Team NP_PROBLEM at SemEval-2024 Task 7: Numerical Reasoning in Headline Generation with Preference Optimization

Pawan Kumar Rajpoot *
SCB DataX, Thailand
pawankumar.rajpoot@data-x.ai

Nut Chukamphaeng *
SCB DataX, Thailand
nut.chukamphaeng@data-x.ai

Abstract

While large language models (LLMs) exhibit impressive linguistic abilities, their numerical reasoning skills within real-world contexts remain under-explored. This paper describes our participation in a headline-generation challenge by Numeval at Semeval 2024, which focused on numerical reasoning. Our system achieved an overall top numerical accuracy of 73.49% on the task. We explore the system's design choices contributing to this result and analyze common error patterns. Our findings highlight the potential and ongoing challenges of integrating numerical reasoning within large language model-based headline generation.

1 Introduction

The capacity to understand and manipulate numerical information within natural language text is essential for various NLP applications. Tasks such as news summarization, report generation, and the creation of data-driven narratives increasingly rely on the accurate interpretation and generation of numerical expressions. SemEval 2024 Task 7 (Chen et al., 2024) addresses these challenges through two intriguing subtasks: numerical headline generation and numerical headline number fill-in-the-blanks.

Generating numerical headlines necessitates models capable of synthesizing a succinct and attention-grabbing title that accurately reflects a news article's core numerical quantities and trends. Conversely, the fill-in-the-blanks subtask tests the model's ability to comprehend numerical relationships and infer the missing value to complete a provided headline. These tasks present a complex intersection of numerical reasoning and natural language generation/understanding.

Existing text generation and numerical understanding work often leverage sequence-to-sequence architectures and specialized pre-trained language models. However, SemEval 2024 Task 7's emphasis on numerical reasoning within headlines creates a distinct demand for techniques capable of accurately grounding representations of numbers and quantities within the linguistic context. This paper describes our approach to SemEval 2024 Task 7. We worked on both tasks separately and created two separate models. We used techniques such as parameter-efficient fine-tuning of large language models and then doing Direct Preference Optimization on top to align models better.

The remainder of this paper proceeds as follows. Section 3 reviews related work in numerical reasoning and headline generation. Section 4 details our models and methodology. Section 4 presents our experimental evaluation of the SemEval 2024 Task 7 dataset and includes a thorough analysis of our results. Finally, Section 5 summarizes our findings and outlines potential future research directions.

2 Background and Related Work

Headline generation within NLP has a rich history, evolving from early extractive techniques towards modern abstractive generation methods. Initial extractive approaches primarily focused on selecting the most salient sentences from the source document to compose the headline (Dorr et al., 2003) (Erkan and Radev, 2011; Mihalcea and Tarau, 2004). These methods offered interpretability but lacked the fluency and novelty often desired in generated headlines. The advent of deep learning and sequence-to-sequence models enabled abstractive headline generation, empowering models to synthesize new phrases and expressions (Rush et al., 2015; Nallapati et al., 2016). Attention mechanisms (Bahdanau et al., 2015) proved pivotal in aligning source text and headline generation. Recent advancements in Generative AI have led to significant improvements in this field with state-of-the-

[*] Equal Contribution

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 716–720
June 20-21, 2024 ©2024 Association for Computational Linguistics

art results (Zhang et al., 2019). *GSum* (Dou et al., 2020), for example, initially performs extractive summarization and then incorporates the extractive summaries into the input for abstractive summarization. *SEASON* (Wang et al., 2022) adopts a dual approach, learning to predict the informativeness of each sentence and using this predicted information to guide abstractive summarization Notably, most of these works focus on the selection of words and the structure of sentences.

3 Numeval

Numeval is part of Semeval 2024; the task we focused on and worked on requires models to generate concise and informative headlines that accurately reflect the core numerical information in news articles. Systems must demonstrate an understanding of how numbers convey meaning and should prioritize the most relevant numerical aspects for inclusion in the headline.

Subtask 1: Numerical Reasoning - models are required to compute the correct number to fill the blank in a news headline.

Subtask 2: Abstractive Headline Generation - models must construct a headline based on the provided news; this headline should incorporate the numerical reasoning within. The organizers released a novel dataset designed to facilitate research on numeral-aware headline generation. The NumHG (Huang et al., 2023) dataset addresses the issue of inaccurate numeral generation in headline creation. It provides over 27,000 news articles with detailed annotations designed to facilitate the development of models that accurately understand and summarize numerical information. For subtask 1, each data point has an answer operator added, which signifies how the numerical answer is obtained, which includes Copy (direct retrieval), Trans, Span, Round, Paraphrase, Add, Subtract, Multiply, and Divide. Meanwhile, subtask 2 requires the model to generate complete headlines from given news content.

News:
At least **30** gunmen burst into a drug rehabilitation center in a Mexican border state capital and opened fire, killing **19** men and wounding **four** people, police said. Gunmen also killed **16** people in another drug-plagued northern city. The killings in Chihuahua city and in Ciudad Madero marked one of the bloodiest weeks ever in Mexico and came just weeks after authorities discovered **55** bodies in an abandoned silver mine, presumably victims of the country's drug violence. More than **60** people have died in mass shootings at rehab clinics in a little less than **two** years. Police have said **two** of Mexico's **six** major drug cartels are exploiting the centers to recruit hit men and drug smugglers, ...

Headline (Question): Mexico Gunmen Kill ____

Answer: 35

Annotation: Add(19,16)

Figure 1: Numeval Task 3-1 examples.

News:
(Apr 18, 2016 1:02 PM CDT) Ingrid Lyne, the Seattle mom allegedly murdered while on a date, left behind three daughters—and a GoFundMe campaign set up to help the girls has raised more than $222,000 so far, Us reports. A friend of the family set up the campaign, and says that all the money raised will go into a trust for the girls, who are ages 12, 10, and 7. Lyne's date was charged with her murder last week.

Headline: $222K Raised for Kids of Mom Dismembered on Date

Figure 2: Numeval SubTask 2 examples.

4 Methodology

4.1 subtask 1

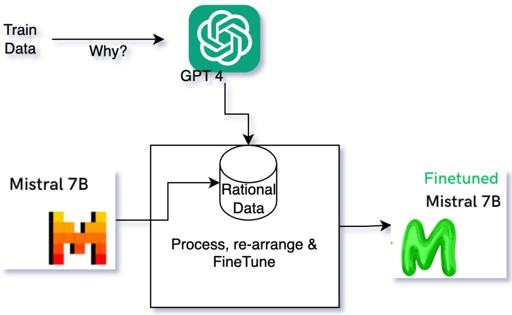

Figure 3: Overall Stage 1 train subtask 1

For this subtask, we start by passing the passage and question in the blank statement with an aligned answer to GPT 4 and ask it to generate a rationale for this given answer. The below prompt is used in this process.

```
PASSAGE: ARTICLE-HERE
QUESTION: FILL_IN_THE_BLANKS-HERE
WHY ANSWER TO THIS IS ANSWER-HERE ?
EXPLAIN\nRESULT:
```

Once we get the reasoning from this module, we restructure the training data in the following manner.

```
PASSAGE: ARTICLE-HERE
QUESTION: FILL_IN_THE_BLANKS-HERE
```

```
WHY ANSWER TO THIS IS ANSWER-HERE ?
REASON: GPT_RATIONALE_HERE
```

We use this to train our main model for subtask 1.

4.2 subtask 2

1) **Numerical Reasoning**: The model must demonstrate fluency in numerical calculations.
2) **Headline Matching**: Generated headlines must stylistically align with the data.

We worked on an end-to-end solution leveraging a Large Language Model (LLM) to address these challenges. First, we fine-tune the LLM to enhance its mathematical reasoning capabilities. This approach targets accurately interpreting numerical data and producing suitable headlines.

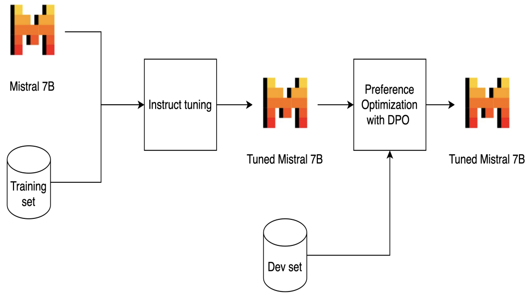

Figure 4: Overall workflow for subtask 2

4.3 Fine-tuning

For both tasks separately, we selected Mistral-7B (Jiang et al., 2023) as our trained LLM based on its strong performance on diverse benchmarks, including those focused on numerical reasoning benchmarks like GSM8K, which suggests a solid foundation for further fine-tuning on our specific numerical headline generation task.

As illustrated in Figure 5, Mistral-7B achieve a lower fine-tuning loss than BART (Lewis et al., 2019), a state-of-the-art text summarization model. This reinforces its suitability for our task.

For efficient fine-tuning, we employ Parameter Efficient Fine Tuning (PEFT). Due to memory constraints, we employed 4-bit QLoRA(Dettmers et al., 2023) quantization (cite reference) with a rank of 128 and an alpha of 256. This quantization technique was applied specifically to the self-attention Query, Key, and Value matrices along with the Linear layers of the model. To optimize the process, we used gradient accumulation (steps=2), a paged 32-bit Adamw optimizer, a cosine learning rate schedule (LR=2e-5), a decay rate of 0.01, and a short 5-step warmup period. The entire fine-tuning process was facilitated using the axolotl library. This technique reduces the model's memory footprint while minimizing performance degradation. This is particularly advantageous when working with LLMs.

Prompt template

Given the news article, please write an appropriate headline
{news content}

Headline:

4.4 Direct Preference Optimization (DPO)

For both subtasks, we further aligned our fine-tuned models to learn better using the dev set. We did not use the dev data split in the first train stage. While aligning, we used dev data to first run through the model. We realigned the fine-tuned model with incorrect outcome results, that is, the dev results where the predicted number in the generated headline was incorrect or rejected. We still use the rationale (for subtask 1) for DPO, while DPO training for the second subtask only contains the predicted headline(wrong/rejected) and the correct/choosen headline. One example of the DPO train data for subtask 1 is below.

PASSAGE:
Stocks made gains today, extending a winning streak into its fourth day, MarketWatch reports. Merck rose 12.5% on announcement of its merger with Schering-Plough, while General Motors built on recent gains with a 22.9% jump. The Dow closed up 53.92 at 7,223.98.,
QUESTION:
"Dow Up ____, Gains 9% for Week
Choosen:54 REASON: rounding off to nearest integer
Rejected: 53.92 REASON:copy from text

To align the generated fill-in-the-blanks/headlines style with the target dataset, we utilize the Direct Preference Optimization (DPO) alignment technique (Rafailov et al., 2023). Direct Preference Optimization (DPO) is a novel approach for aligning large language models (LLMs) with human preferences. Unlike traditional methods that rely on reward models and reinforcement learning, DPO leverages human feedback through preferred and dispreferred outputs to directly train the LLM. This simplifies the training process and avoids the complexities of reward model design. This helps us

Table 1: Automatic evaluation results subtask 2.

	Num Acc.			ROUGE			BERTScore			MoverScore
	Overall	Copy	Reasoning	1	2	L	P	R	F1	
ClusterCore	38.233	51.571	13.942	33.467	11.837	28.927	31.876	42.232	37.026	56.405
Noot Noot	38.393	57.481	3.6331	31.47	11.139	27.284	25.389	43.977	34.539	55.559
Infrrd.ai	65.840	68.354	61.263	46.789	22.36	42.095	51.005	47.260	49.134	59.731
hinoki	62.347	66.284	55.177	43.072	19.719	38.999	47.223	43.444	45.342	58.711
Challenges	72.956	**82.170**	56.176	31.220	12.235	26.859	19.530	47.559	33.132	55.362
NCL_NLP	62.122	65.536	55.904	43.506	19.388	38.878	46.402	45.039	45.734	58.861
YNU-HPCC	69.044	73.018	61.807	**48.852**	**24.681**	**44.175**	**51.553**	**50.095**	**50.381**	**60.551**
NoNameTeam	55.715	57.681	52.134	40.646	17.261	35.745	44.256	40.387	42.324	57.736
np_problem (ours)	**73.487**	76.908	**67.257**	39.816	17.577	34.339	27.800	48.557	37.816	57.024

fine-tune the model beyond numerical correctness to produce stylistically suitable headlines.

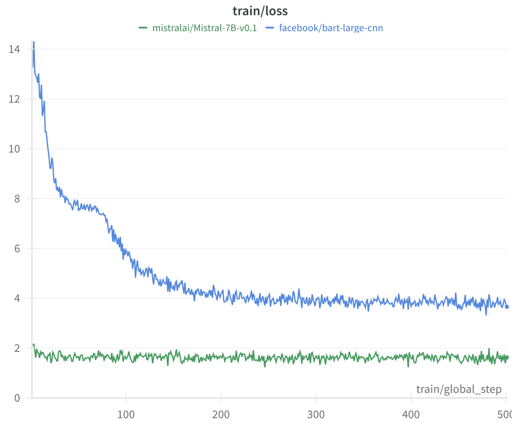

Figure 5: Comparing fine-tuning losses between BART and Mistral-7B

5 Results

For subtask 1, we evaluate our model on two fill-in-the-blank types. One accuracy on Copy, where the answer is directly copied from the article. Second is Reasoning, which includes different reasoning techniques such as addition, subtraction, multiplication, and paraphrasing.

Table 2: Accuracy on tasks type subtask 1

Copy	Reasoning
0.922	0.784

For comparative semeval two-stage evaluations, we scored 0.89 in the open and 0.86 in the hidden stage, respectively.

Table 3: Comparitive num accuracy on subtask 1

Team	Open-Score	Hidden-Score
CTYUN-AI	**0.95**	**0.95**
zhen qian	0.94	0.94
YNU-HPCC	0.93	0.94
NP-Problem(ours)	0.89	0.86

For subtask 2 we performed best in numerical accuracy overall and reasoning scores, we out-shined in reasoning accuracy as difference between 1st and 2nd rank was around 7 points. We open-source our final models on Huggingface [1].

6 Limitations

Since our method is based on 7B LLM, our performance is capped by this model's ability to draw rationale for the numerical reasoning. This limitation is in line with the hardware resource as well; We used an RTX 4090 24GB GPU-based machine for our work, which can load and fine-tune the models with upto 7-10 B parameters as well.

Conclusion

We present a modular solution to the numeral problem with an alignment module to increase the model's ability to understand numerical reasoning across both tasks. We thank the organizing committee of SemEval-2024, along with the task-setting team of Numeval, for allowing us to work on this problem.

[1] https://huggingface.co/lingjoor/numeval-task7-1, https://huggingface.co/lingjoor/numeval-task7-2

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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)*. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 1–8.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. *CoRR*, abs/2010.08014.

Günes Erkan and Dragomir R. Radev. 2011. Lexrank: Graph-based lexical centrality as salience in text summarization. *CoRR*, abs/1109.2128.

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*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *CoRR*, abs/1509.00685.

Fei Wang, Kaiqiang Song, Hongming Zhang, Lifeng Jin, Sangwoo Cho, Wenlin Yao, Xiaoyang Wang, Muhao Chen, and Dong Yu. 2022. Salience allocation as guidance for abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6094–6106, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.