NCL_NLP at SemEval-2024 Task 7: CoT-NumHG: A CoT-Based SFT Training Strategy with Large Language Models for Number-Focused Headline Generation

Junzhe Zhao[1] and **Yingxi Wang**[2] and **Huizhi Liang**[3] and **Nicolay Rusnachenko**[3]
[1]Hangzhou Zero Matrix Intelligence Co., Ltd., China
[2]Huawei Technologies Co., Ltd., China
[3]School of Computing, Newcastle University, Newcastle upon Tyne, UK
zhaojunzhe_bit@163.com wangyingxiclaire@163.com
huizhi.liang@newcastle.ac.uk, rusnicolay@gmail.com

Abstract

Headline Generation is an essential task in Natural Language Processing (NLP), where models often exhibit limited ability to accurately interpret numerals, leading to inaccuracies in generated headlines. This paper introduces CoT-NumHG, a training strategy leveraging the Chain of Thought (CoT) paradigm for Supervised Fine-Tuning (SFT) of large language models. This approach is aimed at enhancing numeral perception, interpretability, accuracy, and the generation of structured outputs. Presented in SemEval-2024 Task 7 (task 3): Numeral-Aware Headline Generation (English), this challenge is divided into two specific subtasks. The first subtask focuses on numerical reasoning, requiring models to precisely calculate and fill in the missing numbers in news headlines, while the second subtask targets the generation of complete headlines. Utilizing the same training strategy across both subtasks, this study primarily explores the first subtask as a demonstration of our training strategy. Through this competition, our CoT-NumHG-Mistral-7B model attained an accuracy rate of 94%, underscoring the effectiveness of our proposed strategy, detailed in our project repository[1].

1 Introduction

Headline Generation is a key task in the field of Natural Language Processing (NLP), aimed at condensing the content of a given article into a concise, accurate, and information-rich single-sentence headline. This process requires not only an understanding of the article's core content but also the ability to creatively express this content (Matsumaru et al., 2020). Recently, Huang et al. (2023) conducted an in-depth analysis of the application of models (Lewis et al., 2019; Liu et al., 2022; Raffel et al., 2020; Wang et al., 2022a; Zhang et al., 2020) in the task of headline generation, revealing limitations of these models in processing numerical information. They identified that inaccuracies in the use of numbers significantly contribute to errors in generated headlines, a particularly critical issue in news headline generation where numbers often carry key information. To further explore the issue of numerical accuracy in news headlines, Huang et al. (2023) introduced a new dataset, NumHG, focused on the accuracy of numerical usage within news headlines. Their analysis revealed that news headline generation typically involves nine distinct methods for handling numbers—Copy, Translate, Round, Paraphrase, Add, Subtract, Divide, Multiply, and Span—each varying in complexity. These techniques enhance the interpretability and clarity of the headline generation process, showcasing a sophisticated blend of precision and creativity in distilling numerical information. Based on these insights, Chen et al. (2024) designed two independent tasks: the first requires models to mask numbers in given news articles and their headlines, then to accurately predict the masked numbers; the second involves generating news headlines with accurate numerical information based on the provided news content.

In the domain of NLP, Large Language Models (LLMs) have gained recognition for their capability to execute a wide array of tasks, including text generation, summarization, and question answering, using straightforward instructions. This demonstrates their remarkable versatility (Guo et al., 2023; Sahoo et al., 2024). To further enhance the adaptability of LLMs, fine-tuning techniques (Zhang et al., 2023) have been extensively applied, improving model performance on specific tasks while preserving a wide scope of application. LLMs typically utilize a decoder-only architecture (Radford et al., 2018) and adopt primarily two strategies for task-specific challenges: prompt engineering and fine-tuning. Prompt engineering enables the direct execution of tasks

[1]https://github.com/GavinZhao19/CoT-NumHG

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 261–269
June 20-21, 2024 ©2024 Association for Computational Linguistics

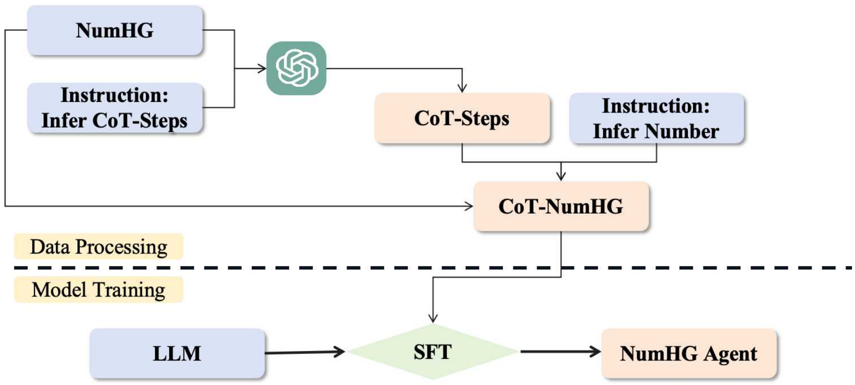

Figure 1: CoT-Based SFT Training Strategy Framework: The framework comprises two main parts: data processing and model training. In the first phase, data processing involves two steps. The first step combines specific instructions with the NumHG dataset, and through knowledge distillation using GPT-3.5-Turbo, new CoT-Steps are generated. These steps are then integrated with the corresponding instructions and the original dataset to produce the CoT-NumHG dataset. In the second phase, the CoT-NumHG dataset is utilized for the full-parameter SFT of the base model.

through various techniques, such as zero-shot (Radford et al., 2019), few-shot (Brown et al., 2020), chain of thought (CoT) (Wei et al., 2022), CoT with self-consistency (Wang et al., 2022b), and tree of thought (Yao et al., 2024), without requiring additional training. This underscores the models' ability to quickly adapt to new tasks by leveraging existing knowledge. Fine-tuning, via further training, refines the models' performance on specific tasks, particularly through Supervised Fine-Tuning (SFT) methods. To improve SFT efficiency, Parameter-Efficient Fine-Tuning (PEFT) techniques (Hu et al., 2023) including LoRA (Hu et al., 2021), prompt-tuning (Lester et al., 2021), and prefix-tuning (Li and Liang, 2021) have been introduced. These significantly enhance the models' adaptability and the quality of outputs for specific tasks without substantially increasing the model size or computational demands. This approach not only preserves the versatility of LLMs but also boosts their output quality and the ability to generate structured outputs in specific domains. Despite these approaches achieving certain levels of performance enhancement, there remains room for improvement in perceiving numerical information, reasoning ability, and generating structured outputs (Ouyang et al., 2023). Particularly in the task of news headline generation, reliance solely on prompt engineering may lead to uncontrollable outputs and insufficient structuring. Meanwhile, SFT, despite its ability to improve performance, shows limitations in the interpretability of the reasoning process and suffers from attention decay, potentially leading to the omission of important information.

To address these challenges, we propose a training strategy based on the CoT approach, designed to significantly enhance LLMs in the task of number-focused headline generation. Our method consists of two key components. First, drawing on the concept of knowledge distillation (Dasgupta et al., 2023), we utilize GPT-3.5-Turbo (Brown et al., 2020) and instructions to process the original NumHG dataset, generating a series of reasoning steps. Given the issue of attention decay when handling long-distance information (Xiao et al., 2023), we created a new CoT-NumHG dataset by combining the question statement with reasoning steps. This process aims to bolster the model's attention mechanism and improve the interpretability of the reasoning process (Wang et al., 2023). Secondly, we selected three LLMs as base models and performed full- parameters SFT using the constructed CoT-NumHG dataset on these base models. Through this approach, we not only significantly improved performance on the specific task, but also optimized structured outputs while maintaining the models' versatility. Our research contributions are threefold:

1. Based on the NumHG dataset, we developed the CoT-NumHG dataset, enhancing model interpretability and structured output capabilities. Importantly, we introduce a dataset construction technique specifically designed for the CoT-NumHG.

2. We demonstrate the enhancement of model performance through task-oriented SFT train-

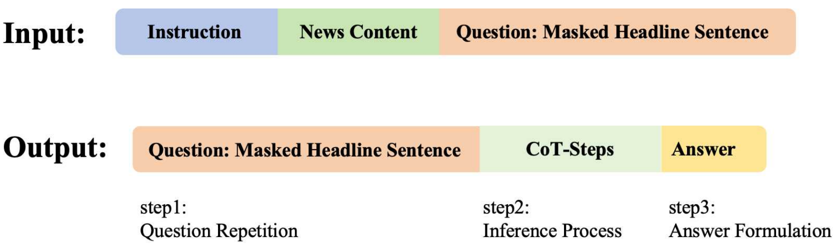

Figure 2: CoT-NumHG Dataset: The input of the dataset consists of three parts: Instruction, News Content, and Masked Headline Sentence (Question). The output is comprised of three components: Question Repetition, Inference Process, and Answer Formulation.

ing on the CoT-NumHG dataset across three base models, significantly improving news headline generation while maintaining general-purpose capabilities.

3. Through ablation studies, we demonstrated that the CoT-based training strategy effectively boosts the model's performance.

2 CoT-Based SFT Training Strategy Design

The training strategy of this study is divided into two main parts, as shown in Figure 1: the construction of the CoT-NumHG dataset and model training. Initially, through a knowledge distillation strategy, we enhanced the original dataset to improve the model's interpretability in handling the task of generating news headlines. Subsequently, the selected base LLMs were trained using full-parameter SFT techniques to achieve performance optimization and structured output for specific tasks.

2.1 CoT-Dataset Generation

To enhance the model's understanding of the relationships among news content, headline sentences, and answers, we employed a knowledge distillation approach during the data construction and optimization phases. Utilizing the original NumHG dataset and instructions, we generated inference processes through the GPT-3.5-Turbo model, termed CoT-Steps. CoT-Steps consist of three steps:

Step 1: **Identifying the Relevant Information**: This involves analyzing semantic relevance to pinpoint sentences in news articles that are closely related to the masked headline sentences and answers. This step ensures that the selected sentences are crucial for understanding the content of the news articles and for generating headlines.

Step 2: **Interpreting the Numerical Information**: For each identified key sentence, its direct numerical relevance to the generation task and the reasons for its selection are interpreted.

Step 3: **Choosing and Applying the Math Method**: For the numerical information in key sentences, appropriate methods are used for transformation and completion to accurately reflect in the generated headline sentences while maintaining logical consistency and accuracy.

This approach aims to bolster the model's data understanding and information processing capabilities by emulating the human problem-solving thought process, thereby enhancing attention scores and interpretability. Then, We integrated the reasoning process (CoT-Steps) into the training set to build a dataset specifically for SFT. Figure 2 shows that the input of this training set includes the instruction, news content, and the masked headline sentence; the output covers question restatement, CoT-Steps, conversion methods, and the final answer. This design aims to train the model to generate answers following given logical steps, thereby improving the accuracy and reliability of the generated results.

2.2 Model Training

For the model training part, we selected three large language models with a decoder-only architecture: ChatGLM3-6B (Du et al., 2022), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), and Zephyr-7B-Beta (Tunstall et al., 2023), for full-parameter fine-tuning. These models were chosen for their outstanding performance in text generation and comprehension, as illustrated in benchmarks comparing their capabilities to other LLMs of similar size (Zheng et al., 2023). During the fine-tuning process, we focused on enhancing the models' comprehensive understanding and generation capabilities, especially in handling news headlines that contain numerical information. Through meticulous training methods, we ensured that the models could achieve higher performance on specific tasks.

3 Data Construction

3.1 CoT-Steps Generation

The primary source of the NumHG dataset is Newser[2], a news aggregation platform that provides headline news from both American and international media. News articles typically contain between 200 and 300 words. The entire NumHG dataset consists of news articles with titles that integrate numerical information, comprising 21,157 news articles for training and 2,572 for validation, totaling 23,729 articles. The data includes four keys: news, masked headline sentence, answer, and calculation. Initially, we employ a few-shot approach to distill the reasoning steps. The complete prompt given to the model comprises three parts: instruction, news, masked sentence (question), calculation, and answer. Figure 3 shows the instruction content used.

Instruction: Infer CoT-Steps

I will give you article, question sentence, chosen mathematical methods and answer. You need think and write three step sentences to explain why I chose the specific mathematical methods to arrive at the answer. Three step sentences are required:
1. Identifying the Relevant Information: write your thought
2. Interpreting the Numerical Information: write your thought
3. Choosing and Applying the Math Method: write your thought

#All mathematical methods
- Copy(v): Copy v from the article (e.g. Copy(60) ans:60)
- Trans(e): Convert e into a number (e.g. Trans(three) ans:3)
- Paraphrase(v0,n): Paraphrase the form of digits to other representations (e.g. Paraphrase(5,000,K) ans: 5)
- Round(v0,c): Hold c digits after the decimal point of v0 (e.g. Round(305.45,0) ans: 305)
- Subtract(v0,v1): Subtract v1 from v0 (e.g. Subtract(Sep 5,July 8) ans: 59)
- Add(v0,v1): Add v0 and v1 (e.g. Add(2,2) ans: 4)
- Span(s): Select a span from the article (e.g. Span(top draft pick) ans: 1)
- Divide(v0,v1): Divide v0 by v1 (e.g. Divide(700,2) ans: 350)
- Multiply(v0,v1): Multiply v0 and v1 (e.g. Multiply(0.65,100) ans: 65)

Figure 3: Instruction Prompt of Inferring CoT-Steps

Table 1 lists the detailed examples from the NumHG dataset, along with the inference steps obtained using knowledge distillation techniques. These steps not only reveal the key reasoning pathways in the news headline generation process but also provide clear guidance for models to more effectively handle numerical information and generate structured headlines.

3.2 CoT-NumHG Datasets Generation

In preparing the CoT-NumHG datasets for model training, we have adopted an approach inspired by

Table 1: Example of NumHG datasets and CoT-Steps

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.
Masked Headline (Question): $____K Raised for Kids of Mom Dismembered on Date
Calculation: Paraphrase(222,000,K)
Answer: 222
CoT Steps: 1. Identifying the Relevant Information: The relevant information in this question is the amount of money raised for the kids of the mom who was dismembered on a date. 2. Interpreting the Numerical Information: The numerical information given is $222,000. 3. Choosing and Applying the Math Method: I chose the Paraphrase method to convert the numerical information from the form of digits to other representations. By paraphrasing 222,000 as K, I am representing the amount as 222 thousand dollars.

the methodologies outlined in the "Never Lost in the Middle" study (Junqing et al., 2023). This strategy ensures that the model can efficiently identify and utilize key information within extended texts. Our dataset is specifically tailored to enhance the models' attention mechanisms, thereby improving their reasoning capabilities and their ability to produce structured outputs for complex tasks.

The dataset (see the example in Table 4) is meticulously organized, comprising three elements in its input section: instruction (as referenced in Figure 4), news content, and masked headline sentences (posed as questions). This configuration is designed to keep the model focused during the processing of information and to encourage a logical and structured approach to output generation. In the output section, we employ a stepwise methodology to formulate answers. Initially, the model is instructed to repeat the question, a step that not only deepens its understanding of the query, but also counteracts attention drift by enhancing attention scores. Following this, the model boosts the interpretability of its reasoning process by executing CoT-Steps, which involve generating a sequence of intermediate reasoning steps. These steps are designed to mimic human problem-solving processes, thereby clarifying the model's reasoning pathway. Ultimately, the model presents the final answer, ensuring the creation of a structured and precise headline while preserving the integrity of the news content. Through this dataset design, our

[2] https://www.newser.com/

objective is for the model to demonstrate enhanced accuracy and interpretability in news headline generation tasks, in addition to maintaining consistent performance when managing information over long distances.

Instruction: Infer Number

Your task is to read and understand Article and Question. Focus on the numerical information in the article. Choose the suitable mathematical methods to arrive at the answer. The mathematical methods include copy a value from the article, trans a word into a number, paraphrase the form of digits, round a number, subtract, add, select a span from the article, divide, and multiply. Think and write three step sentences to fill in the blank in the question sentence:
1. Identifying the Relevant Information
2. Interpreting the Numerical Information
3. Choosing and Applying the Math Method

Finally, output the answer in the blank. Remember, the final answer is immediately followed by "Answer:"!

Figure 4: Instruction Prompt of Inferring Number

By incorporating these strategies into our dataset design, we aim to equip models with the ability to achieve superior accuracy and interpretability, particularly in tasks such as news headline generation, while also ensuring steady performance in the processing of long-range information. This methodology embodies the proverb "the pen is mightier than memory," emphasizing the importance of a structured and considered compilation of training data to bolster model capabilities in the NLP field.

4 Experiments and Results

4.1 Experiments Settings

Ultimately, we constructed the CoT-NumHG dataset, which comprises 23,729 samples. To ensure the randomness and fairness of the experiments, the dataset was randomly shuffled and 3,000 samples were selected as the validation set, leaving 20,729 samples for the training set. We chose three state-of-the-art LLMs as our base models: ChatGLM3-6B, Mistral-7B-Instruct-v0.2, and Zephyr-7B-Beta. SFT was performed on these base models.

For each base model, training was conducted over 2 epochs on 16 $\times$ A100 GPUs (80GB). During training, we set the batch size to 32, the learning rate to $1 \cdot 10^{-6}$, and the warmup ratio to 0.01. A cosine annealing schedule was used for learning rate adjustment, and training was accelerated using `bfloat16` precision. Cross-entropy loss was used to compute the loss. We monitored the evaluation loss (eval loss) of each model on the validation set during training. ChatGLM3-6B achieved the lowest eval loss (0.176) at epoch 1.95, with a corresponding training loss (train loss) of 0.121. Mistral-7B-Instruct-v0.2 reached its lowest eval loss (0.153) at epoch 1.81, with a training loss of 0.111. Zephyr-7B-Beta achieved the lowest eval loss (0.151) at epoch 1.96, with a training loss of 0.110.

Table 2: Example of CoT-NumHG-Mistral-7B's output

News: (Jul 20, 2009 3:41 PM CDT) The special inspector overseeing Treasury's TARP program says federal assistance to banks and other financial entities could end up costing taxpayers $23.7 trillion, Bloomberg reports. Aside from the $700 billion bailout, Neil Barofsky says in testimony prepared for told Congress tomorrow, other trillion-dollar federal programs could balloon...
Masked Headline (Question): TARP Watchdog: Bailouts May Cost US $____T
CoT-NumHG-Mistral-7B Output: To fill in the blank in the question sentence: TARP Watchdog: Bailouts May Cost US $____T. 1. Identifying the Relevant Information: The relevant information is the cost of the bailouts, which is given as $23.7 trillion. 2. Interpreting the Numerical Information: The numerical information is in the form of a decimal, with two digits after the decimal point. 3. Choosing and Applying the Math Method: I chose the math method Round(23.7,0) to round the number to the nearest whole number.This is because the question asks for the answer to be filled in the blank as a whole number, and rounding 23.7 to the nearest whole number gives us 24. Summary: math methods: Round(23.7,0) Answer: 24

To ensure the accuracy of the results, we performed ablation studies. Specifically, we used only the instruction prompt (also CoT) to generate outcomes with three base models, along with a benchmark model, GPT-3.5-Turbo. Additionally, we trained these three base models solely with the NumHG dataset, comparing the results against those trained using the CoT-NumHG dataset. The vLLM framework (Kwon et al., 2023) was consistently employed for inference.

4.2 Results

The competition provided a dataset containing 4,921 samples. The results showed that the first-place participant achieved an accuracy of 95%, while the participants in second to fourth places all reached an accuracy of 94%. Our team's submission, the CoT-NumHG-Mistral-7B model, also achieved an accuracy of 94% in this task, demonstrating the effectiveness of the CoT training strategy in enhancing model performance. Table 2 presents an example of the result.

To further analyze model performance, we conducted ablation studies across all models. We observed incremental improvements in accuracy, starting from models prompted solely by CoT, progressing through those trained on the NumHG dataset, and culminating with those trained on the CoT-NumHG dataset. The accuracy of the CoT-NumHG-Mistral-7B model increased from 0.58 to 0.94, surpassing the untrained baseline, and improved from 0.73 to 0.94 compared to NumHG-Mistral-7B, showcasing significant improvements. This indicates that the CoT-Based SFT training strategy not only enhances model accuracy, but also improves the stability of generating structured outputs. Models without fine-tuning produce less stable outputs, sometimes requiring manual intervention to identify generated answers. Furthermore, their reasoning processes exhibit a higher degree of interpretability.

Table 3: Accuracy of Different LLMs; the result of the final submission is bolded

Model Name	Accuracy
ChatGLM3-6B	0.51
Mistral-7B-Instruct-v0.2	0.58
Zephyr-7B-Beta	0.56
GPT-3.5-Turbo	0.74
NumHG-ChatGLM3-6B	0.62
NumHG-zephyr-7b	0.71
NumHG-Mistral-7B	0.73
CoT-NumHG-ChatGLM3-6B	0.83
CoT-NumHG-Zephyr-7B	0.90
CoT-NumHG-Mistral-7B	**0.94**

5 Conclusion and Future Work

In this paper, we have introduced a CoT-based SFT training strategy aimed at enhancing the performance of LLMs in the task of news headline generation. Initially, we constructed the CoT-NumHG dataset, based on the existing NumHG dataset through knowledge distillation techniques. By simulating the human thought process, this dataset enhances the interpretability of the reasoning path from problem to answer. Subsequently, we utilized the CoT-NumHG dataset to perform SFT on a selected baseline model and verified significant improvements in model performance through ablation studies. The competition results further validated the efficacy of our approach, with the CoT-NumHG-Mistral-7B model achieving an accuracy rate of 94%. However, a manual review of the competition outcomes revealed some uncertainties in the model's handling of numerical information in titles, such as the need for approximations. This indicates that there is still room for improvement in understanding numerical information and generating structured outputs.

Future work will focus on the following directions: further optimization of the dataset by deduplicating and enhancing data diversity to improve the model's generalization capabilities. This includes identifying and removing duplicate or low-quality data samples, as the current proportion of copied methods is excessively high. We will adjust the proportions through sampling to address this issue. To align the generated headlines more closely with the standards of human editors, we will explore constructing a dataset in DPO (Rafailov et al., 2024) format from incorrect generation outcomes.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Sayantan Dasgupta, Trevor Cohn, and Timothy Baldwin. 2023. Cost-effective distillation of large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7346–7354.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

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, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

He Junqing, Pan Kunhao, Dong Xiaoqun, Song Zhuoyang, Liu Yibo, Liang Yuxin, Wang Hao, Sun Qianguo, Zhang Songxin, Xie Zejian, et al. 2023. Never lost in the middle: Improving large language models via attention strengthening question answering. *arXiv preprint arXiv:2311.09198*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

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*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804*.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. *arXiv preprint arXiv:2005.00882*.

Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. 2023. Llm is like a box of chocolates: the non-determinism of chatgpt in code generation. *arXiv preprint arXiv:2308.02828*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Fei Wang, Kaiqiang Song, Hongming Zhang, Lifeng Jin, Sangwoo Cho, Wenlin Yao, Xiaoyang Wang, Muhao Chen, and Dong Yu. 2022a. Salience allocation as guidance for abstractive summarization. *arXiv preprint arXiv:2210.12330*.

Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. 2023. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Example of CoT-NumHG Datasets

Table 4: Example of CoT-NumHG Datasets

Input: Your task is to read and understand Article and Question. Focus on the numerical information in the article. Choose the suitable mathematical methods to arrive at the answer. The mathematical methods include copy a value from the article, trans a word into a number, paraphrase the form of digits, round a number, subtract, add, select a span from the article, divide, and multiply. Think and write three step sentences to fill in the blank in the question sentence: 1. Identifying the Relevant Information 2. Interpreting the Numerical Information 3. Choosing and Applying the Math Method Finally, output the answer in the blank. Remember, the final answer is immediately followed by Änswer:Ï #Article (Dec 25, 2020 2:25 PM) Around 1,000 British soldiers were spending Christmas Day trying to clear a huge backlog of truck drivers stuck in southeast England after France briefly closed its border to the UK, then demanded coronavirus tests from all amid fears of a new coronavirus variant. Even though an estimated 4,000 or so international truck drivers are spending yet another day cooped up in their cabs, some progress was evident Friday, with traffic around the English Channel port of Dover moving in an orderly fashion toward the extra ferries that were put on to make the short crossing across to Calais in northern France. Rail operator Eurotunnel was also back in action, the AP reports, offering a way back into France. The military personnel were directing traffic and helping a mass testing program for the drivers, who must test negative to enter France. French firefighters have been drafted to help the military test drivers for coronavirus. Poland's Territorial Defense Force also sent reinforcements to help with testing and food distribution. #Question Fill in the blank: British Troops Work to Clear ____K Trucks
Output: To fill in the blank in the question sentence: British Troops Work to Clear ____K Trucks. 1. Identifying the Relevant Information: The question asks for the number of trucks that the British troops are working to clear. 2. Interpreting the Numerical Information: The article mentions that there is an estimated 4,000 international truck drivers stuck in southeast England. 3. Choosing and Applying the Math Method: To arrive at the answer, I chose the Paraphrase method to convert the numerical information of 4,000 into a different representation. By paraphrasing 4,000 as K̈,Ï can represent the number of trucks as 4. Summary: math methods: Paraphrase(4,000,K) Answer: 4