ClusterCore at SemEval-2024 Task 7: Few Shot Prompting With Large Language Models for Numeral-Aware Headline Generation

Monika Singh, Sujit Kumar, Tanveen and Sanasam Ranbir Singh
Indian Institute of Technology Guwahati
{s.monika, sujitkumar, t.tanveen , ranbir}@iitg.ac.in

Abstract

The generation of headlines, a crucial aspect of abstractive summarization, aims to compress an entire article into a concise, single line of text despite the effectiveness of modern encoder-decoder models for text generation and summarization tasks. The encoder-decoder model commonly faces challenges in accurately generating numerical content within headlines. This study empirically explored LLMs for numeral-aware headline generation and proposed few-shot prompting with LLMs for numeral-aware headline generations. Experiments conducted on the *NumHG* dataset and NumEval-2024 test set suggest that fine-tuning LLMs on *NumHG* dataset enhances the performance of LLMs for numeral aware headline generation. Furthermore, few-shot prompting with LLMs surpassed the performance of fine-tuned LLMs for numeral-aware headline generation.

1 Introduction

News articles are one of the most important sources of information in everyday life. News headlines are vital in selecting which news seems relevant to read. As delineated in studies (Wei and Wan, 2017; Gabielkov et al., 2016), headlines play a significant role in making news viral on social media and influencing readers' opinions (Tannenbaum, 1953). Inaccurate, incongruent or misinformation headlines also lead to the spread of misinformation and disinformation over digital platforms (Chesney et al., 2017; Kumar et al., 2022, 2023). Consequently, generating an accurate headline for a news body is essential. Therefore, ensuring the accuracy of headlines is essential for maintaining the credibility and usefulness of news publications. The task of headline generation, which is a form of text summarization, aims to condense a lengthy source text into a concise summary. This summary, typically presented as a headline, encapsulates the main points of the original text, providing readers with a quick overview of the content (Huang et al., 2023).

In earlier studies on headline generation, various sequence-to-sequence and encoder-decoder methods have been employed to extract relevant headlines from news articles (Nallapati et al., 2016; Chen et al., 2020; Paulus et al., 2018; Song et al., 2019). However, encoder-decoder methods faced challenges in processing large sequences of text. To address these limitations, recent studies (Radford et al., 2018; Devlin et al., 2018; Lewis et al., 2019; Liu et al., 2019; Raffel et al., 2020) have proposed transformer-based models for headline generation by summarizing news articles. While transformer-based models have indeed showcased enhanced capabilities in handling longer text sequences and have exhibited promising outcomes in headline generation tasks; it is noteworthy that their performance in numeral-aware headline generation tasks need to be consistently superior. Despite their overall advancements, transformer-based models may face challenges in accurately incorporating and representing numeric information within generated headlines. Motivated by such observations, the study (Huang et al., 2023) proposed numeral aware headline generation datasets.
This paper introduces our proposed approach and provides a comprehensive analysis of the task of *Numeral-Aware Headline Generation* (Task 3 (2)). Our proposed methodology leverages Few-shot prompting with LLMs, which involves applying few-shot learning techniques to large language models (LLMs) for numeral-aware headline generation tasks. We conduct our experiments using the NumHG dataset (Huang et al., 2023) and the test set provided by the organizer of NumEval Task-3(2). Our experimental results suggest that few-shot prompting-based methods with LLMs are efficient for numeral-aware headline generation.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1719–1726
June 20-21, 2024 ©2024 Association for Computational Linguistics

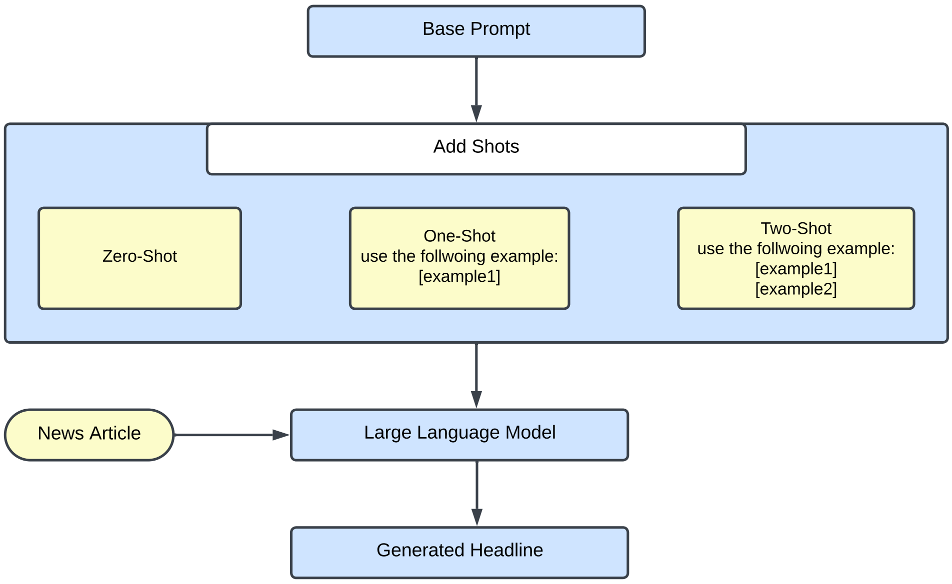

Figure 1: Working diagram of the proposed method.

2 Related Work

Headline generation, a type of text summarization, condenses lengthy source text into a brief summary, usually presented as a headline. This summary captures the main points of the original text, offering readers a quick overview (Huang et al., 2023). Summarization involves extractive and abstractive methods: Extractive selects key sentences, while abstractive generates novel summaries. In prior research investigating headline generation, a range of sequence-to-sequence and encoder-decoder approaches were employed to derive relevant headlines from news articles (Nallapati et al., 2016; Chen et al., 2020; Paulus et al., 2018; Song et al., 2019). However, these approaches encountered challenges, particularly in processing lengthy text sequences. The limitations of encoder-decoder methods in handling large sequences of text hindered their effectiveness in accurately summarizing news articles. To address these shortcomings and enhance the capability of headline generation models, recent research has focused on developing transformer-based architectures (Radford et al., 2018; Devlin et al., 2018; Lewis et al., 2019; Liu et al., 2019; Raffel et al., 2020). Similarly, Large Language Models LLMs such as GPT (Radford et al., 2019), BART (Lewis et al., 2020), T5 (Raffel et al., 2020) and LLaMA (Touvron et al., 2023) have also shown promising state-of-the-art models performance for text generation and summarization task.

Most studies above emphasize word selection and sentence structure, overlooking the significance of accurate numeric values in news headlines. Addressing this gap in the literature, a study (Huang et al., 2023) introduced numeral-aware headline generation datasets to facilitate the development of numeral-aware headline generation methods. Considering the superior performance of Large Language Models (LLMs) in text generation and summarization tasks (Basyal and Sanghvi, 2023), this study conducts an empirical study of LLMs for numeral-aware headline generation. Additionally, an error analysis is performed on the responses of LLMs for numeral aware headline generation. Subsequently, we propose Few-shot prompting with Large Language Models (LLMs) for numeral-aware headline generation.

3 Proposed Method

As mentioned above, the paper aims to study the effect of two important aspects of numeral aware headline generations. First, we study the effectiveness of large language models (LLMs) for numeral-aware headline generations. Second, we propose a few prompting-based methods for numeral-aware headline generations.

3.1 Large Language Models (LLMs):

Considering the effectiveness of LLMs in text summarization (Basyal and Sanghvi, 2023) and headline generations task (Ding et al., 2023). We fine-

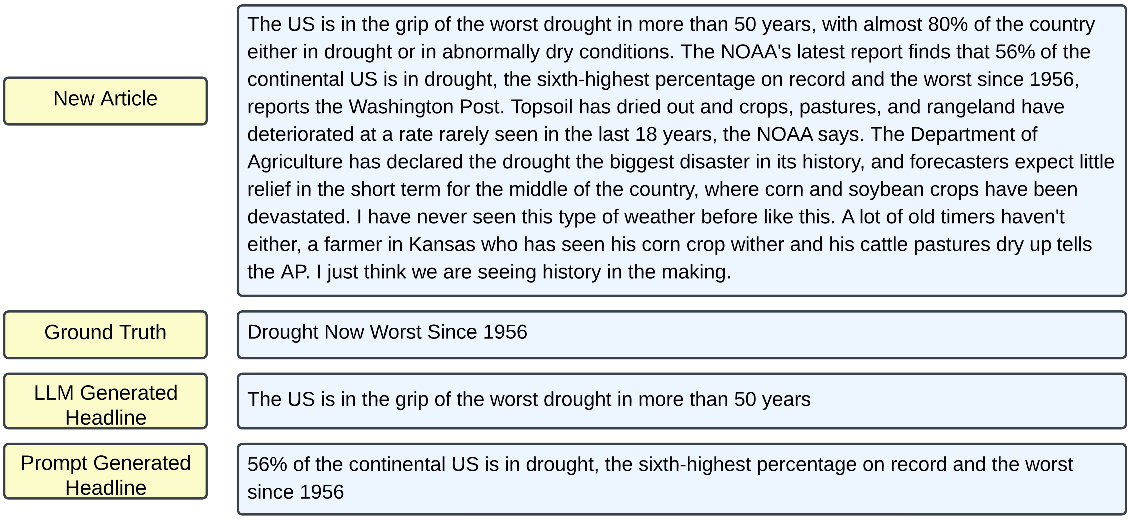

Figure 2: Presents an example comparison of a headline generated by a fine-tuned $T5$ model and a headline generated by a $T5$ model with three shot prompt

tune RoBERTa[1] (Rothe et al., 2020), *Generative Pre-trained Transformer* (GPT-2)[2] (Radford et al., 2019), *Bidirectional and Auto-Regressive Transformers* **BART**[3] (Lewis et al., 2020) and ***Text-To-Text Transfer Transformer*** **T5**[4] (Raffel et al., 2020) for numeral aware headline generations.

3.2 Few Shot Prompting:

In-context learning denotes a methodology whereby language models acquire proficiency in tasks by utilizing a limited number of examples provided as demonstrations (Dong et al., 2022). The utilization of shot prompting guides the model's output. This approach encompasses three distinct strategies: zero-shot, one-shot, and few-shot prompting. Zero-shot prompting, also called direct prompting, entails assigning a task to the model without providing specific examples, relying solely on the knowledge the model has gained through its training. In contrast, one-shot and few-shot prompting involve presenting examples or 'shots' to the model during runtime, which are references for the expected response's structure or context (Reynolds and McDonell, 2021). The model then utilizes these examples to perform the task. Because these examples are presented in natural language, they offer an accessible method for interacting with language models and facilitate the integration of human knowledge into these models through demonstrations and templates. As evidenced by the findings of several recent studies (van Zandvoort et al., 2023; Schick and Schütze, 2021; Luo et al., 2022), the integration of few-shot learning techniques coupled with prompt instructions has demonstrated a noteworthy enhancement in the quality of text generated or summarized by large language models (LLMs). These observations underscore the potential effectiveness of leveraging few-shot learning methodologies alongside prompt guidance to augment the capabilities of LLMs in generating text of higher quality and relevance. Motivated by such observations regarding few-shot learning with quick text generation and summarization instructions, this study proposes few-shot and prompt engineering-based methods for numeral-aware headline generations. Figure 1 presents the working diagram of our few shot prompting with LLMs-based numeral aware headline generation method. There are three key components of our proposed method, namely:

1. **Few Shot:** We explore three distinct strategies of few-shot prompting: zero-shot, one-shot, and few-shot prompting. These strategies encompass varying degrees of example provision to guide the model's output, allowing for a comprehensive analysis of their respective efficacy in facilitating model performance across different tasks. We have used three examples for methods in our study, which will

[1] https://huggingface.co/google/roberta2roberta_L-24_gigaword

[2] https://huggingface.co/MU-NLPC/CzeGPT-2_headline_generator

[3] facebook/bart-large-cnnÂůHuggingFace

[4] https://huggingface.co/Michau/t5-base-en-generate-headline

be considered three-shot prompting.

2. **Base Prompt:** Here, we provide instruction to a model which guides the model in numeral-aware headline generations. Below is one example of prompt instruction we provided to LLMs for generating numeral-aware headlines.
   Prompt (P1) : `Generate a short headline for a given news article. The headline should be concise and small but represent the content of the news body. The headline may contain a number that could be obtained by performing simple arithmetic operations like addition, subtraction, division, and multiplication or obtained by copying the same valid number from the news article if required to summarize the article.`

3. **Large Language models (LLMs):** This study considers three prominent large language models: GPT (Radford et al., 2018), T5 (Raffel et al., 2020), and LLaMA [5] (Touvron et al., 2023). These models generate headlines that accurately represent given news bodies, utilizing input consisting of the news body itself, prompt instructions, and a few-shot example.

4 Experimental Results and Discussions

4.1 Dataset

We consider the NumHG dataset curated by study (Huang et al., 2023) for training models and the test set provided by *NumEval* organizers for evaluating models. Table 4 presents the characteristics of experimental datasets.

4.2 Experimental Setups

This study incorporates several performance evaluation metrics to assess the effectiveness of models, namely *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE)[6] (Lin, 2004), BERTScore[7] (Zhang* et al., 2020), MoverScore(Zhao et al., 2019) and Num Acc. (Huang et al., 2023) as performance evaluation metrics to evaluate the performance of models. These metrics provide comprehensive insights into various aspects of model performance, including linguistic quality, content overlap, semantic similarity, and numeral accuracy, respectively. Table 3 presents the details of experimental hyperparameters. To replicate the findings in this work, visit GitHub `https://github.com/MONIKASINGH16999/ClusterCore_SemEval2024Task7` to access our code repository.

4.3 Results and discussion

Table 1 illustrates the performance metrics of large language models (LLMs) across various configurations, including *Pretrained*, *Fine-tuned*, and *Shot Prompting*, evaluated on a designated test dataset. This evaluation aims to provide insights into the efficacy and adaptability of LLMs in different settings for numeral-aware headline generation. Initially, we examine the response of LLMs in both the *Pretrained* and *Fine-tuned* setups for numeral-aware headline generation. From Table 1, it is evident that the $T5$ model consistently outperforms the $RoBERTa$, GPT, and $BART$ models across the test dataset in both the *Pretrained* and *Fine-tuned* setups. From such observations, we can claim that the $T5$ model is more suitable for the headline generation tasks compared to $RoBERTa$, GPT, and $BART$. Referring to Table 1, it becomes apparent that fine-tuning these models over the training set enhances their performance and headline generation capability. Subsequently, we curate a subset of the dataset consisting of fifty news headlines and corresponding news bodies. This subset is formed by selecting pairs from the validation dataset where the presence of numeral figures in the headline is deemed particularly significant in accurately representing the content of the news body. Upon manual inspection of the news headlines generated by fine-tuned $RoBERTa$, GPT, $BART$, and $T5$ models over the subset of the dataset comprising fifty samples, our observations suggest that while the generated headlines are contextually similar to the ground truth headlines and effectively represent the content of the news body, the accuracy in representing numeral figures is notably average. From these observations, we can conclude that fine-tuned $RoBERTa$, GPT, $BART$, and $T5$ models exhibit high efficiency in headline generation but display slightly lower efficiency in numeral-

[5] LLaMA

[6] `https://huggingface.co/spaces/evaluate-metric/rouge`

[7] `https://huggingface.co/spaces/evaluate-metric/bertscore`

Table 1: presents the performance of the models over test datasets

	Model	Num Acc.			ROUGE			BERTScore			MoverScore
		Overall	Copy	Reasoning	1	2	L	P	R	F1.	
Pretrained	**RoBERTa**	20.761	31.943	9.579	18.558	10.325	17.394	83.611	84.728	84.158	53.258
	GPT	24.028	34.529	11.527	18.596	12.356	16.879	81.192	76.925	79.058	54.217
	BART	24.137	35.529	12.746	15.7	11.72	14.846	84.264	84.382	84.323	55.321
	T5	23.988	35.995	11.982	**19.023**	9.365	**17.152**	**85.985**	**85.355**	**85.638**	**57.298**
Finetuned	**RoBERTa**	21.726	32.594	10.859	18.558	10.325	17.394	85.5	86.355	85.907	54.258
	GPT	23.265	34.952	11.578	31.896	14.256	29.854	86.935	81.325	84.13	55.941
	BART	25.623	35.621	13.291	32.64	13.587	30.466	86.435	88.324	87.377	57.689
	T5	36.985	37.514	12.852	**34.352**	**13.876**	**32.365**	**87.383**	**89.682**	**88.532**	**59.982**
Shot Prompting	**GPT**	37.259	37.594	12.589	31.746	12.653	29.356	87.659	86.926	87.292	54.989
	T5	37.569	37.295	12.958	30.245	10.941	29.596	89.111	86.922	87.988	58.364
	LLaMA	**38.233**	**38.233**	**13.942**	**37.985**	**14.854**	**33.592**	**90.359**	**89.856**	**90.107**	**59.983**

Table 2: Presents the human evaluation of headlines generated by our proposed system (few shot prompting with LLMs) submitted to NumEval-224. The organizer of NumEval-2024 does this human evaluation of generated headlines.

Submission	Num Acc. (50 Headlines)	Recommendation (100 News)
ClusterCore	**1.60**	31
Noot Noot	1.68	11
Infrrd.ai	**1.81**	**22**
np$_{p}roblem$	1.57	14
hinoki	1.67	16
Challenges	1.70	10
NCL$_{N}LP$	1.73	16
YNU-HPCC	1.69	15
NoNameTeam	1.59	12

Table 3: Details of Experimental Setups and Hyperparameters

Hyperparameters	**Value**
Batch Size	16
Learning Rate	0.01
Maximum #word in news body	250
Maximum #word in headline	15

aware headline generation. One possible reason behind this discrepancy could be the requirement for complex mathematical reasoning capabilities in numeral-aware headline generation tasks. To enhance the performance of models in numeral-aware headline generation tasks, this study employs shot prompting methods. Shot prompting methods offer several advantages, primarily providing prompts to models that serve as instructions, guiding them on what specific task needs to be performed and how to approach it. Additionally, shot prompting methods supply a few examples to the models, aiding them in inference and comprehension for the underlying task. This approach enables the models to better understand the task and generate more

Table 4: present the characteristics of experimental datasets. Here, #sample indicates the number of news headlines and body pairs in the dataset. Similarly, #head and #Word indicate the average number of words in the headline and news body. Whereas #sent indicates the average number of sentences in the news body and #num indicates the average number of numeric figures in the news body.

	#sample	#head	#sent	#Word	#num
Train	21157	7.769	9.851	179.116	9.884
Dev	2367	7.723	9.719	178.396	9.595
Test	5227	8.082	10.427	190.006	10.186

accurate and contextually relevant headlines containing numeral figures. We consider GPT, $T5$ and $LLaMA$ in few shot prompt settings. From Table 1 it is apparent that $LLaMA$ the model outperforms GPT and $T5$ with few shot prompting. Similarly, it is also evident that $LLaMA$ a model with few shot prompting outperforms $RoBERTa$, GPT, $BART$, and $T5$ models in *Pretrained* and *Fine-tuned* setups. Our manual inspection of the news headlines generated by the GPT, $T5$, and $LLaMA$ models utilizing few-shot prompting over a subset of the dataset containing fifty samples suggests that the implementation of few-shot prompting enhances the efficiency of numeral-aware headline generation by the models. Based on the findings presented in Table 1, it's clear that few-shot prompting using the $LLaMA$ model outperforms both few-shot prompting with $T5$ and GPT. As a result, we chose to submit headlines generated by the few-shot prompting with the $LLaMA$ model as our final system for evaluation at NumEval-2024. We could have fine-tuned the $LLaMA$ model for better results, but we have only used the pre-trained $LLaMA$ model due to resource constraints.

Table 2 presents the human evaluation of headlines generated by our proposed system (few-shot prompting with $LLaMA$), which were submitted to NumEval-2024. The organizers of NumEval-2024 conducted this human evaluation of the generated headlines. It is apparent from Table 2 that our proposed system (few-shot prompting with $LLaMA$) achieved the top rank in recommending 100 news.

5 Error Analysis

This study also conducts an error analysis to examine the strengths and weaknesses of large language models (LLMs) across different setups for numeral-aware headline generation. Through this analysis, we aim to identify patterns of errors and limitations inherent in the models, providing valuable insights into areas for improvement and optimization in future model development and training methodologies. We selected fifty news headline-body pairs, where numeral figures in the headline are crucial for accurately representing the news content. Our examination of the generated headlines by the models revealed the following insights: (i) $RoBERTa$ the model generates a headline, which is representative of the news body, however in some instances is failed to consider the numeric value for headline generation. Consequently, $RoBERTa$ is deemed unsuitable for numeral-aware headline generation. However, fine-tuning the $RoBERTa$ model enhances generated headline quality, which is also evident by the performance comparison between its *Pretrained* and *Fine-tuned* setups. (ii) The $BART$ models, whether in the *Pretrained* or *Fine-tuned* setups, demonstrate proficiency in generating efficient headlines that include valid numeric values. However, it is noteworthy that the inclusion of valid numeric values in headlines is more prevalent in the fine-tuned models compared to those without fine-tuning. (iii) The $T5$ models, in both the *Pretrained* and *Fine-tuned* setups, consistently produced headlines with more efficient and valid numerical values compared to $RoBERTa$, $BART$, and GPT. This indicates that $T5$ models are particularly more effective in numeral aware headline generations. (iv) The $LLaMA$ model stands out for its ability to generate accurate and efficient headlines containing valid numerical values when compared to $RoBERTa$, $BART$, GPT, and $T5$. This suggests that the $LLaMA$ model excels in incorporating precise numeric information into its generated headlines, surpassing other models in this aspect. Figure 2 presents an example' comparison between headline generated by fine-tuned $T5$ model and headline generated by $T5$ with three shot prompt. From Figure 2, it is apparent that the headline generated by $T5$ with three three-shot prompts better represents the news body compared to the headline generated by the fine-tuned T5 model. This further validates our claim that a few-shot prompt helps the LLMs generate headlines.

6 Conclusion and Future work

This study conducted an empirical research on LLMs for numeral-aware headline generation and proposed a few shots prompting with LLMs for numeral-aware headline generation. We conduct our experiments on *NumHG* and test set data provided by the organizer of NumEval-2024. Our experimental results suggest that finetuning LLMs over *NumHG* dataset improves the performance of numeral-aware headline generation. Further, few shot prompting with LLMs outperform fine-tuned LLMs for numeral-aware headline generation. This study identifies prompt tuning using LLMs for numeral-aware headline generation.

References

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *arXiv preprint arXiv:2310.10449*.

Wang Chen, Hou Pong Chan, Piji Li, and Irwin King. 2020. Exclusive hierarchical decoding for deep keyphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1095–1105, Online. Association for Computational Linguistics.

Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver. 2017. Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 56–61, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. Harnessing the power of llms: Evaluating human-ai text co-creation through the lens of news headline generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3321–3339.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on twitter? In *Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science*, pages 179–192.

Jian-Tao Huang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Numhg: A dataset for number-focused headline generation. *arXiv preprint arXiv:2309.01455*.

Sujit Kumar, Durgesh Kumar, and Sanasam Ranbir Singh. 2023. Gated recursive and sequential deep hierarchical encoding for detecting incongruent news articles. *IEEE Transactions on Computational Social Systems*.

Sujit Kumar, Gaurav Kumar, and Sanasam Ranbir Singh. 2022. Detecting incongruent news articles using multi-head attention dual summarization. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 967–977.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yutao Luo, Menghua Lu, Gongshen Liu, and Shilin Wang. 2022. Few-shot table-to-text generation with prefix-controlled generator. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6493–6504.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Timo Schick and Hinrich Schütze. 2021. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402.

Shengli Song, Haitao Huang, and Tongxiao Ruan. 2019. Abstractive text summarization using lstm-cnn based deep learning. *Multimedia Tools and Applications*, 78:857–875.

Percy H Tannenbaum. 1953. The effect of headlines on the interpretation of news stories. *Journalism Quarterly*, 30(2):189–197.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Daphne van Zandvoort, Laura Wiersema, Tom Huibers, Sandra van Dulmen, and Sjaak Brinkkemper. 2023. Enhancing summarization performance through transformer-based prompt engineering in automated medical reporting. *arXiv preprint arXiv:2311.13274*.

Wei Wei and Xiaojun Wan. 2017. Learning to identify ambiguous and misleading news headlines. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4172–4178.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.