Numeral-Aware Language Generation

Abhilash Neog *
Department of Computer Science
Virginia Tech
abhilash22@vt.edu

Aruj Nayak *
Department of Computer Science
Virginia Tech
aruj@vt.edu

Mridul Khurana *
Department of Electrical and Computer Engineering
Virginia Tech
mridul@vt.edu

Abstract

The proliferation of Large Language Models (LLMs) has significantly impacted various sectors, ranging from customer service to education. Despite their adeptness in generating human-like text, these models face challenges in tasks requiring nuanced numerical understanding, such as interpreting and processing numerical data within text. This limitation is particularly evident in generating headlines for news articles that include numerical information, a task requiring both semantic comprehension and numerical precision. Despite the emergence of various techniques to improve LLMs' performance in numerically intensive tasks, these techniques often lack generalizability across different domains. Our work specifically explores the capabilities of LLMs in the context of generating numerically-aware headlines for news articles, as part of the upcoming SemEval'24 workshop's "Numeral-Aware Headline Generation (English)" challenge. Our study presents interesting insights into the current abilities and limitations of LLMs in this domain, contributing to the ongoing discourse in natural language processing. The code can be found at *https://github.com/mridulk97/numeval*

1 Introduction

Large Language Models (LLMs) like OpenAI's GPT series have become incredibly popular and are changing many areas of our lives. These models are good at creating and understanding text that sounds like it was written by a human. They have been integrated into a wide range of applications, revolutionizing the way we interact with technology. From powering sophisticated chatbots that enhance customer service experiences to aiding in creative writing and content generation, LLMs have demonstrated remarkable versatility. In the business world, these models help with automating tasks, generating insightful data analyses, and even creating marketing materials. In education, they assist in tutoring and language learning, showcasing their expansive utility. This widespread adoption highlights not only the technological advancements in natural language processing but also the growing reliance on AI-driven solutions in various industries.

Despite their advanced capabilities, LLMs exhibit notable limitations in understanding numerical data in text and capturing the complex relationships between numbers. These models often struggle in tasks requiring precise numerical understanding and sometimes understanding simple mathematical operations in the given text, which creates an interesting area of research to overcome these shortcomings. Prompting techniques like "chain of thought" (Wei et al., 2022) have addressed some of these shortcomings with the help of creating prompts for LLMs to think like the specific prompts provided. These techniques have highlighted and presented a significant step in improving the efficacy of LLMs in handling tasks that require more than just understanding text. However, prompting techniques may still be restricted in terms of creating task-specific prompts and are not generalizable across domains.

In our work, we explore the LLMs capabilities in generating headlines of news articles (Huang et al., 2023). This task is much more challenging than simply extracting sentences for summarization as it requires generating a headline that captures the semantic meaning of the article i.e. capturing the core ideas. This becomes a harder problem when these headlines contain numerical information and generate numeral-aware headlines. Recognizing these limitations and opportunities for improvement, the upcoming SemEval workshop has introduced a new task named NumEval focused on headline generation from news articles. In our work, we focus on the third task "Numeral-Aware Headline Generation (English)". This task aims to provide a robust

[*] Equal Contribution. Names listed in Alphabetic order

testing ground for developing and refining models to better handle tasks like headline generation, pushing the frontiers of what is achievable with current NLP technology.

2 Related Work

The integration of numerical values in text and automated headline generation are pivotal tasks in natural language processing (NLP). Recent advancements in large language models (LLMs) have significantly contributed to these areas. Notably, Google's PaLM (Chowdhery et al., 2023) and Minerva (Lewkowycz et al., 2022) have demonstrated exceptional capabilities in numerical reasoning and text processing. This review categorizes methodologies into fine-tuning, prompt-tuning, and zero-shot approaches for headline generation, alongside zero-shot techniques, masked fine-tuning, and additional methods for numerical generation. The inclusion of these advanced models provides a broader perspective on how the latest developments in NLP are shaping approaches in these two tasks.

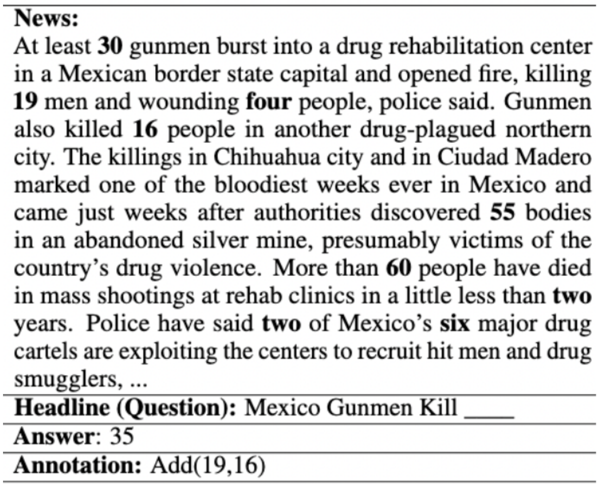
In the realm of headline generation, various methodologies have been explored to enhance the effectiveness and accuracy of these models. Fine-tuning techniques like prefix-tuning (Lisa Li and Liang, 2021) and LoRA (Hu et al., 2021) have been instrumental in adapting language models to specific tasks. Text generation models such as T5 (Raffel et al., 2020), Llama (Touvron et al., 2023), and GPT-2 (Radford et al., 2019) are particularly suited for this, each offering unique attributes for headline generation. For instance, the T5 model, as demonstrated by Raffel et al. (2020), has been effectively used for summarization tasks, a core component of headline generation. Similarly, Llama's parameter-efficient approach allows for nuanced adaptation to the stylistic demands of news headlines.

Prompt-tuning methods, including least-to-most prompting (Zhou et al., 2023) and chain-of-thought prompting (Wei et al., 2022), provide structured guidance to models in headline generation. Chain-of-thought prompting breaks down the generation process into comprehensible steps, improving the headlines' relevance and clarity. The least-to-most approach builds up on the chain-of-thought work by breaking the prompts into sub-prompts and incrementally builds the complexity of tasks, enhancing the model's focus on relevant details.

Moreover, zero-shot learning, particularly in models like GPT-4 (Achiam et al., 2023), offers the potential for generating headlines without task-specific training. The inherent abilities of these models in understanding and generating contextually relevant text, as noted by OpenAI (2023), make them suitable for creating informative headlines without extensive fine-tuning.

Parallelly, in the field of numerical generation, the roles of models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are significant. Their ability to understand context and predict numerical values makes them apt for extracting numerical information directly from text, as shown in studies by Devlin et al. (2018) and Liu et al. (2019). Moreover, masked fine-tuning, particularly with models like BERT, involves training the model to predict masked tokens, including numerical values, in a given text. This method is advantageous for generating accurate numerical information in contexts like financial reporting or statistical data summarization, leveraging the model's deep understanding of context.

With the advent of latest models, like GPT-4, Gemini Pro, PaLM, LLMs have shown a significant improvement in numerical reasoning performance. However, as they are not yet open-sourced, we have not used them in our experiments.



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: **Illustration of Numerical Reasoning Task**:
(a) The news article reports the killing of 19 men by gunmen at a drug rehabilitation center.
(b) Additionally, it mentions another incident where 16 people were killed.
(c) The annotation suggests adding these two numbers together (19 and 16) to get the total.
(d) As the answer provided in the screenshot is 35, the numerical reasoning would be the sum of the two separate incidents.

3 Tasks

The tasks defined as part of NumEval@SemEval-2024 are Numerical reasoning and headline generation. They are described in the following subsections:

3.1 Numerical Reasoning Task

The Numerical Reasoning task within this dataset presents instances where models are required to demonstrate accurate numerical comprehension as shown in Figure 1. The structured components of this part of the dataset include:

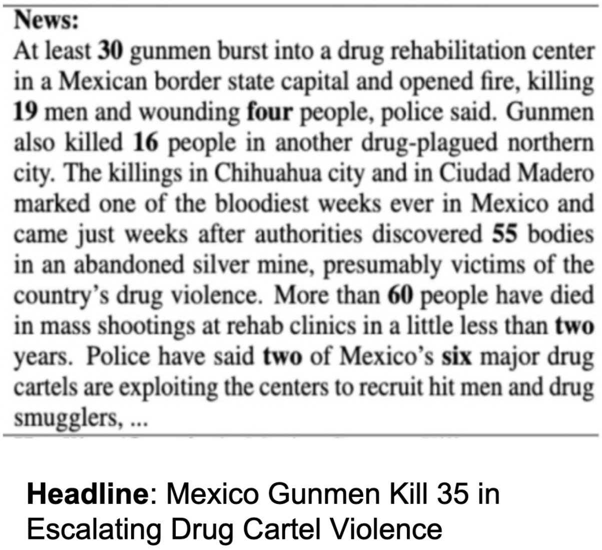
- **"news"**: The contextual news article providing information for numerical reasoning.
- **"masked headline"**: A headline with a numerical value replaced by a mask, challenging models to fill in the correct number.
- **"calculation"**: A representation of the numerical calculation involved, providing insight into the reasoning required.
- **"ans"**: The correct numerical answer.

3.2 Headline Generation Task

The Headline Generation task involves crafting headlines based on the information provided in news articles. The structured components include:

- **"news"**: The contextual news article serving as the basis for headline generation.
- **"headline"**: The target headline that models need to generate based on the provided news content.

Leveraging the critical insights from the provided dataset, the intricacies of masked headline and final headline generation tasks have been explored in this paper. These details encompass temporal elements, such as the date and time of the news article, coupled with a thorough understanding of the primary event articulated in the news. Vital numerical specifics, including casualty counts, are imperative inputs for the model. Additionally, contextual information contributes crucial background and event-specific details. Subsequently, the final headline generation follows a predefined template or structure, ensuring coherence and relevance within the news context. This nuanced process, meticulously depicted in Figures 1 and 2, underscores the imperative for the model's adept comprehension of the news article, precise extraction of relevant numerical details, and proficient integration into the headline template. The significance of these tasks resonates with the broader objectives outlined in the research.



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: Mexico Gunmen Kill 35 in Escalating Drug Cartel Violence

Figure 2: Illustration of Headline Prediction: **Summarize the main event**: (a) Gunmen attack at a drug rehabilitation center. (b) Include the key figure: 35 total killed in recent violence. (c) Reflect the broader context: Indicate a trend or pattern in the violence related to drug cartels. **Resulting Headline Prediction**: "Mexico Gunmen Kill 35 in Escalating Drug Cartel Violence"

4 Approaches

In this work, we evaluate multiple approaches for the task of numeral-aware headline generation. We use zero-shot prompting, few-shot prompting, and model fine-tuning. Through these methodologies, we aim to explore the capabilities and boundaries of LLMs in a task that requires not only understanding the gist of a news article but also conciseness essential for headline creation.

4.1 Zero-Shot Prompting

In this approach, we evaluate the capabilities of new advanced open-source LLMs like Llama2 (Touvron et al., 2023), to generate headlines without prior exposure to specific examples of the task. We use the Llama 2 -7B parameter model for zero-shot prompting. We present the model with an input prompt that clearly outlines the task of headline generation for the given news article. The input to the model is:

```
"""
<s> [INST] <<SYS>>
```

Table 1: Sample example of headline generation using T5-3b model and Llama 78 on zero-shot

<table>
<tr><td colspan="2">(Nov 4, 2008 3:19 PM) Stocks rallied on Election Day as investors applauded the looming conclusion to the presidential race, the Wall Street Journal reports. Continuing declines in interbank lending, and strong third-quarter earnings reports also fueled optimism, with the Dow rising 305.45 to close at 9,625.28. The Nasdaq climbed 53.79 closing at 1,780.12, while the S&P 500 gained 39.45 to settle at 1,005.75. The three-month US Libor dropped 0.10% to 2.706%, its 17th consecutive daily decline, signaling easing lending costs. Archer Daniels Midland rose 14.83% after the agricultural firm reported a doubling in third-quarter profit. MasterCard also shot up 18.2% despite reporting a quarterly loss of $194 million, because the loss was due to an antitrust penalty.</td></tr>
<tr><td><b>Llama2</b></td><td>Dow Soars 305 Points on Election Day, Libor Rates Drop</td></tr>
<tr><td><b>T5</b></td><td>Dow Up 305 on Election Day</td></tr>
<tr><td><b>Ground Truth</b></td><td>Stocks Up 305 in Election Rally</td></tr>
</table>

```
Create a headline for the provided news
    article, which is enclosed within
    triple backquotes. Ensure that the
    headline includes numerical
    information derived from the article
    . The numerical value should be
    processed through one of these
    operations: Copy, Add, Trans, Span,
    Round, or Paraphrase. When
    paraphrasing, represent numbers in
    the thousands using 'K' (e.g., 30000
     as 30K). The headline must be
    concise, not exceeding 10 words.
<</SYS>>
ARTICLE: ```{news}```
[/INST]
HEADLINE:
"""
```

Where $news$ represents the news article. We limited the number of output tokens as Llama was generating longer headlines and the ground truth data had headlines smaller than 10 words. By setting a word limit, we guide the model towards producing headlines contextually accurate and capture only the main context of the news article. Another addition to the prompt was how the numbers were being represented in the ground truth dataset as the headlines shortened the numbers in thousands to be replaced by ’K’, for example, 750,000 was represented as 750K in the headline.

4.2 Few-shot Prompting

We further evaluate the efficacy of Llama2, using few-shot prompting, more specifically two-shot prompting. We provide the model with two random examples from the training set, accompanied by their ground truth headline. The input to the model is:

```
"""
<s> [INST] <<SYS>>
Create a headline for the provided news
    article, which is enclosed within
    triple backquotes. Ensure that the
    headline includes numerical
    information derived from the article
    . The numerical value should be
    processed through one of these
    operations: Copy, Add, Trans, Span,
    Round, or Paraphrase. When
    paraphrasing, represent numbers in
    the thousands using 'K' (e.g., 30000
     as 30K). The headline must be
    concise, not exceeding 10 words.

Develop the headline by following the
    structure and approach demonstrated
    in two examples provided below:
<</SYS>>
Article 1: {news1} \n
Headline 1: {headline1} \n
Article 2: {news2} \n
Headline 2: {headline2} \n

ARTICLE: ```{text}```
[/INST]
HEADLINE:
"""
```

Where $news1, news2$ are two news articles and $headline1, headline2$ are the respective headlines for the two articles. These examples serve as a learning aid, offering the model concrete instances of successful headlines.

4.3 Fine-tuning

The final approach involves fine-tuning two variants of the T5 model (Text-to-Text Transfer Transformer) (Raffel et al., 2020) - ’t5-large’ and ’t5-3b’ - on our headline generation dataset. The T5 model has shown its effectiveness in various sequence-to-sequence tasks. We used a pre-trained model using the transformers library of Hugging face (Wolf et al., 2019) and fine-tuned the model to adapt to our specific task of headline generation. For the T5 model, the input to the model is:

```
summarize: {news}
```

Table 2: Performance comparison for **Headline Generation**: Zero Shot Prompting, Few Shot Prompting, and Fine Tuning

Model	**Rouge Score**			**BERT Score**		
	Rouge 1	**Rouge 2**	**Rouge L**	**P**	**R**	**F1**
BRIO (baseline)	48.93	24.09	44.12	52.17	50.84	51.43
Llama2-7B (zero-shot)	36.14	13.33	30.63	32.70	45.20	38.89
Llama2-7B (few-shot)	37.56	14.23	32.78	33.42	45.47	41.49
T5-large	46.42	21.97	41.64	49.25	47.06	48.17
T5-3B(2 Epochs)	47.89	22.94	42.9	50.9	48.81	49.8

Where $news$ is the news article. The prefix *summarize* is added to the input for the model to understand that it is a summarization task. We also limited the output number of tokens to 32 to ensure the model generates smaller headlines to match the ground truth.

5 Results

The investigation into numerical-aware headline generation encompasses an examination of three key approaches: zero-shot prompting, few-shot prompting, and fine-tuning(cite). These approaches are evaluated using prominent language models, including LlaMa2 7B, T5-Large, and T5-3B. The findings presented herein shed light on the nuanced performance of each approach and the subsequent implications for advancing the field.

5.1 Data Pre-processing

The initial step involves pre-processing the dataset to remove irrelevant information, particularly focusing on the date and timestamp. This step is crucial as it ensures that the model's attention is directed towards the content relevant to headline generation, rather than extraneous details. The preprocessing involves scrubbing the data of any temporal metadata that might skew the model's learning towards time-sensitive contexts rather than the core news content.

5.2 Qualitative results

Table 1 shows qualitative results for headlines generated for one of the dev set examples. We can see that the model generates similar headlines and correctly picks up the number from the given news article. T5 produces a smaller headline as we had limited the number of output tokens for the model and Llama2 understands the prompt correctly. Table 6 shows some not-so-good samples.

Table 3: Performance comparison for **Numerical Reasoning**: Zero Shot Prompting, Few Shot Prompting, and Fine Tuning

Model	**Accuracy Score**
BRIO (baseline)	66.56
Llama2-7B (zero-shot)	40.13
Llama2-7B (few-shot)	41.08
T5-large	62.18
XLM-Roberta (Zero-shot)	58.7
T5-3B	63.65

5.3 Comparative Analysis of LLMs

Our exploration of headline generation using large language models (LLMs) began with a zero-shot prompting approach using LlaMa2 7B, which, despite its robust pre-training, fell short of improving headline generation metrics significantly, as evidenced in Table 2. Progressing to few-shot learning with the same model offered only marginal gains and failed to outperform the established BRIO baseline. Subsequent fine-tuning of LlaMa2 7B, T5-Large, and T5-3B models led to notable performance enhancements; particularly, the T5-3B model stood out with superior Rouge scores, overtaking the BRIO model (Liu et al., 2022) in efficacy after fine-tuning, detailed in Table 2. Even for numerical reasoning tasks, the fine-tuned T5 3B model emerged as the best-performing variant. Although it did not exceed the baseline, the results, as outlined in Table 3, exhibit significant promise and suggest a viable avenue for future exploration. It should be noted that the training of the T5-3B model was limited to two epochs due to computational resource constraints, which presents an area for potential improvement with extended training.

6 Future Work

Fine-tuning pre-trained text generation models shows great promise, with the performance almost

Table 4: Zero-shot results of xlm-roberta for numerical generation

<table>
<tr><td colspan="2"><b>News Article</b>: "(Mar 6, 2016 10:50 AM) Nancy Reagan, the helpmate, backstage adviser, and fierce protector of Ronald Reagan in his journey from actor to president—and finally during his 10-year battle with Alzheimer's disease—died Sunday at the age of 94, reports the AP, via CBS News. The cause was congestive heart failure, notes ABC News. In addition to her famous campaign against drugs, the one-time actress promoted several causes while she was in the White House and even in the years after. She was a passionate advocate for lifting restrictions on stem cell research and promoting better treatment of America's veterans. But above all, Nancy Reagan was a fiercely devoted wife. My life began with Ronnie, she told Vanity Fair magazine in 1998. The first lady's public life had its share of controversy but also earned the respect of the nation, making Nancy Reagan one of America's most admired women in the 1980s and beyond. Anne Frances Nancy Robbins was born on July 6, 1921 in New York City to Kenneth Robbins, a car salesman, and Edith Luckett Robbins, an actress. She met Ronald Reagan in 1950, when he was president of the Screen Actors Guild and she was seeking help with a problem: Her name had been wrongly included on a published list of suspected communist sympathizers. They discussed it over dinner, and she later wrote that she realized on that first blind date he was everything that I wanted. They wed two years later, on March 4, 1952. She was thrust into the political life when her husband ran for California governor in 1966 and won. She found it a surprisingly rough business. The movies were custard compared to politics, she said. The couple had two children together, Patricia Ann and Ronald Prescott. Reagan will be buried next to her late husband at the Ronald Reagan Presidential Library in Simi Valley, California. The New York Times has a full obituary here."</td></tr>
<tr><td><b>Masked Headline</b></td><td>'Nancy Reagan Dead at _____'</td></tr>
<tr><td><b>Ground-truth</b></td><td>94</td></tr>
<tr><td><b>Prediction</b></td><td><b>score</b>: 0.875 | <b>token</b>: 16064 | <b>token_str</b>: '94'</td></tr>
</table>

reaching the baseline scores. However, there's still room for a lot of improvement and it is clear from the experiments that with longer fine-tuning even better results can be achieved. As part of future work, various fine-tuning techniques like prefix-tuning, LoRA can be employed to fine-tune the larger models, which otherwise is a difficult task owing to their huge parameter sizes.

For numerical generation task, where we are given a news article and a masked headline, and we need to fill the mask token, similar experiments as text generation can be conducted. While prompting can be the most effective way of filling the masked token, it was interesting to see the zero-shot results of RoBERTa. This prompted us to consider a masked fine-tuning approach for numerical value generation using RoBERTa, which can potentially lead to further improved results.

The proposed approach is shown in Figure 3. The news article and masked headline are concatenated with the blank replaced by the MASK token. This forms the input for the pre-trained model. Additionally, minor architectural changes are proposed. Two prediction heads are attached to the model - one, a binary classification head for classifying the reasoning process, i.e. whether answer was generated by a copy (direct extraction from text) operation or a not copy (reasoning) operation; two, a MLM head for predicting the MASK token. The MLM head produces a probability vector of size of the vocabulary, the prediction for which simply becomes the token with the highest probability. Both the prediction heads are simple linear layers with softmax activation. The whole model can be fine-tuned in a BERT-like masked modeling fashion with Negative log likelihood or cross entropy loss functions.

Furthermore, the pre-trained models can be initially fine-tuned on a mathematical dataset like GSM-8K (Cobbe et al., 2021), which consists of math word problems. This would enable the models to develop better reasoning capabilities by learning to solve arithmetic problems for generating the answers. Based on the experimental results from numerical generation task, it was observed that, while it is easy for models like RoBERTa to fill the mask token when the answer was already present in the article and just had to be extracted, it performs

Table 5: Zero-shot xlm-roberta for numerical generation. Cases where it fails

News Article: "(Feb 8, 2013 9:05 AM) With the manhunt in Big Bear Lake for ex-cop Christopher Dorner coming up empty so far, authorities are admitting they have no idea where Dorner is, reports Fox News. Thousands of police officers are now involved in the hunt, in California, Nevada, Arizona, and northern Mexico. He could be anywhere at this point, says the San Bernardino county sheriff. Meanwhile, the AP reports that as many as 16 San Diego County sheriff's deputies spent the night surrounding and searching a rural home after a hoaxer reported Dorner was there. There were people at home but Dorner wasn't one of them. Investigators have a pretty good idea who made the call and will seek criminal charges."	
Masked Headline	'Thousands of Cops Look for Dorner in _____ States'
Ground-truth	3
Prediction	**score**: 0.75 | **token**: 70 | **token_str**: 'the'

Table 6: Example of Failure Cases

(Oct 3, 2011 9:10 AM CDT) Tobacco companies were well aware that their products contained radiation2̌014and they discovered this decades ago, UCLA researchers studying 27 historical documents have found. The firms learned of the presence of polonium-210 in cigarettes in 1959, and they examined the radioactive material's effects during the 1960s, documents show. They found that it caused cancerous growths in users2̌019 lungs, and figured out just how much radiation a typical smoker would inhale, but the companies didn2̌019t warn the public. Using the original calculations, the researchers determined that the radiation would lead to the deaths of 138 of every 1,000 smokers over 25 years, ABC News reports. Not only are you inhaling it, but you're emitting radiation when you smoke, and your family, your dog, your cat are all inhaling that radiation, says an expert. What2̌019s more, some of these radiation particles hang around for decades. The timing of the original discovery is notable, says an anti-smoking advocate: It happened in an era when Americans were crawling under our desks during school radiation drills, meaning the news would have likely had a huge impact. Polonium remains in cigarettes, and a rep for Philip Morris says it2̌019s no secret to public health experts. It2̌019s a naturally occurring element found in the air, soil, and water, says the spokesman.	
Llama2	Radiation in Cigs: Tobacco Companies Knew, (138 Deaths per 1,000 Smokers)
T5	Tobacco Firms Knew of Cancer-Causing Radiation in Cigarettes
Ground Truth	Tobacco Firms Knew of Radiation in Cigs—in 1959

poorly on samples where the answer needs to be computed using arithmetic operations or that needs reasoning. In table 4., we can see that, the answer involved copying the number from the text, which the model could capture quite easily with high confidence. But, in table 5, where it had to compute the result by performing Add(Add(Span(California), Span(Nevada)), Span(Arizona)), it fails. Hence, proper understanding of solving numerical problems would significantly help these LLMs. .

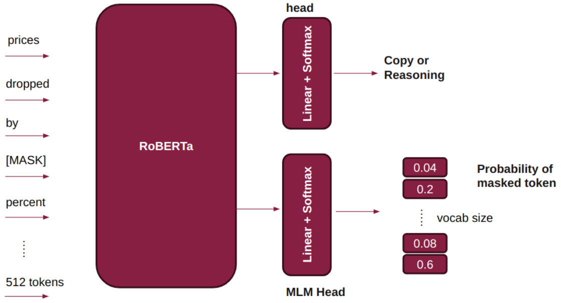


Figure 3: MLM fine-tuning RoBERTa

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, F.L. Aleman, Diogo Almeida, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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. *International Conference on Learning Representations*.

JT Huang, CC Chen, HH Huang, and HH Chen. 2023. Numhg: A dataset for number-focused headline generation. *arXiv preprint arXiv:2309.01455*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems 35*, pages 3843–3857.

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

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*.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog 1, no. 8*.

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*, (1):5485–5551.

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

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, et al. 2023. Least-to-most prompting enables complex reasoning in large language models. In *Proceedings of the Eleventh International Conference on Learning Representations*.