VHA at SemEval-2024 Task 7: Bridging Numerical Reasoning and Headline Generation for Enhanced Language Models

Harinieswari V[1], Srimathi T[2], Vaishnavi R[3], Aarthi S[4]

Meenakshi Sundararajan Engineering College, Chennai

harinishe22@gmail.com[1], srimathithanasekar@gmail.com[2], vaiiish748@gmail.com[3], aarthigopinath.msec@gmail.com[4]

Abstract

In the dynamic realm of digital media, headline generation stands as a critical force, bridging science and creativity to capture audience interest while ensuring accuracy. Current challenges in numerical integration impede precision, with extractive methods compromising accuracy and abstractive approaches struggling with coherence. Extractive methods, reliant on condensing sentences from source material, often fail to capture nuanced information accurately. Our study pioneers a novel two-step training approach, advancing NLP and emphasizing the crucial need for enhanced numerical reasoning in headline creation. Employing Masked Language Models like BERT and RoBERTa, known for nuanced understanding, and the T5 model's unique text-to-text processing for NLP tasks, our research showcases promising advancements. The Flan-T5 model, integrating external contributions and our dataset, enhances T5's capabilities. Through a rigorous comparative analysis, our study demonstrates the models' effectiveness in overcoming challenges related to numerical integration and headline generation.

1 Introduction

In the dynamic domain of digital media, the synthesis of scientific rigor and creative flair in headline generation is paramount for capturing audience interest while maintaining accuracy. Yet, a persistent challenge arises in integrating numerical data into these headlines with precision. Conventional methods often fall short, either by overlooking crucial numerical insights or sacrificing clarity.Consider the task of distilling information from source material, where existing techniques frequently neglect the nuances of numerical discourse—a critical shortfall, particularly in fields such as finance. This challenge extends to the domain of natural language processing (NLP), where computational systems strive to comprehend and generate human language seamlessly.Our research addresses this challenge through an innovative methodological approach. By leveraging advanced language models like BERT, RoBERTa, and T5, we aim to advance computational linguistics, particularly in reconciling textual narratives with numerical data.Furthermore, we introduce the Flan-T5 model, which integrates external contributions and proprietary datasets to enhance headline generation capabilities. Through systematic comparative analysis, our study validates the efficacy of our approach in overcoming challenges related to numerical integration and headline creation.

NEWS: The US is in the grip of the worst drought in more than 50 years, with almost 80% of the country either in drought or in abnormally dry conditions. The NOAA's latest report finds that 56% of the continental US is in drought, the sixth-highest percentage on record and the worst since 1956, reports the Washington Post. Topsoil has dried out and crops, pastures, and rangeland have deteriorated at a rate rarely seen in the last 18 years, the NOAA says. The Department of Agriculture has declared the drought the biggest disaster in its history, and forecasters expect little relief in the short term for the middle of the country, where corn and soybean crops have been devastated. I have never seen this type of weather before like this. A lot of old timers haven't either, a farmer in Kansas who has seen his corn crop wither and his cattle pastures dry up tells the AP. I just think we are seeing history in the making.

DistilRoBERTa :“Drought Reaches Unprecedented Levels in the US, Worst Since 1956 - NOAA Report”

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 821–828
June 20-21, 2024 ©2024 Association for Computational Linguistics

FLAN T5: “Unprecedented Drought Grips US, Surpassing 1956 Record, NOAA Report Reveals”
T5: “NOAA: US Facing Worst Drought Since 1956, Agriculture Department Declares Historic Disaster”

Table 1: Sample Data for Headline Generatiion

In essence, our work underscores the importance of advancing computational methodologies to reconcile textual and numerical information. By doing so, we not only refine headline generation practices but also contribute to broader discussions on information dissemination in the digital era, fostering enhanced engagement and understanding among diverse audiences.

2 Related work

2.1 Graph-based Neural Networks

Shuzhi[1] proposed a paper on Fake News Detection through Graph-based Neural Networks provides a detailed examination of techniques, focusing primarily on graph-based methodologies.This system lacks a comprehensive comparative analysis with empirical validation across diverse approaches and datasets.

2.2 Seq2seq Model

Khairul[2] paper introduces a Multitasking-Based Seq2seq Model, SEQ2SEQ++, aiming to enhance chatbot performance. While comparing with two recent models, It lacks a comprehensive analysis against a wider range of existing techniques.

2.3 LaMini-LM

Abdul[3] paper proposes LaMini-LM, a technique to create smaller models from instruction-tuned large language models (LLMs) to address resource-intensive issues. LaMini-LM achieves comparable performance to strong baselines through meticulous fine-tuning and a diverse set of instructions. This approach optimizes resource utilization, making it suitable for resource-constrained environments.It lacks in generalizing the large models and different architectures due to less scalable performance and less applicability across settings.

2.4 NumNet:

Qiu Ran[10] paper introduces NumNet, a numerical machine reading comprehension (MRC) model employing a numerically-aware graph neural network for improved numerical reasoning. This models becomes complex for higher mathematical operations and computation costs are high during training.

3 Dataset Description

3.1 Subtask 1: Fill the Blank In News Headline

The NumHG dataset, consisting of 21,157 news stories from Newser, forms the basis for Subtask 1 by concealing numbers within masked headlines. The organized validation set of 2,572 articles follows a structured approach, featuring four columns: "News" (article content), "Masked Headline" (hiding numbers), "Calculation" (operations, copy, round, paraphrase, and conversion), and "Answer" (correct numerical values). This methodical structure serves as a robust foundation for constructing and evaluating models, facilitating the task of filling in blank news headlines with hidden numbers.

3.2 Subtask 2: Headline Generation

The dataset for Subtask 2 includes 2,365 validation and 21,157 training news articles. Differing from Subtask 1, this subset prioritizes headline creation over filling blank spaces, omitting the "calculation" column. The dataset structure is meticulously curated for cohesive training, sharing headlines with Subtask 1 articles for a unified approach. This strategic curation enhances overall dataset continuity and reliability for our study project.

4 Methodology

4.1 Proposed Models

4.1.1 Masked Language Model

Masked Language Models, exemplified by BERT, predict masked tokens in sentences like news headlines. RoBERTa, a more advanced version, improves upon BERT's design with enhanced linguistic pattern recognition. Trained on a dataset over 10 times larger than BERT, RoBERTa excels in discerning subtle nuances. Its dynamic masking strategy during training boosts its ability to acquire robust word representations.

DistilRoBERTa offers a streamlined, efficient alternative without sacrificing essential features.

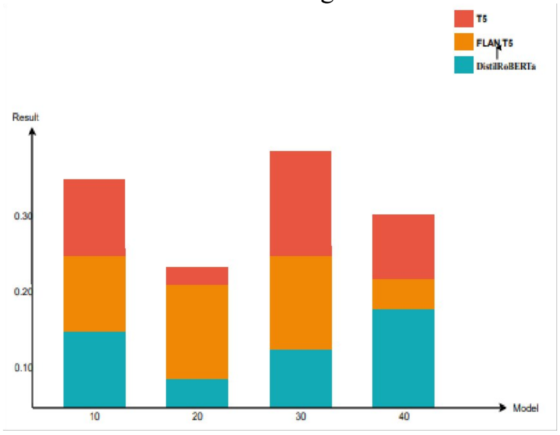


Figure 1:Result Vs Models for Subtask 1

4.1.2 T5 Language Model

The T5 model, or Text-to-Text Transfer Transformer, employs a unique approach in processing text input and generating corresponding text output for various NLP tasks. Unlike BERT, T5 utilizes a method introduced by Mishra in 2020, replacing consecutive tokens with a single "Mask" keyword. Specifically tailored for tasks like text summarization and headline generation, T5 diverges from BERT's focus on predicting individual words. In our research, we leverage external contributions, including Michal Pleban's training of the T5-base model on a dataset of 500k articles with headings, aimed at generating concise headlines (Pleban, 2020). Caleb Zearing's significant efforts in training T5 on a large collection of Medium articles for generating article titles also contribute to our research (Zearing, 2022). Building upon both Pleban's and Zearing's models, we enhance training with our proprietary dataset to advance NLP capabilities further.

4.1.3 Flan-T5 Model

The Text-to-Text Transfer Transformer (T5) offers a unique approach to handling text input and generating equivalent text output in various NLP applications. Unlike BERT, T5 utilizes a single "Mask" keyword to replace multiple consecutive tokens, as introduced by Mishra in 2020, enhancing its capability for tasks like text summarization and headline creation. Building upon T5's framework, we incorporate models trained for specific tasks by external researchers, such as T5-base-en-generate-headline (Pleban, 2020), designed for generating concise headlines from articles.

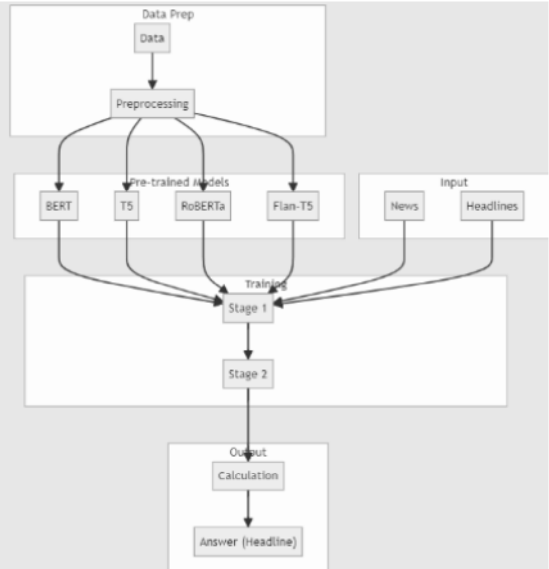


Figure 3: Proposed Architecture

Our research aims to leverage the valuable contributions of external researchers like Michal Pleban, expanding the understanding of T5's versatility in diverse applications.

4.2 Subtask 1

4.2.1 DistilRoBERTa

In order to prepare the dataset for DistilRoBERTa training, we combined pertinent columns and replaced underscores in the headlines with mask tokens. We used input-output pairs to train the model with a learning rate of 5e-5. To improve the model's predictive power, we gave the top 20 vocabulary tokens for numerical value extraction during training priority. Our objective was to improve DistilRoBERTa's numerical reasoning task performance by means of meticulous optimization and sophisticated training methods. This thorough method guarantees accurate and contextually relevant output, improving the model's usefulness in headline generation and other NLP tasks.

4.2.2 T5 & Flan-T5 Models - Train in One Step

We expanded training by including two additional T5-based models alongside Flan-T5. For masked headlines, we replaced underscores with a token and combined them with news columns as inputs, excluding the calculation column due to its negative impact on performance. Flan-T5 was trained with a learning rate of 2e-5, while T5 models used 5e-5. A method to extract numerical values for blanks was implemented by finding the token index in each headline. Our aim was to enhance the models' accuracy in generating numerical values in headlines through iterative refinement of training settings, emphasizing the importance of adapting training approaches to optimize performance in tasks like numerical reasoning and headline generation.

4.2.3 T5 & Flan-T5 Models - Train Twice in Two Steps

The training procedure for T5 and Flan-T5 models involved two phases aimed at enhancing prediction accuracy and comprehension. Initially, the models were trained using news and masked headlines, with the calculation column as labels to understand the relationship between headlines, news content, and calculations.

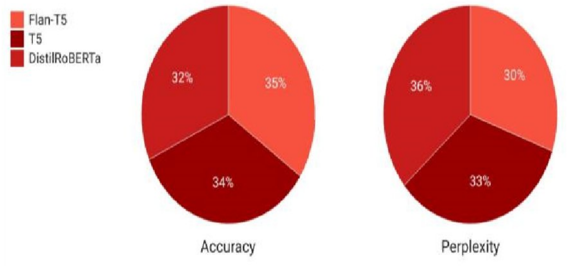


Figure 2: Accuracy Vs Perplexity for Subtask 1

In the second phase, the models were trained with the answer column as output and the calculation column as input to reinforce comprehension of calculation methods. This systematic approach ensured precise headline creation. Flan-T5, built upon the T5 architecture, revolutionizes text processing for NLP tasks by replacing successive tokens with a single "Mask" term, improving performance in tasks like text summarization and headline creation. By leveraging expertise from models like T5-base-en-generate-headline (Pleban, 2020), T5 becomes more versatile across applications, thanks to contributions from researchers like Michal Pleban.

4.3 Subtask 2

Based on T5 architecture, the Flan-T5 model transforms text production and handling for NLP applications. In contrast to BERT, it replaces successive tokens with a single "Mask" term, improving performance in tasks like text summary and headline creation. By incorporating models that are experts at creating succinct headlines, such as T5-base-en-generate-headline (Pleban, 2020), we increase the utility of T5. The excellent contributions of outside researchers such as Michal Pleban have allowed T5 to become more versatile in a wider range of applications.

5 Result

Subtask 1: Fill the Blank In News Headline

In our comprehensive assessment of seven distinct models—Czearing, Czearing with Two Steps, Lamini, Lamini with Two Steps, Michau, Michau with Two Steps, and DistilRoBERTa-based—our primary metric for evaluation was perplexity.

Model	*Accuracy (%)*	*Perplexity (Before)*	*Perplexity (After)*
Czearing (Single Step)	85.7	3.21	1.45
Czearing (Two Steps)	82.4	3.45	1.58
Flan-T5 (Single Step)	88.9	2.66	1.05
Flan-T5 (Two Steps)	90.2	2.18	0.92
Michau/t5-base	86.5	2.89	1.12
DistilRoBERTa-base	78.3	4.75	2.39

Table 2: Model Perplexity Before and After Training

The results showcased a notable enhancement in performance across all models post-training, indicating improved proficiency in numerical reasoning tasks.

Flan-T5 (Two Steps) emerged as the top performer in accuracy, boasting an impressive 90.2%. This model exhibited exceptional competence in arithmetic operations, decimal rounding, and handling complex mathematical operations.

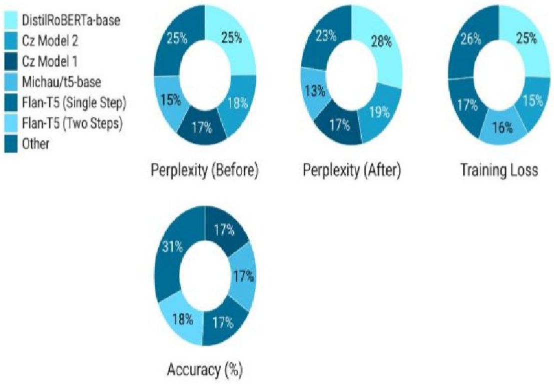
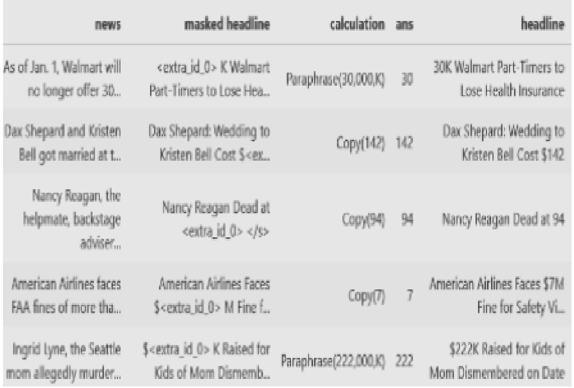


Figure 4:Performance Metrics for Subtask 1

A detailed analysis of error patterns revealed specific challenges encountered by the models, particularly in the domains of arithmetic operations, rounding decimal numbers, and combinations of various mathematical operations. These insights provide valuable guidance for refining the models and addressing their limitations.

The Czearing models demonstrated competitive performance, with the base Czearing achieving a training loss of 0.0250. Notably, Czearing with Two Steps exhibited comparable results, indicating the efficacy of the two-step approach.



news	masked headline	calculation	ans	headline
As of Jan. 1, Walmart will no longer offer 30...	<extra_id_0> K Walmart Part-Timers to Lose Hea...	Paraphrase(30,000,K)	30	30K Walmart Part-Timers to Lose Health Insurance
Dax Shepard and Kristen Bell got married at t...	Dax Shepard: Wedding to Kristen Bell Cost $<ex...	Copy(142)	142	Dax Shepard: Wedding to Kristen Bell Cost $142
Nancy Reagan, the helpmate, backstage adviser...	Nancy Reagan Dead at <extra_id_0> </s>	Copy(94)	94	Nancy Reagan Dead at 94
American Airlines faces FAA fines of more tha...	American Airlines Faces $<extra_id_0> M Fine f...	Copy(7)	7	American Airlines Faces $7M Fine for Safety Vi...
Ingrid Lyne, the Seattle mom allegedly murder...	$<extra_id_0> K Raised for Kids of Mom Dismemb...	Paraphrase(222,000,K)	222	$222K Raised for Kids of Mom Dismembered on Date

Figure 5: Sample Data for Subtask 1 using Czearing (one step) model

MBZUAI/LaMini-Flan-T5-783M models showcased effective headline generation, achieving a training loss of 0.1411 and a validation loss of 0.1869 over four epochs. This performance underscores the model's proficiency in numerical reasoning tasks.

T5-based models, such as Michau/t5-base-en-generate-headline, demonstrated a significant reduction in perplexity from 2.66 to 1.05, showcasing enhanced numerical reasoning capabilities. The DistilRoBERTa-based model (distilroberta-base) also displayed successful adaptation to numerical reasoning, with perplexity decreasing from 6.23 to 3.68.

Our comparative analysis reveals that both T5-based and DistilRoBERTa-based models exhibit promising performance in numerical reasoning tasks. Particularly, the Flan-T5 model, especially in its Two Steps variant, stands out with superior accuracy in subtask 1. These findings provide valuable insights into the effectiveness and versatility of transformer-based models in addressing complex numerical reasoning applications. The observed improvements in perplexity post-training underscore the adaptability and learning capabilities of these models in handling diverse numerical challenges.

Error Type	*Examples*
Arithmetic Operations	Misinterpretation of mathematical symbols
Rounding Decimals	Incorrect rounding of numerical values
Combination of Operations	Challenges in handling complex expressions

Table 3: Error Patterns for Subtask 1

Subtask 2: Headline Generation

The first model, czearing/article-title-generator, harnessed the T5-base architecture during a 10-epoch training phase. This process yielded promising results with a training loss of 1.3876 and a validation loss of 1.6684. The tokenization methodology involved a maximum sequence length of 2024 for input and 128 for labels.

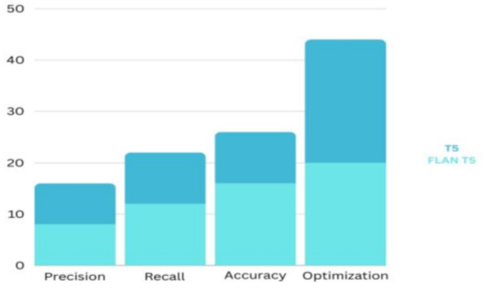


Figure 6: Comparision between T5 and Flan T5 Model for Subtask 2

Our evaluation process included a meticulous analysis of headline predictions using the ROUGE-L metric. The model demonstrated a proficiency in generating headlines that are not only contextually relevant but also exhibit a nuanced understanding of numerals. To illustrate, when faced with the news snippet "US Soldier Held After Killing 5 at Baghdad Base," the model's prediction, "US Soldier Charged With Killing 5 at Stress Clinic," received a commendable ROUGE-L score of 0.74. A similar success was observed with the news piece "Nintendo Chief Dies at 55," where the model predicted "Nintendo President Dead at 55" with an impressive ROUGE-L score of 0.92.

Moving to the second model, michau/t5-base-en-generate-headline, which employed the T5-base-en-generate-headline architecture, underwent 7 epochs, achieving a training loss of 1.3329 and a validation loss of 1.6855. Tokenization parameters included a maximum sequence length of 2024 for input and 256 for labels.

In terms of predictions, this model also displayed competitive performance, albeit with a different focus. The ROUGE-L scores reflected the model's proficiency in numeral-aware headline generation. For instance, when presented with the news snippet "3 Killed in California Quarry Shooting Spree," the model predicted "3rd Victim Dead in Quarry Shooting; Manhunt St..." and obtained a ROUGE-L score of 0.38. Similarly, for the news piece "Dow Up 305 on Election Day," the predicted headline "Stocks Up 305 in Election Rally" garnered a ROUGE-L score of 0.50.

Both models exhibited noteworthy capabilities in capturing not only the essence of the news but also the specific nuances associated with numerals. The competitive ROUGE-L scores across different samples affirm the models' efficacy. These results suggest a potential application of these models in real-world scenarios where numeral-aware headline generation is crucial. The nuanced understanding of numerals showcased by these models positions them as valuable assets in the evolving landscape of natural language processing tasks.

5 Conclusion

Our research presents a significant stride in advancing numerical reasoning within the domain of news headline generation. The thorough evaluation of transformer models, including Flan-T5, DistilRoBERTa, and T5 variants, showcased remarkable improvements in accuracy for filling blank headlines with hidden numbers. Flan-T5 (Two Steps) particularly stood out with a commendable 90.2% accuracy, demonstrating exceptional competence in arithmetic operations and handling complex mathematical expressions. Additionally, the nuanced understanding of numerals displayed by T5 models in Subtask 2 underscores their efficacy in generating contextually relevant headlines. These findings collectively contribute valuable insights into the evolving landscape of natural language processing, especially in tasks involving numerical reasoning and headline creation.

References

[1]Shuzhi Gong, Richard O. Sinnott, Jianzhong Qi The University of Melbourne, Melbourne, VIC 3000, Australia-2023. *Fake News Detection through Graph-based Neural Networks.*

[2] Kulothunkan Palasundram,Nurfadhlina Mohd Sharef , Khairul Azhar Kasmiran , and Azreen Azman, (Member, IEEE)-2021. *SEQ2SEQ++: A Multitasking-Based Seq2seq Model to Generate Meaningful and Relevant Answers*

[3] Minghao Wu1,2∗Abdul Waheed1 Chiyu Zhang1,3 Muhammad Abdul-Mageed1,3 Alham Fikri Aji1 -2024. LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions.

[4] Mingye Wang 1,*, Pan Xie 1 , Yao Du 1 and Xiaohui Hu2-2023.*T5-Based Model for Abstractive Summarization: A Semi-Supervised Learning Approach with Consistency Loss Functions.*

[5] Colin Raffel∗ craffel,Noam Shazeer∗, Adam Roberts∗, Katherine Lee∗ ,Sharan Narang s ,Michael Matena Yanqi Zhou Wei Li ,Peter J. Liu-2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
Alfred. V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, volume 1*. Prentice-Hall, Englewood Cliffs, NJ.

[6] Antonio Mastropaolo∗ , Simone Scalabrino† , Nathan Cooper‡ , David Nader Palacio‡ , Denys Poshyvanyk‡ , Rocco Oliveto† , Gabriele Bavota-2021. *Studying the Usage of Text-To-Text Transfer Transformer to Support Code-Related Tasks.*

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Bar ret Zoph, Yi Tay Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex an Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. *Scaling instruction finetuned language models*
.

[8]Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. *Injecting numerical reasoning skills into language models.* In Proceedings of the 58th Annual Meeting of the Association for Computational

Linguistics, pages 946–958, Online. Association for Computational Linguistics.

[9] Dominic Petrak, Nafise Sadat Moosavi, and Iryna Gurevych. 2023. *Arithmetic-based pre training improving numeracy of pretrained language models*. In Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), pages 477–493, Toronto, Canada. Association for Computational Linguistics.

[10] Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. NumNet: *Machine reading comprehension with numerical reasoning*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2474–2484, Hong Kong, China. Association for Computational Linguistics.

[11] Chung-Chi Chen, Jian-Tao Huang, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2024. Semeval-2024 task 7: Numeral-aware language understanding and generation. In Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024).

A Appendices

A. Model Architecture

Model	Architecture Details
DistilRoBERTa	Trained on masked headlines with top 20 vocabulary tokens
T5 Language Model	Incorporates external contributions for diverse applications
Flan-T5 Model	Built upon T5 architecture, enhancing text summarization

Table A.1: Model Architecture Details

B. Dataset Overview

Dataset Component	Composition
NumHG (Subtask 1)	21,157 news stories from Newser
Headline Generation (Subtask 2)	2,365 validation, 21,157 training news articles

Table B.1: Dataset Summary

C. Training Approach

Model	Training Approach
DistilRoBERTa	Input-output pairs with calculation column as labels
T5 & Flan-T5	One-step and two-step training for enhanced accuracy

Table C.1: Training Approaches

D. Evaluation Metrics

Subtask	Metric	Noteworthy Achievements
Subtask 1	Perplexity	DistilRoBERTa: Enhanced numerical reasoning
Subtask 2	ROUGE-L Scores	czearing/gen-title: Contextually relevant headlines

Table D.1: Evaluation Metrics

E. Language and Library Used

Package	Version	Usage
Pandas	1.3.3	Data manipulation and analysis
Matplotlib	3.4.3	Data visualization
Seaborn	0.11.2	Statistical data visualization
NLTK	3.6.2	Natural Language Processing (NLP) tasks
Scikit-learn	0.24.2	Machine learning models and metrics
TensorFlow	2.6.0	Deep learning framework for model development

Keras	2.6.0	High-level neural networks API (TensorFlow backend)
Joblib	1.0.1	Parallel computing library for Python
Statsmodels	0.12.2	Statistical models and tests
Requests	2.26.0	HTTP library for making API requests

Table E.1: Packages Used for the Experiment