Temporally-Varied News Headline Generation with GPT-3

Victoria Graf **Edward Tian** **Richard Zhu**

Princeton University

{vgraf,ejtian,ryzhu}@princeton.edu

Abstract

We evaluate the ability of GPT-3 to generate news headlines in zero-shot and few-shot settings across a wide range of years. We introduce our novel temporal analysis of headline generation through a dataset of Daily Princetonian articles across five decades which allows for analysis of temporal misalignment. We observe that demonstrations have little effect on task performance, including in settings with temporally mismatched examples. Additionally, we show that short article lengths are sufficient for performing headline generation and that topical domain does not affect performance on the BBC news article dataset. Lastly, we evaluate the alignment of generated headlines on the basis of human preference, and we find that generations were well-aligned relative to our 4 ranked criteria for good headlines (informative, harmless, honest, and interesting).

1 Introduction

Large language models such as GPT-3 are inherently pre-trained on text corpora from a specific temporal range. Namely, they see primarily data from recent years due to the exponential growth of internet data, and they are limited in the data they see by the year of their training. This inherent skew in the temporal distribution of data may degrade model performance on tasks involving understanding temporally misaligned (older or younger) text.

One genre with a significant portion of text that predates the internet is news articles. The Daily Princetonian, for example, dates back to print articles from the 1870s. Not only has the terminology of the English lexicon evolved across decades, certain linguistic and stylistic practices have also evolved in the news industry across time (Westin, 2016), which could also be cause for temporal lexical shifts. Thus, news headline generation is a natural task with an inherent temporal component that could be used in testing the effects of temporal misalignment in GPT-3 generations. To our knowledge, this study is the first to evaluate temporal misalignment for news headline generation.

Additionally, news headline generation is an application of latent text summarization abilities of GPT-3 with potentially significant economic impact. Millions of news articles are published every day, and it is the task of human editors to write associated headlines. As a result, headline generation tools are of significant economic interest to the journalism industry.

The application of news headline generation technology continues to be an open problem. The need for automated headline generation is significantly larger in newsrooms that publish thousands of articles per day, such as Bloomberg, compared to newsrooms that publish closer to hundreds of articles daily, such as the New York Times. Bloomberg has been a leader in the space of automated news generation, releasing Bloomberg Automated Intelligence (BAI) in 2020 which writes shorts articles and headlines for specific use-cases on business and finance articles (Fesanghary and Verma, 2021).

In this paper, we attempt to further improve understanding effects of temporal misalignment, by specifically studying temporal and domain mismatch between demonstrations and inputs in an in-context news headline generation setting. The contributions of our paper are the following:

- Evaluating the effect of temporal misalignment on news headline generation
- Determining the necessary truncated article length for news headline generation
- Evaluating the alignment of GPT-3 generated English news headlines with human preferences

2 Related Work

Temporal Misalignment Luu et al., 2022 demonstrated the effects of temporal alignment in a diverse range of domains across periods of time spanning five years or more. Following intuition, they concluded that when a language model is trained on text from one time period and deployed on data from another, the resulting temporal misalignment can significantly degrade task performance. They note that continued pretraining on temporally aligned data can improve this gap but that finetuning on temporally correct task data produces greater gains.

Other studies have noted that this degradation of task performance may be due to lexical shifts in text over time. Namely, semantic shifts occur naturally in language causing inherent issues with temporal misalignment in older models. Su et al., 2022 introduce a pipeline for detecting semantic change in words through a Lexical-based Masked Language Model (LMLM) objective and demonstrate that this kind of post-training approach produces improved results on temporally misaligned settings. Additionally, semantic changes can occur quickly during major events in addition to gradual shifts over time. These quick shifts during crises are studied by Pramanick et al., 2021, and they pose additional challenges in adaption due to the reduced data involved in shorter time frames. Pramanick et al., 2021 propose using methods from domain adaptation to enhance performance on these challenging cases of temporal misalignment.

Large language models such as GPT-3 are pretrained on text corpora from the internet resulting in a bias toward more recent text due to the exponential growth of internet data. An analysis of the C4 training data in Dodge et al., 2021 revealed a heavy bias towards more recent data in the corpus, specifically toward data generated closer to 2019, the year when C4 was first introduced by Raffel et al., 2020. GPT-3 was pretrained on 2016-2019 Common Crawl data (though not C4 specifically) indicating that GPT-3 would likely have a similar temporal data skew (Brown et al., 2020). This inherent skew in the temporal distribution of data may degrade model performance on tasks involving understanding older text.

Headline Generation Headline generation has a strong resemblance to text summarization tasks. Literature focusing on news headline generation categorize it as a special kind of text summarization task (Bukhtiyarov and Gusev, 2020), where models need to have strong natural language understanding that goes beyond the meaning of individual words and sentences.

In headline writing, long-form text is distilled into a short phrases while retaining key components of the original text, which is the essence of text summarization. Thus, news headline generation is analogous but not identical to heavily reduced summarization tasks on news corpora such as in Chen et al., 2016 which exploits existing summarization in CNN and Daily Mail news articles from associated bullet point summaries at the beginning of articles.

In addition to generating headlines for news articles, researchers have studied the generation of short, representative headlines for news stories (sets of multiple related articles) (Gu et al., 2020). Gu et al., 2020 create NHNet, a multi-document news-headline generation model which outperforms other summarization models such as WikiSum, SinABS, Concat, and SinABS on headline generation, and they release the first manually curated dataset for news-story headline generation, NewHead, which contains more than 367K stories. Our headline generation approach focuses on generating headlines for individual articles.

More similarly to our approach, in Bukhtiyarov and Gusev, 2020, researchers fine-tune pre-trained Transformer-based models for headline generation and achieve new state-of-the-art results on Russian news datasets. A more recent study by Anastasiu et al., 2021 leveraged a fine-tuned BERT model to generate headlines for German news articles. Notably, both of these papers evaluated their generated headlines with ROUGE scores similarly to established summarization practices. However, Anastasiu et al., 2021 recognized the limitations of ROUGE scores for measuring the quality of generated headlines and conducted a small amount of human evaluations for 100 article and headline pairs. In our study, we also conducted human evaluations on a selected subset of the articles to measure alignment to human preferences.

There are some potential industry applications of applying GPT models for news headline generation such as Mutiny, a marketing company, experimenting with GPT-3 headline suggestions (Hoey). However, per our understanding, this study is the first robust academic evaluation of the performance of

in-context English headline generation with GPT-3.

3 Data

3.1 Datasets

We primarily use the Daily Princetonian corpus for our temporal investigations of headline generation. The data is from Princeton's Mudd Library and was constructed using optical character recognition (OCR) data from the Larry DuPraz archive of digital articles. The data contains articles from the Daily Princetonian between 1900 to 2015, categorized by year. The Daily Princetonian is a campus newspaper written by students.

We also use a dataset of professional news articles for headline generation, specifically a set of BBC news articles (Greene and Cunningham, 2006). This dataset contains 2,225 articles from 2004 to 2005 and is divided by article type into five domains: business, sports, entertainment, politics, and tech. Evaluating on the BBC dataset allows comparison of headline generation for student written articles with professionally written articles. Furthermore, the breadth of this dataset allows us to investigate variations in topical accuracy of headline generation across difference news domains (e.g. sports and business).

3.2 Processing

We perform several processing steps on the Daily Princetonian dataset. For each file, we filter blank spaces and irregular characters. Then, for all articles, we remove and store the headline and filter the byline or author name by removing up to the first name to appear in the first 50 characters after the headline. Furthermore, we remove any repeated headlines from the dataset since these generally reflected weekly or monthly specials rather than standard articles (e.g. "editor's note" which repeats multiple times; theoretically no two standard news articles should have the same headline).

The dataset for each year was then divided into testing and training sets to separate articles used for demonstrations and those used for testing. For each year, the training set contained 50 articles (where each year contained around 1000-1500 articles). In the year 1985 with 1489 articles, the training set is less than 5% of the data.

4 Methodology

This paper runs experiments on GPT-3 since it is a state of the art LLM. We use the text-davinci-002 model for all our experiments. GPT-3 was prompted to generate headlines for articles truncated to the first l tokens. For the majority of experiments, we use $l = 500$ with the exception of the experiment in which we vary l. For few-shot settings, we use $k = 3$ demonstrations for the majority of experiments.

4.1 Prompting and Demonstrations

k-shot prompting We perform evaluations of GPT-3 generated headlines on testing-dataset articles across years by prompting the model with "Write a headline for the following article\n\n{article text}."

We randomly select sets of k examples (initially, $k = 0$ for zero-shot and $k = 3$ for few-shot) from the training set for demonstrations. We used the label "headline: " before headlines in the demonstrations and append the prompt format noted previously.

For the Daily Princetonian corpus, we also study how temporal misalignment of examples affects performance. Specifically, we have two settings we compare: (1) examples from the same decade as the prompted article and (2) examples from a different decade from the prompted article.

4.2 Evaluation

To evaluate the quality of our generated headlines, we focused on ROUGE scores due to the similarity of our task to summarization. We used ROUGE-1, ROUGE-2, and ROUGE-L scores to compute similarity between the generated and reference headline.

Human Evaluation To align our evaluation of generated headlines to human preferences, we did a round of human evaluation of generated headlines. Specifically, we define four criteria for the quality of headlines:

- **Informative** headlines accurately summarize the content of the article.
- **Harmless** headlines are not offensive, overly biased, or toxic. However, we did not penalize non-toxic bias in articles that appear to be editorials.
- **Honest** headlines do not contradict information in the article.
- **Interesting** headlines make humans want to read the associated article.

The order of priority for these criteria was chosen to reflect the potential application of this task to real news headline generation.

Evaluators were given the generated and original headlines in random order with the first 500 tokens of the associated article and were instructed to select the headline that they preferred using the above criteria. They were instructed to only rank the headlines as equal if the headlines were particularly hard to distinguish. Full instructions for human evaluation are included in the Appendix.

5 Experiments

Zeroshot Performance We conduct experiments on zeroshot performance of headline generation for article across years with results in Figure 1. ROUGE-1, ROUGE-2, and ROUGE-L scores had similar patterns in headline generation performance across years with ROUGE-1 and ROUGE-L scores consistently above ROUGE-2 scores in all years. Zeroshot performance showed a U-shape performing best in 1940 and 1975, with ROUGE-1 scores around 0.55, decreasing significantly in 1985, and then increasing again in years 2000 and 2015. This trend is surprising given that GPT-3 was trained on Common Crawl data from 2016-2019.

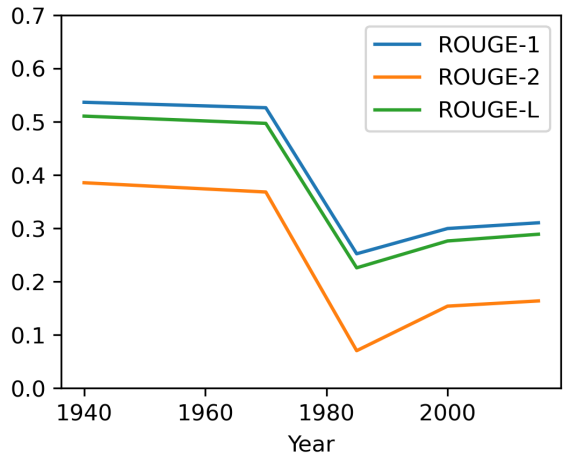

Figure 1: Zeroshot performance varied over article years in a non-linear shape. This is surprising given that GPT-3's training data is from 2016-2019.

Fewshot Performance For the $k = 3$ baseline with temporally matched demonstrations (examples from the same year as the prompted article), a similar trend to zeroshot performance in ROUGE scores is seen (Figure 2) with scores reaching their minimum for examples from 1985. However, scores plateau at the extremes for $k = 0$ while scores increase sharply towards the extremes for $k = 3$ indicating that the low performance for 1985 was not just an outlier year.

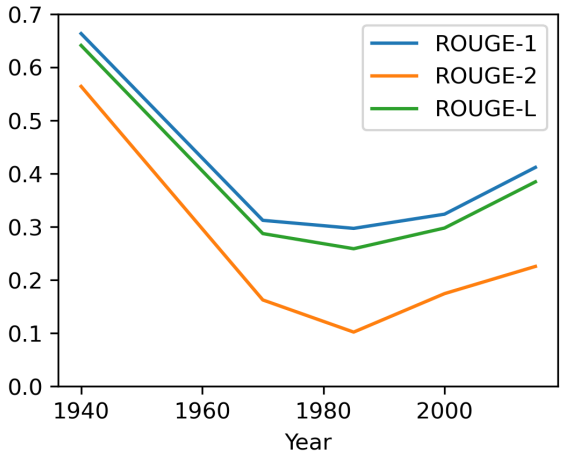

Figure 2: Fewshot ($k = 3$) performance showed a similar trend to zeroshot performance over the tested years.

Temporally Mismatched Demonstrations We investigate the effect on performance of demonstrations that are temporally mismatched from the prompt article. We maintain temporal consistency within the demonstrations - each set of demonstrations for a given prompt comes from reference articles from the same year.

When fixing the demonstrations to be from the year 2000 and varying the year of the prompted headline (Figure 3), we see results that appear to follow the few-shot ($k = 3$) baseline. Thus, it seems to suggest that temporal misalignment of demonstrations has a negligible effect on generated headlines. This may be becuase GPT-3 has seen sufficient data on news articles that the model is largely unaffected by demonstrations.

Meanwhile, when fixing the prompted article year to 2015 and varying the demonstration year (Figure 4), we see a almost no change in performance across demonstration year. This is consistent with the minimal effect we saw comparing Figures 2 and 3. Performance reaches a local maximum for the 2000 data and drops off to either side. Surprisingly, ROUGE scores are near their lowest for demonstrations in the year 2015, suggesting that slightly temporally mismatched prompts may not harm performance or could even cause better performance than temporally matched ones.

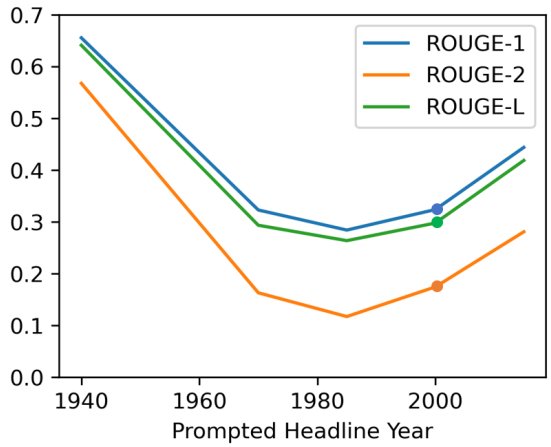

Figure 3: Performance was similar to temporally aligned fewshot when the year of the demonstrations was fixed to 2000 and year of the prompted headline was varied. Marked datapoints represent the year of prompted generations at which demonstrations and prompted generation were aligned.

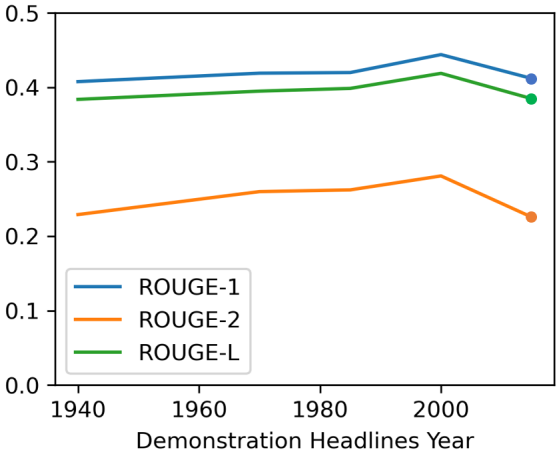

Figure 4: Performance was stable when the year of the prompted headline was fixed to 2015 and the year of demonstrations was varied. Marked datapoints represent the year of demonstrations at which demonstrations and prompted generation were aligned.

Topical Domain Mismatch Since the Daily Princetonian dataset is comprised of student-written articles, we additionally perform experiments on a dataset of professionally written BBC new articles. We evaluate zeroshot headline generation performance on a dataset of BBC news articles in 5. The BBC dataset had ROUGE-1 scores around 0.35, which is close to the Daily Princetonian dataset in 2000. (In comparison, the zeroshot performance for the Daily Princetonian dataset had ROUGE-1 scores ranging from 0.3 to 0.6 depending on the article year.) This is consistent with the timeline of the BBC dataset, consisting of articles from 2004 and confirms that news articles in the early 2000s often have these lower ROUGE scores for GPT-3 headline generation not just for the Daily Princetonian dataset, although more data needs to be evaluated to confirm this possibility.

Additionally, the BBC news dataset contains information on article domain (namely, each article was given a category out of entertainment, business, tech, sports, and politics). ROUGE scores were stable across domains in the BBC news dataset (Figure 5).

Subsequently, the impact of demonstration news domains on headline generation was further tested through cross-domain fewshot generations, where the demonstrations for ($k = 3$) fewshot examples were selected from a different news domain from the prompted generation. Domain mismatch in demonstrations and prompted headlines had negligible impact on our results as observed (Figure 6).

Input Length We conduct additional experiments to determine optimal article lengths. Articles were truncated at $l = 500$ tokens in the previous experiments. We vary input lengths through tokenizing the text and truncating at l tokens.

As seen in Figure 7, truncating the length of the articles did not significantly affect performance. This result is not expected, as it is common practice in news articles to include the most important information in the 'nutgraph' or introductory paragraph. Thus, removing later text would not significantly impact the performance of headline generation since the key information is summarized early in the article.

Additionally, we varied the number of examples k (initially $k = 3$) in our temporally matched fewshot setting and determined that number of demonstrations had little to no effect on performance (Figure 8).

Human evaluation ROUGE scores are an imperfect metric for evaluating headline generations. As such, we use human evaluations to establish a grounded baseline on whether our generated headlines are aligned to humans preferences. An ideal headline generator would produce headlines indistinguishable from journalist-written headlines and thus be preferred by humans equally often as real headlines. The results of our human evaluations on zeroshot, fewshot, and temporally misaligned few-

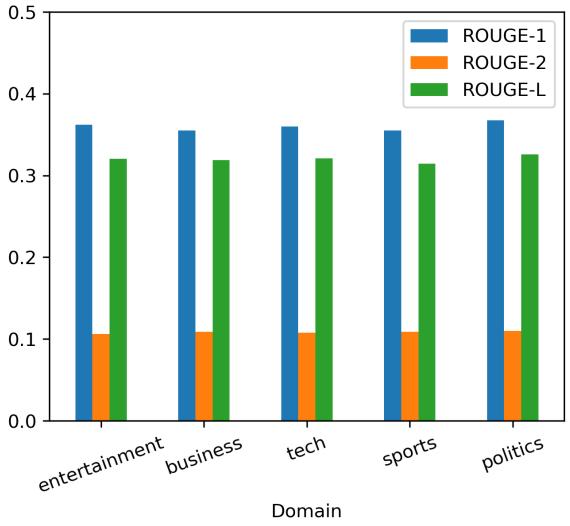

Figure 5: Performance was stable across domains in the BBC news dataset.

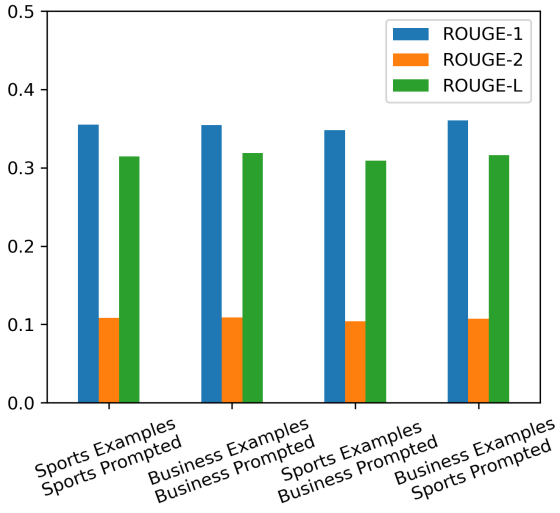

Figure 6: Domain mismatch in the BBC dataset between demonstrations and prompted headlines did not affect performance.

shot setting generations are shown in Figure 9 and Table 1 in the Appendix. Human preference scores demonstrate that GPT-3 performs well in headline generations based on human criteria being approximately equally preferred to actual headlines.

6 Conclusion

This paper makes a number of surprising findings. Most notably, we observe that temporal performance trends are non-linear. Rather, they generally formed a U-shape with the year of prompted headline with higher performance around 1940 (unexpected) and 2015 (expected). Additionally, we found that demonstrations have little effect on performance in any setting, including varying the number of demonstrations (k; Figure 8), the year of demonstrations (Figure 4), and the domain of demonstrations (Figure 5).

In a practical result, we find that short article lengths are sufficient to achieve average performance on headline generation, as increasing the length of articles used as both demonstrations and prompts for evaluation have little effect on ROUGE scores. This suggests that future work may truncate articles at as early as 100 tokens.

Finally, we find that generated headlines are well aligned with human preferences. In fact, human evaluators on the whole prefer model-generated headlines equally to human-written reference headlines (Figure 9). This strong alignment suggests headline generation is among more well aligned outputs even within summarization-type tasks.

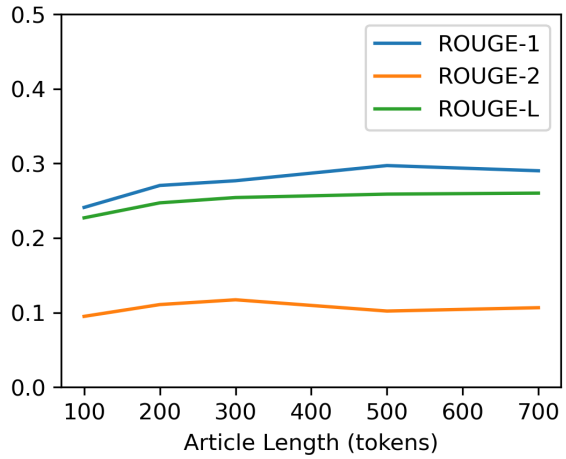

Figure 7: The truncated length of the articles in the context did not affect performance.

Limitations Our experiments were conducted for a limited quantity of data from select years (i.e. 1940, 1970, 1985, 2000, and 2015) due to the available time and resources. Meanwhile, newsrooms practices could shift year-to-year leading to unrepresentative years in the data. Thus, a single year (e.g. 1985) may not be an accurate representation of its entire decade or surrounding time-frame. As such, with additional resources, we would recommend increasing the scale of our study to generate results for all years from 1900 to 2015.

In addition, we are aware of the limitations of ROUGE scores as a evaluation metric for headline generation. Although past papers on headline generation have consistently used ROUGE scores,

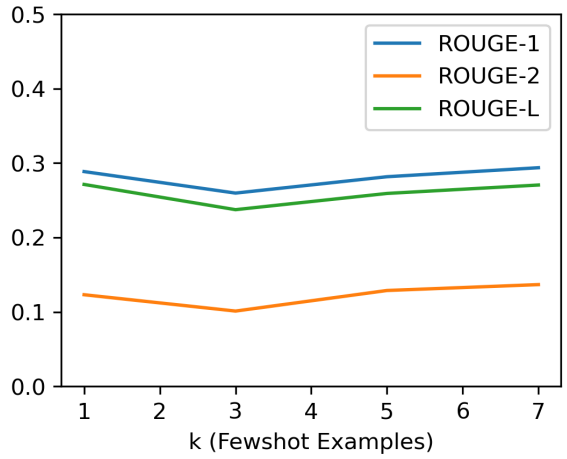

Figure 8: The number of examples k in the context did not affect performance.

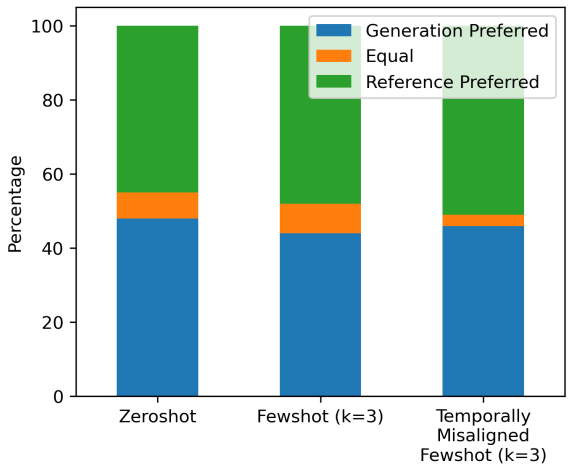

Figure 9: Human evaluation results. Generated and reference headlines performed similarly in the human evaluation.

ROUGE scores are unable to capture critical human criteria for good headlines such as accurate, non-toxic, informative, and "catchy" headlines. We attempt to bridge the shortcomings of ROUGE scores with human evaluation for a subset of our generations. In a study with greater time and resources, we would prefer to conduct more thorough human evaluations on headlines from more years and settings.

Acknowledgements

We would like to thank Danqi Chen and Alex Wettig for all of their guidance and support in writing this paper.

References

Cristian Anastasiu, Hanna Behnke, Sarah Lück, Viktor Malesevic, Aamna Najmi, and Javier Poveda-Panter. 2021. Deeptitle–leveraging bert to generate search engine optimized headlines. *arXiv preprint arXiv:2107.10935*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Alexey Bukhtiyarov and Ilya Gusev. 2020. Advances of transformer-based models for news headline generation. In *Conference on Artificial Intelligence and Natural Language*, pages 54–61. Springer.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Mohammad Fesanghary and Arun Verma. 2021. Predictive analysis of bloomberg automated intelligence.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 377–384, New York, NY, USA. Association for Computing Machinery.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, You Wu, Cong Yu, Daniel Finnie, Hongkun Yu, Jiaqi Zhai, and Nicholas Zukoski. 2020. *Generating representative headlines for news stories*, page 1773–1784. ACM.

Alec Hoey. Generating personalized website headlines using gpt-3 for mutinyhq engineering and marketing purposes. Mutiny Engineering Blog.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Aniket Pramanick, Tilman Beck, Kevin Stowe, and Iryna Gurevych. 2021. The challenges of temporal alignment on twitter during crises.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Zhaochen Su, Zecheng Tang, Xinyan Guan, Juntao Li, Lijun Wu, and Min Zhang. 2022. Improving temporal generalization of pre-trained language models with lexical semantic change.

Ingrid Westin. 2016. *Language change in English newspaper editorials*. Brill.

A Appendix

A.1 Additional Data

	Setting		
Preference	Zeroshot	Fewshot (k=3)	Temporally Misaligned Fewshot (k=3)
Generation Preferred	48	44	46
Equal	7	8	3
Reference Preferred	45	48	51
Total	100	100	100

Table 1: Human evaluation results. Each of 4 annotators were instructed to evaluate 25 headlines from each of three settings.

A.2 Human Evaluation Instructions

(on next page)

Thank you for your help with our study on news article headline generation!

You will be given spreadsheets with four columns, as shown below. For each row, please first read through the entry in the 3rd column ("Article"). This is the beginning of a news article from the Daily Princetonian.

Then read the two headlines in columns A and B ("Headline A" and "Headline B") before indicating, in column D, "A" if you prefer Headline A over Headline B, "B" if you prefer Headline B over Headline A, or "=" if the headlines are too similar for comparison. Try to avoid using "=" unless you are having a particularly difficult time distinguishing between the two headline options (e.g. they differ by one word that does not change their meaning). *How do I assess whether I prefer one headline over the other?* ***Please read the guidelines for high quality headlines below.***

Please only edit the last column of the spreadsheet ("Preferred Headline (A, B, =)").

A1 | Headline A

	A	B	C	D	E
1	Headline A	Headline B	Article	Preferred Headline (A, B, =)	
2	Wood hooks award, casting a line to female fishers	Lillith Wood '00 to use Dale Fellowship to travel America and write book on women who fish for a living	After graduation, many University students follow similar paths, pursuing careers in consulting, becoming investment bankers or going to graduate school. But in a journey financed by the Martin A. Dale Fellowship, Lillith Wood '00 will follow a different route, traveling the coastlines of America to write a book about women who fish for a living. "I'm going to start in Alaska this summer, travel down the west coast, across the gulf coast to Florida and up the east coast to Maine, taking about a year," Wood said in an interview yesterday. "I will stop in selected fishing towns, talk to women who fish commercially for a living, take their stories and create a modest book-length piece of creative non-fiction." The Dale fellowship provides $20,000 each year for a student to pursue an independent project for a year following graduation. Wood said the idea for her project was inspired by her upbringing in a small Alaskan fishing community. "I grew up in a fishing town in southeast Alaska," Wood said. "Fishing has been a big part of my life. I grew up in the middle of a culture that was very involved with fish." Wood's hometown of Petersburg is on Mitkof Island, a small island "about half way between Juneau and Ketchikan." "It's pretty small," she said.	

Guidelines for High Quality Headlines:

A high quality headline is one that is (in order of descending importance)…

- **Informative**. It must accurately summarize the content of the article sample and provide relevant information.
- **Harmless**. The headline should not be offensive, overly biased, or toxic. However, for articles that appear to be editorials, it is alright for headlines to construe the topic in a biased but still non-toxic way.
- **Honest.** The headline should not contradict information in the article.
- **Interesting**. The headline should make you want to read the article.

Please try to finish your assigned articles/headline matching as soon as you can since it gives us more time to understand and evaluate results.

Questions? Please contact Edward (ejtian@), Richard (ryzhu@), or Victoria (vgraf@).

Thank you for your help!!! 😊😊😊

Figure 10: Human evaluation instruction document, provided to all human evaluators for consistency.