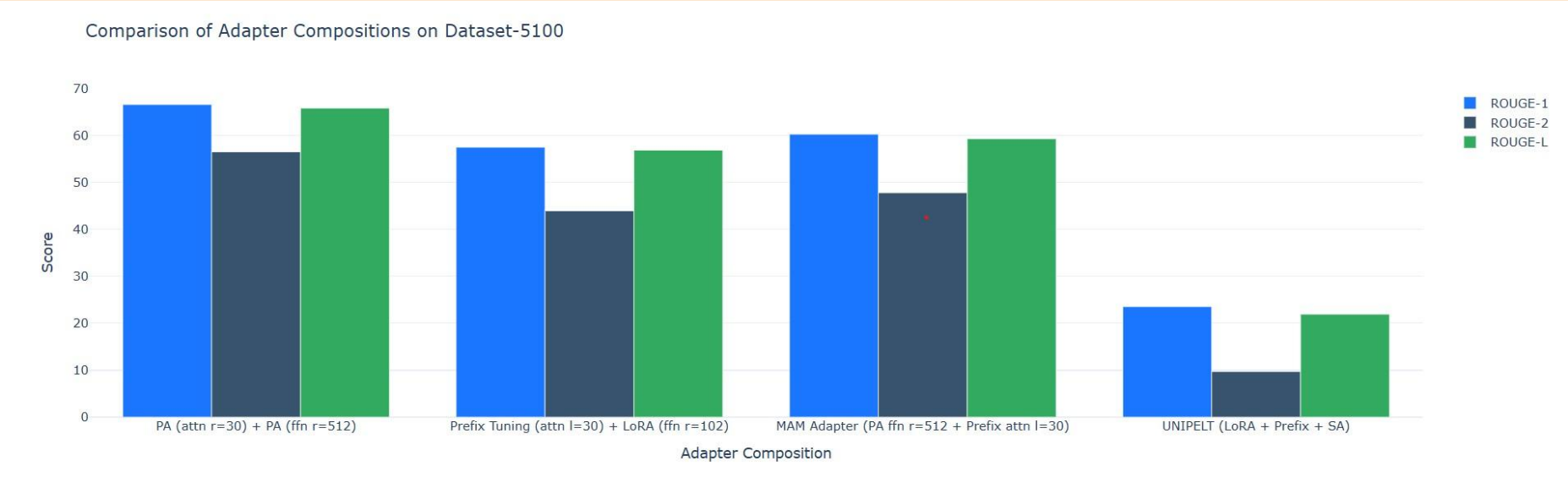
- mBART: Better context, strong seq2seq, relevant headlines
- mT5: Flexible; less stable on Sinhala dataset

Transformer based Encoder-Decoder Models Experimental Results and Discussion

Experiment 2: Mix-And-Match (MAM) Adapter Related Experiments

The findings of He et al. further support this conclusion within the context of the Sinhala news domain

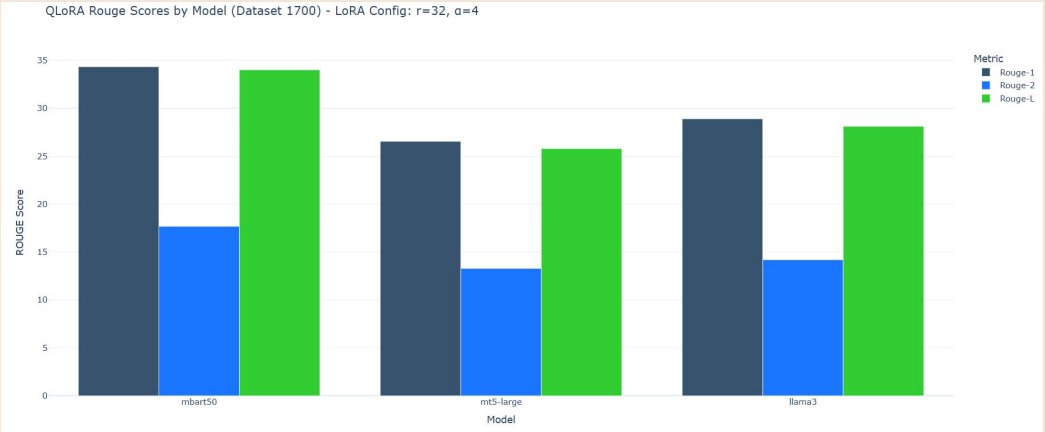
Transformer based Encoder-Decoder Models Experimental Results and Discussion

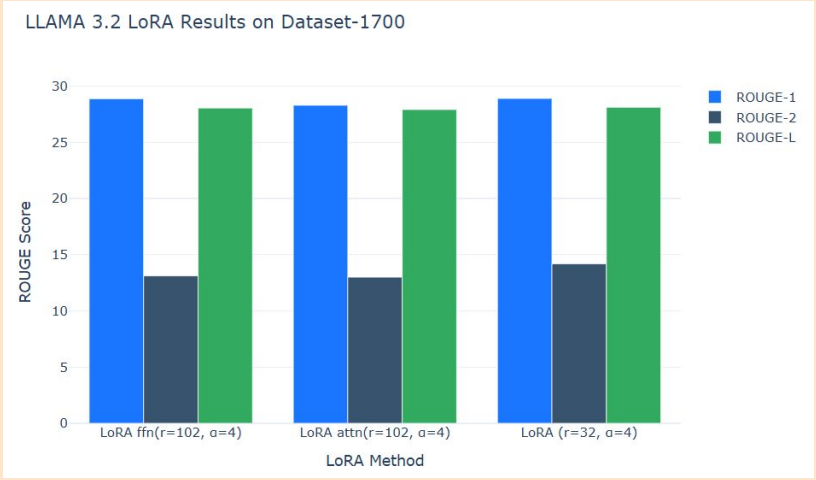


The UNIPELT results lag behind those achieved with other PEFT methods.

Transformer based Decoder-only Models Experimental Results and Discussion

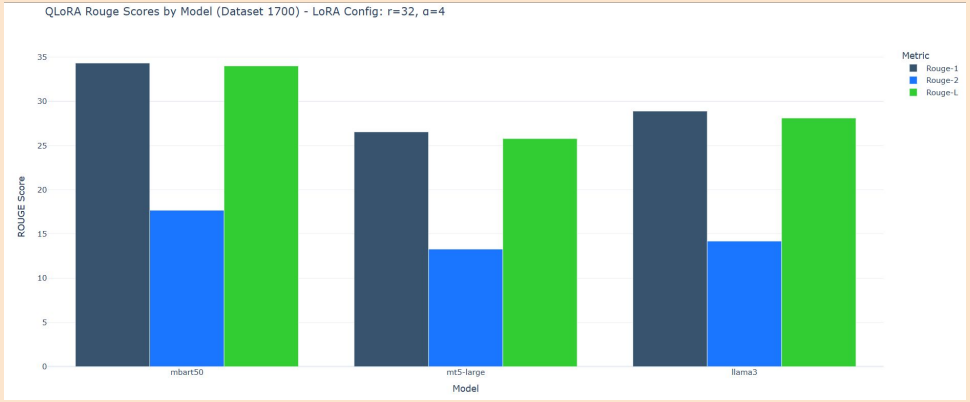
mBART50 vs mT5 vs LLaMA 3: QLoRA Tested on Two Datasets

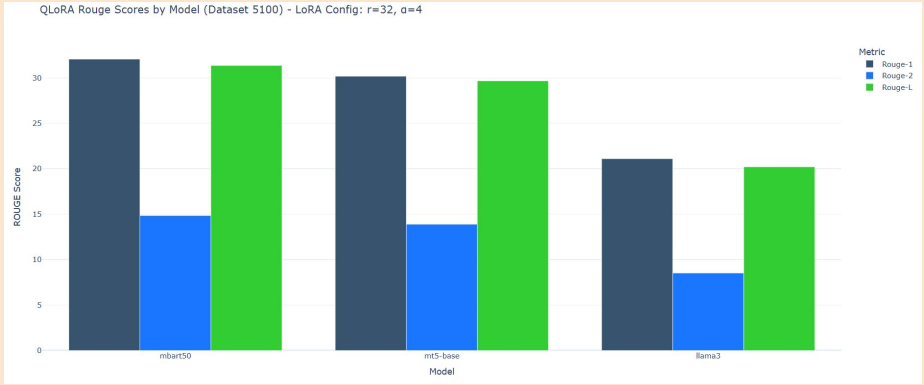




Transformer based Decoder-only Models Experimental Results and Discussion

mBART50 vs mT5 vs LLaMA 3: QLoRA Tested on Two Datasets





Key Findings from Experiments

Efficient Headline Generation

Encoder-Decoder > Decoder-Only on limited hardware

mBART50 > mT5 in accuracy and conciseness

Best Pick: mBART50 for resource-constrained setups

Key Findings from Experiments

Optimizing mBART50 with Adapters

Best: Parallel Adapters on FFN layers

Data helps: More = better tuning

Smarter Scaling: Dynamic > Fixed

Takeaway: Use PA (FFN) + ample data + dynamic scaling

Key Findings from Experiments

Balancing PEFT for Headlines

Prefix Tuning: Lightweight, decent gains

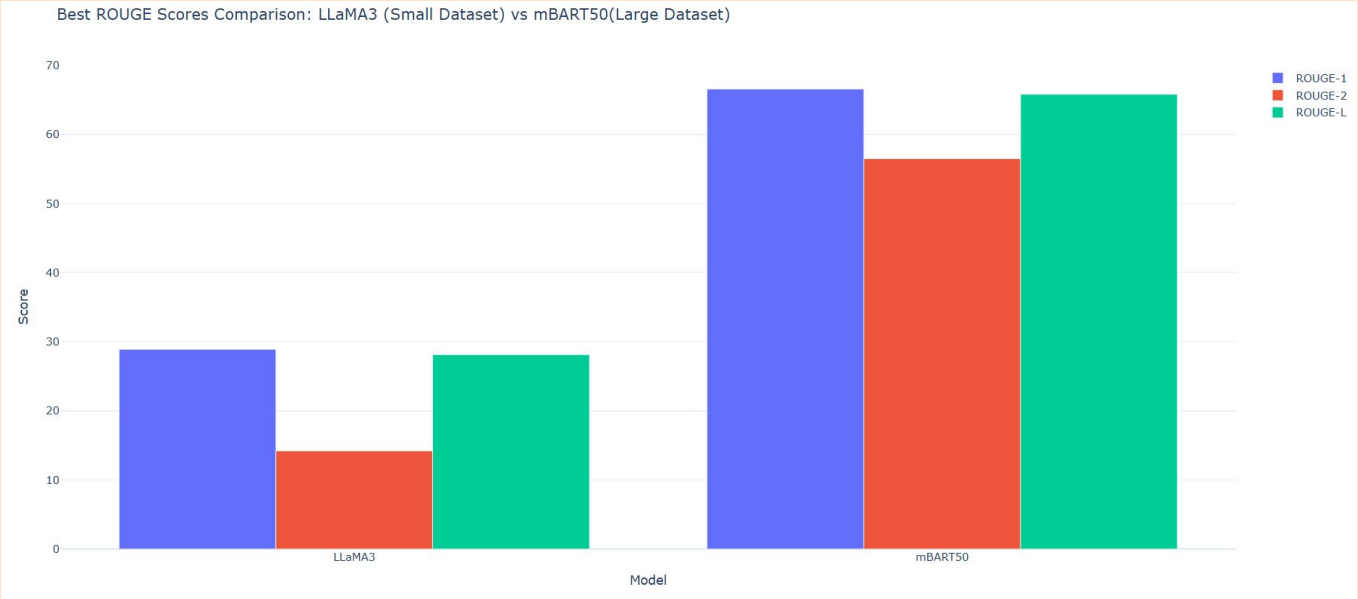
MAM Adapter: Balanced, but not top

Winner: PA on attn layers

Takeaway: Stick to PA (attn) for best results

Transformer based Decoder-only Models Experimental Results and Discussion

Best Results comparison from Encoder-decoder models vs decoder-only models



Proposed Methodology Results and Discussion

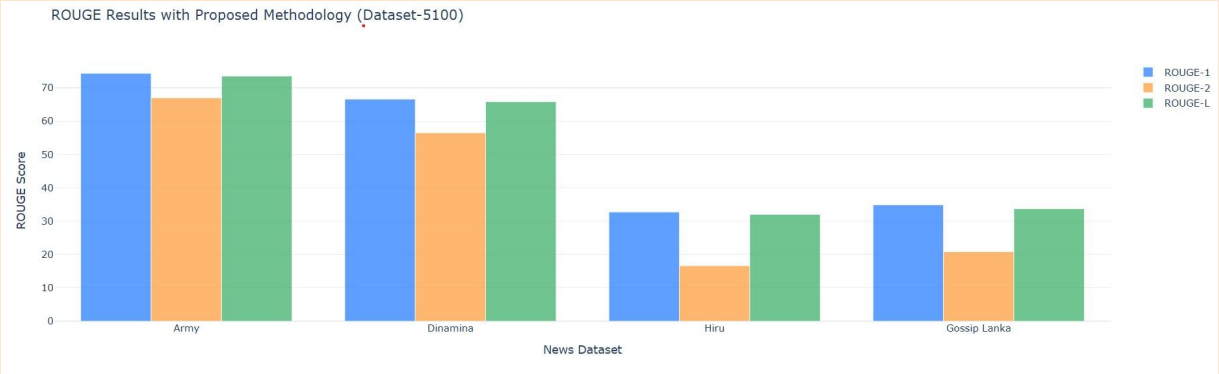
mBART Performance on News Datasets (ROUGE)

Army (Formal): Highest – clear structure helps

Dinamina (Semi-Formal): Strong, concise

Hiru (Mixed): Lower – diverse content hurts

Gossip Lanka (Informal): Mid – struggles with informality



Conclusion

- mBART excels on formal datasets
- Struggles with informal content (tone, metaphors)
- Needs domain-specific fine-tuning
- Current setup suits structured data best

- QLoRA + Unsloth enable low-resource training
- Useful for Sinhala news automation

- Future work: Improve informal text handling, Explore multi-task learning & larger datasets