Media Bias Detection Across Families of Language Models

Iffat Maab[1], Edison Marrese-Taylor[1,2], Sebastian Pado[3], Yutaka Matsuo[1]
[1]The University of Tokyo
[2]National Institute of Advanced Industrial Science and Technology
[3]University of Stuttgart
{iffatmaab, emarrese, matsuo}@weblab.t.u-tokyo.ac.jp
pado@ims.uni-stuttgart.de

Abstract

Bias in reporting can influence the public's opinion on relevant societal issues. Examples include informational bias (selective presentation of content) and lexical bias (specific framing of content through linguistic choices). The recognition of media bias is arguably an area where NLP can contribute to the "social good". Traditional NLP models have shown good performance in classifying media bias, but require careful model design and extensive tuning. In this paper, we ask how well prompting of large language models can recognize media bias. Through an extensive empirical study including a wide selection of pre-trained models, we find that prompt-based techniques can deliver comparable performance to traditional models with greatly reduced effort and that, similar to traditional models, the availability of context substantially improves results. We further show that larger models can leverage different kinds of context simultaneously, obtaining further performance improvements.

1 Introduction

Both mass media and social media are potent channels for expressing viewpoints and shaping decisions. News outlets stand out, exerting a pivotal influence on altering and molding individual and collective perspectives (Entman, 2007; Hamborg et al., 2019). Clearly, biased media has the ability to influence people, for instance, in a study by Fletcher and Park (2017), it is revealed that there is a negative association between trust in the news media and online news participation. Hamborg et al. (2019) highlight that distinctive contributions can be made by computer scientists to study *bias*. We follow them in defining bias as subjective standpoints manifested variously in, for example, word choice, framing, intentional omission or misrepresentation of specific details (Lin et al., 2006; Iyyer et al., 2014; Rashkin et al., 2017).

In NLP, media bias has been scrutinized in different ways using different names (Pan et al., 2018; Pérez-Rosas et al., 2017). An important step towards a common perspective was the development of the Bias Annotation Spans on the Informational Level (BASIL) dataset, which established the specific concepts of *information bias* and *lexical bias* (Fan et al., 2019). Informational bias refers to selective presentation of content in a factual manner to sway opinion of readers (van den Berg and Markert, 2020; Fan et al., 2019), whereas lexical bias refers to linguistic attributes like word selection and syntax (Hube and Fetahu, 2019; Greene and Resnik, 2009; Iyyer et al., 2014).

These types of bias have been studied with supervised learning methods (Lei et al., 2022; van den Berg and Markert, 2020; Lee et al., 2021; Fan et al., 2019; Guo and Zhu, 2022a; Maab et al., 2023a). However, so far the task proves challenging and often results in unsatisfactory performance when trained with a limited set of labeled examples. This can be due to the expensive nature of data annotation and extensive variety of domains, languages, and tasks (Chen et al., 2020; Akhter et al., 2020). This is particularly true given the importance of context in the identification of bias (van den Berg and Markert, 2020; Guo and Zhu, 2022a; Lei et al., 2022; Lee et al., 2021; Maab et al., 2023a,b). For example, in the BASIL dataset, where political bias is identified at the sentence-level, it has been shown that *informational bias* depends fundamentally on the context of the sentence (Guo and Zhu, 2022b; van den Berg and Markert, 2020), arises from manipulation of information or selective presentation of content in a factual way, e.g., use of quotes, to evoke reader's emotions towards news entities (Fan et al., 2019; van den Berg and Markert, 2020).

Thus, recent work on automatic bias detection in news has focused on identifying the right context to present during training. Context ranges from

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4083–4098

whole articles (van den Berg and Markert, 2020), to just sentences surrounding the target sentence (Guo and Zhu, 2022a), and other sentences discussing the same entity sampled from the article (Maab et al., 2023b). Although these approaches lead to substantial improvements, all of them have so far focused on model fine-tuning, which limits their applicability in broader scenarios.

This paper investigates whether bias detection can profit from current developments on large language models (LLMs), which can achieve excellent performance on a wide variety of downstream tasks utilizing zero-shot or few-shot approaches (Brown et al., 2020; Kojima et al., 2022), i.e., without fine-tuning. Previous work (Beltagy et al., 2022; Shin et al., 2020; Schick and Schütze, 2020) shows the significance of in-context learning for a large variety of tasks. Our work continues in this direction, and we propose to test prompting strategies for bias detection using LLMs. Our prompt design is motivated to a large extent by its *flexibility* as an experimental platform: we can investigate the interaction between various experimental variables, the most important of which are: (a) the style of the prompt (concise vs. detailed); (b) the amount of document context presented to the model; (c) the amount of supervision afforded to the model (zero-shot vs. few-shot). We carry out experiments regarding these variables on a range of current LLMs, empirically evaluating zero-shot and few-shot learning approaches using task-specific prompts to detect political bias across a wide range of LLMs. To the best of our knowledge, our work is the first to attempt this kind of investigation.

Our work shows the importance of the specificity and detail in prompt engineering, which is a highly task-dependent and a potentially cumbersome process. We additionally show that providing an LLM with more appropriate and consistent bias contexts, either in the form of examples, or of related sentences, can approximate conventional supervised learning approaches, setting up a research direction for the future.

2 Related Work

The societal consequence of misinformation in various fields are substantial. News articles can contain biased opinion, resulting in misleading views (Gentzkow and Shapiro, 2010). Political scientists have determined that bias in news reporting can be defined by the choices in content selection and organization of information within articles (Prat and Strömberg, 2013; Gentzkow et al., 2015). According to Shapiro (2016), bias can also be introduced when less information or facts on news articles is provided because for journalists who try to appear neutral, avoids conveying ample information. In another study, Fletcher and Park (2017) show a negative association between trust in the news media and online news participation.

Since news media advance their interests by devoting resources to control reporting (Entman, 2007; Chang et al., 2019), the NLP field has devoted attention to predicting the political leaning and trustworthiness of news media outlets (Baly et al., 2019; Mehta et al., 2022). Regarding bias in news, the introduction of the BASIL dataset (Fan et al., 2019) marks the onset of interesting paradigms. Within BASIL, sentences are annotated with their corresponding bias type, a designated target (the primary entity), and several other labels. Prior studies on BASIL have shown that the incorporation of *contextual information* enhances supervised learning models: van den Berg and Markert (2020) utilized article and event-level context; Guo and Zhu (2022a); Maab et al. (2023b) integrate three levels of context, i.e., article, event, and adjacent sentences, and Chen et al. (2020) utilized second order bias features to detect article-level bias.

Generative models are outlined as complex systems by (Holtzman et al., 2023) due to their tendency to exhibit emergent behaviors, for example, the ability of LLMs to perform in-context learning (Törnberg, 2023). Along these lines, instructing LLMs with prompt in zero or few-shot settings enables models to perform tasks with no or minimal task-specific training, which is evidence of the LLMs' generalization capability (Gao et al., 2020b), and has put LLMs at the center of recent progress in NLP (Devlin et al., 2018; Brown et al., 2020; Thoppilan et al., 2022; Rae et al., 2021; Liu et al., 2022).

In essence, LLMs offer a broad spectrum of beneficial applications that extend beyond their generative functionalities. Task division into multiple concise steps and sequentially introducing them to the large language model yields improved performance across an array of reasoning tasks, spanning word problems, arithmetic operations, and code execution (Anil et al., 2022; Hao et al., 2022). Notably, GPT models have achieved success in diverse

language-related endeavors, showcasing their capacity to produce text resembling human-written content (Radford et al., 2019; Brown et al., 2020).

3 Method

Recent studies have shown that fine-tuning of LLMs to instruction-style prompts is a common approach to achieve gains in performance (Ouyang et al., 2022; Wei et al., 2021; Min et al., 2021; Sanh et al., 2021). However, language model fine-tuning is also computationally very expensive. Consequently, based on our comprehensive analysis on how models for language representation have been incorporating broader contextual scopes into their predictive processes, for example, models have evolved to consider neighboring words (Mikolov et al., 2013), sentences (Peters et al., 2018; Kiros et al., 2015; Maab et al., 2023b), paragraphs (Devlin et al., 2018; Radford and Mikolov, 2018), and even articles (van den Berg and Markert, 2020; Guo and Zhu, 2022a), we align analogous contexts with prompts to learn the capacity of context dependence across different large language models.

Our basic setting adopts a zero-shot approach with prompts shown in Table 1. We follow Maab et al. (2023b) by modeling the two bias classification tasks of INF/OTH and INF/LEX, where 'INF' refers to sentences with informational bias, 'LEX' refers to sentences with lexical bias and 'OTH' denotes the combination of neutral and lexically biased sentences. We employ two prompts, as detailed in Table 1, a simple one (CONCISE), and one with a definition of the two bias types (DETAILED). Our goal is to assess how the difference between these prompts influences the LLMs' capability for bias classification. We evaluate the two prompts in the following settings.

Zero-shot In this approach, using zero-shot means we do not give the model any knowledge of question-answer pairs as to how a certain sentence gets classified as a particular label, nor fine-tuned the language models in any aspect.

Context-augmented (+CTX) Given the context-sensitive of the task, we find that a large portion of previous work in detection of political bias on BASIL has focused on introducing contextual information into classification (Cohan et al., 2019; van den Berg and Markert, 2020; Guo and Zhu, 2022a), for example, by mixing contexts of informational and lexical bias at both the article-level (entire article encompassing target sentence) and event-level (triplet of articles discussing the same event). In this context, recent work by Maab et al. (2023b) proposed a comprehensive framework to generate more consistent and similar bias contexts to improve performance when fine-tuning models. In this paper, we propose to adapt these techniques for the task of zero-shot and few-shot political bias detection. Concretely, we take advantage of three context-augmented techniques:

Bias-Aware Neighborhood (BANC) extends the target sentence with neighboring spans, i.e., combining former and next sentence with target sentence.

Article-Based Target-Aware (ABTA) extends this idea by extending the target sentence with spans from the same article which share both the same bias label (bias type of BASIL, i.e., informational and lexical) and the same target, i.e., main entity or topic of the target sentence already annotated in BASIL (cf. Figure 1a).

Event-Based Target-Aware (EBTA) extends the ABTA approach by gathering spans for the same bias label and target across articles, i.e., triplets of BASIL (cf. Figure 1b).

Utilizing annotated bias spans gathered through these approaches are added to the input prompt, which effectively mix relevant local and global contexts of informational and lexical bias at the article and event levels.

Few-shot (+K-SHOTS) Our experiments with few-shot learning are parallel to the zero-shot case, except that we present the model with question–answer (i.e., sentence–label) pairs. Since the available data is limited, the choice of parameter k (the number of examples provided) is dependent on the task. More details about this are given in §4.

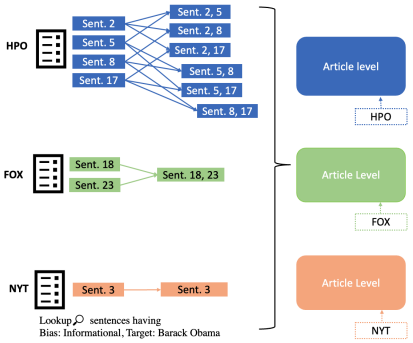
4 Experimental Setup

4.1 Tasks and Data

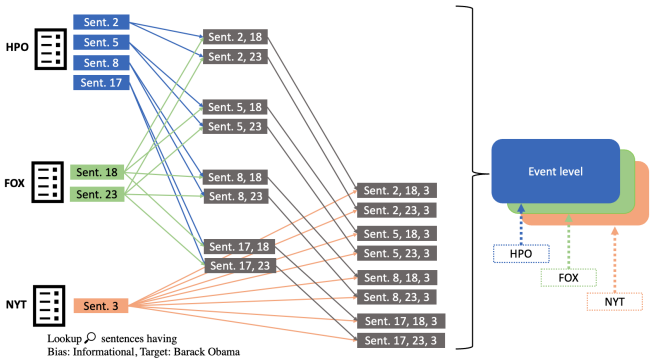
Utilizing the target-aware contextualization techniques by Maab et al. (2023b), we incorporate a total of 4,987 informational and 1,551 lexical contextualized sentences. The remaining 6,250 sentences named "OTH", originating from the foundational work by Fan et al. (2019), do not exhibit bias. Our approach aligns with the task formulations outlined in Maab et al. (2023a,b) for both the INF/OTH and INF/LEX tasks, as mentioned in §3. For comparison of sentences devoid of bias contexts, we

Prompt Style	Prompt Template	Task	
		INF/OTH	INF/LEX
CONCISE	Given the following text, classify it as either c_a or c_b, Text: s, Classification:	c_a: "bias" c_b: "no bias"	c_a: "informational" c_b: "lexical"
DETAILED	I want you to act as a bias detector. The definition of "bias" is it uses speculative and tangential clauses in a factual and neutral way to sway readers. In contrast, 'no bias' means that the information is presented in a fair and impartial manner, without any undue influence or prejudice. Classify if the following text exhibits c_a or c_b. Text: "s", Classification:	c_a: "bias" c_b: "no bias"	-
	I want you to act as a bias detector. The definition of "Informational bias" is it uses speculative and tangential clauses in a factual and neutral way to sway readers. In contrast, the definition of "Lexical bias" is it depends on linguistic attributes like word choice and syntax. Given the definitions, classify if the following text exhibits c_a or c_b. Text: "s", Classification:	-	c_a: "informational" c_b: "lexical"

Table 1: The strategy for formulating two prompts for the INF/OTH and INF/LEX tasks, where c_a and c_b denote the labels associated to each class for each task, and s denotes the target sentence.



(a) Article-based Target-aware ABTA obtains contexts within the article.



(b) Event-based Target-aware EBTA obtains contexts across articles.

Figure 1: Two contextualization methods for a target sentence with informational bias on the target "Barack Obama" and three news sources of BASIL i.e., FOX, HPO, and NYT.

CONTEXT	Per-class Examples	
	INF/OTH	INF/LEX
Yes	4,987 / 6,250	4,987 / 1,551
No	1,221 / 6,250	1,221 / 462

Table 2: Task datasets for INF-vs.-OTH and INF-vs.-LEX with and without context-augmented examples.

employ the original BASIL sentences, comprising 1,221 examples of informational and 462 examples of lexical bias (Fan et al., 2019). Table 2 summarizes the details of the examples available for each task and setting.

Regarding K-SHOTS experiments, in the INF/OTH task, we employ 5-shot examples, comprising 3 examples derived from INF sentences and 2 examples from OTH. For the INF/LEX task, we utilize 3 INF examples and 3 LEX examples. The slight variation in the K-SHOTS of the INF/LEX task aims to ensure that the model receives equal exposure to LEX bias information, considering its limited size.

4.2 Models

For all tasks, large language models are used without any fine-tuning or gradient updates, and the zero-shot and few-shot experiments are specified only as prompts. Furthermore, supplementary experiments are also conducted to involve k-shot experiments to delve into the influence of k examples on model performance. We consider a range of LLMs with various parameter counts and in-context learning abilities as measured by standard benchmarks. We also consider models that are instruction fine-tuned. Within our proposed framework, concretely, we work with the family of FLAN instruction-tuned models which includes FLAN-T5-Base, FLAN-T5-XL, FLAN-T5-XXL, (Chung et al., 2022) and FLAN-UL2 (Tay et al., 2022). We also consider two publicly-available regular (causal) LLMs, namely GPT-Neo (Black et al.,

Model	Training Data	# Par.
Flan-T5 (Google)	Based on T5 and Raffel et al. (2019) and later fine-tuned on 1,836 tasks by combining four mixtures from prior work, Muffin (80 tasks) (Wei et al., 2021), T0-SF (193 tasks) (Sanh et al., 2021), NIV2 (1,554 tasks) (Wang et al., 2022), and CoT (manual annotations for reasoning tasks)	80M–11B
FLAN-UL2 (Google)	Model pre-trained on hundreds of gigabytes of clean English text named "Colossal Clean Crawled Corpus" (C4) (Raffel et al., 2019), acquired by web scraping, and then same as above.	20B
GPT-3.5 -turbo (OpenAI)	Reportedly, trained on hundreds of billions of data from a diverse range of data sources, including Common Crawl, web documents, books, and Wikipedia. (Brown et al., 2020).	175B
GPT-4 (OpenAI)	Reportedly, a model based on Mixture-of-Experts, which consists of 16 different experts working together, each with approx. 110b parameters and trained for a specific task/field.	1.76T
OPT (Meta AI)	Causal language model pre-trained on five datasets: three RoBERTa datasets (Liu et al., 2019), the Pile (Gao et al., 2020a), and Pushshift.io Reddit (Baumgartner et al., 2020).	125M–175B
GPT-Neo (EleutherAI)	Causal language model pre-rained on a 800GB diverse English text corpus, the Pile (Gao et al., 2020a)	1.3B–20B

Table 3: Details on key properties of various LLMs, including pre-trained data and parameter count.

2021) and OPT (Zhang et al., 2022). Finally, we consider the two popular black-box models ChatGPT/ GPT-3.5-turbo and GPT-4. Table 3 provides detailed properties on the LLMs that we use in our study.

4.3 Answer Parsing

Some models, like the ones from the FLAN family, which are instruction-tuned solely, produce the necessary target label, which makes interpretation trivial. Other model families, like GPT-Neo and OPT, sometimes return long and intricate answers which require additional mapping onto a label. We select the first portion of the response that is interpretable as an answer, and if no such answer exists, we treat the prediction as incorrect. See Appendix A for details.

4.4 Hardware

Our experiments were performed on two kinds of GPUs, 16-GB NVIDIA V100 and 40-GB NVIDIA A100. We access these GPUs by means of nodes on a large cluster, where each node has several such GPUs. Some experiments were performed using data parallelism to speed up inference time.

	INF/ OTH			INF/ LEX		
		F1 Score			F1 Score	
Model	Acc.	INF	OTH	Acc.	INF	LEX
FLAN-T5-Base	46.98	35.32	87.11	39.98	32.77	11.43
+ CTX	59.61	36.14	67.43	51.67	38.99	19.33
FLAN-T5-XL	78.36	40.66	71.23	55.98	40.45	20.12
+ CTX	76.79	39.32	63.38	59.10	46.84	27.98
FLAN-T5-XXL	71.13	41.87	69.43	52.55	47.09	38.39
+ CTX	73.67	44.89	58.34	60.86	48.22	33.00
FLAN-UL2	73.09	43.34	65.22	69.34	48.83	35.17
+ CTX	73.06	44.96	60.34	68.99	**51.02**	38.64
GPT-3.5-turbo	74.46	47.87	68.09	68.98	48.11	37.53
+ CTX	75.50	49.55	70.34	70.07	48.91	36.05
GPT-4	77.11	51.66	71.56	72.21	47.03	38.27
+ CTX	77.20	**54.04**	69.89	**74.45**	49.90	**40.71**
GPT-Neo	62.24	50.01	78.23	50.36	39.91	30.23
+ CTX	71.38	47.43	64.32	58.27	42.93	28.26
OPT	**83.01**	44.07	**79.03**	48.78	43.12	31.82
+ CTX	76.24	46.11	53.98	69.33	45.25	33.15

Table 4: Summary of our results with the CONCISE prompt in terms of accuracy and micro-F1 scores. Results in bold indicate the best performance on each task and metric across, models and settings.

4.5 Implementation Details

We use PyTorch to implement models, borrowing from HuggingFace (Face, 2021) for FLAN-T5 (Chung et al., 2022), FLAN-UL2 (Tay et al., 2022), GPT-Neo (Black et al., 2021), and OPT (Zhang et al., 2022). For ChatGPT/ GPT-3.5-turbo and GPT-4, we rely on the official OpenAI API. GPT-3.5-turbo and GPT-4 were tested with low and high settings of temperature (0, 0.5, and 0.8), however we did not observe much variation between the generated texts.

5 Results

On our first set of experiments, we use the CONCISE prompt, which we test on regular baseline (zero-shot) and also in the CTX setting. Results are shown in Table 4. After pilot experiments, we observed that models could only provide marginal performance improvements when provided with long contexts, so we did not consider few-shot learning using this prompt. In Table 4, overall we observe that with target-aware contextual information CTX, advance LLMs result in enhancements in F1-scores and accuracy for both INF/OTH and INF/LEX bias tasks. For instance, when context is utilized for FLAN-T5 and its variants, FLAN-UL2, GPT-3.5-turbo, and GPT-4, INF/OTH task shows a rise in INF-F1 scores against regular, respectively. It can be seen that GPT-4 substantially

outperforms other models, where INF/OTH task shows INF F1-score of 54.04 against 51.66 of regular, and INF/LEX task shows INF F1-score of 49.90 against 47.03 of regular, respectively. For GPT-Neo, it is found that they perform well even with no context augmented information. However, GPT-Neo encounter challenges when working with extended contexts (Yang et al., 2022). OPT also seems to have increased performance with CTX given concise prompts as discussed in section 5.3.

Owing to the fact that contextual information raise performance in LLMs, our findings are well-aligned with prior work, which implies that LLMs are capable of classifying certain kinds of bias. We include additional experiments on context sensitivity to assess LLM capacities in comprehending contextual information. These experiments involve extensive prompts, and additional tests are conducted with k-shot context examples. Table 5 shows a summary of our experiments. Through this study, we aim to determine whether context enhances the model's ability to efficiently process more extensive prompts when combined with k-shot examples.

When we combine k-shot with context examples, i.e., CTX + K-SHOTS, we note that the performance starts improving against zero-shot and simple k-shot experiments. Using an extensive prompt, our results further demonstrate the effectiveness of context-augmented zero-shot experiments. For instance, for the CTX scenario, FLAN-UL2 shows INF F1- score of 58.21 against 53.08 of K-SHOTS in INF/OTH task, whereas we observe INF F1-scores of 59.55 against 58.03 of K-SHOTS for the INF/LEX task, respectively. Similarly, an increase in the LEX F1-score is noticeable in the INF/LEX task for most of the LLMs as we go down from regular (zero-shot) to contextualized information, i.e., following the trend from regular to K-SHOTS to CTX and finally to both CTX + K-SHOTS. Consequently, we note that that best performance in INF F1-score of 69.07 in INF/OTH and 64.55 in INF/LEX, is achieved by GPT-4 when presented with context-augmented k-shot examples.

Most notably, using a more detailed definitions of bias types (cf. Table 5) led to consistent performance improvements over results in Table 4, especially in instruction-tuned models such as FLAN-T5 and its variants, GPT-3.5-turbo, and GPT-4. Our findings also align with the notion that compact models, such as GPT-Neo and OPT, featuring 2.7B

	INF/ OTH			INF/ LEX		
		F1 Score			F1 Score	
Model	**Acc.**	**INF**	**OTH**	**Acc.**	**INF**	**LEX**
FLAN-T5-Base	60.48	35.81	87.43	42.69	39.97	23.14
+ K-SHOTS	72.81	40.66	87.00	49.39	45.55	31.96
+ CTX	73.03	42.81	89.28	49.67	49.88	28.90
+ CTX + K-SHOTS	69.61	42.14	87.43	61.67	51.18	33.88
FLAN-T5-XL	64.08	39.32	82.38	53.20	54.21	40.90
+ K-SHOTS	65.89	43.45	84.11	63.01	52.39	38.18
+ CTX	69.13	44.09	80.24	69.29	57.23	36.76
+ CTX + K-SHOTS	71.79	53.32	82.38	67.00	54.11	40.35
FLAN-T5-XXL	74.47	46.56	77.34	59.16	53.83	42.42
+ K-SHOTS	80.34	51.68	78.77	70.82	55.35	38.65
+ CTX	78.35	48.21	79.09	72.88	56.74	38.73
+ CTX + K-SHOTS	76.67	60.71	77.34	75.34	55.09	38.12
FLAN-UL2	78.50	49.99	79.00	73.08	54.77	37.33
+ K-SHOTS	80.71	53.08	73.32	78.37	58.03	36.51
+ CTX	77.29	58.21	74.31	71.34	59.55	39.10
+ CTX + K-SHOTS	78.87	61.08	69.40	81.89	62.91	41.36
GPT-3.5-turbo	70.44	50.90	73.13	71.08	52.54	37.77
+ K-SHOTS	71.71	54.54	74.33	74.37	54.88	38.87
+ CTX	74.09	62.08	74.40	72.69	59.69	41.36
+ CTX + K-SHOTS	75.98	65.86	75.83	76.08	60.89	40.58
GPT-4	73.08	53.16	75.88	70.45	57.16	40.04
+ K-SHOTS	71.98	55.23	**76.01**	72.45	59.21	38.20
+ CTX	75.54	67.87	73.08	75.01	63.06	40.89
+ CTX + K-SHOTS	78.88	**69.07**	73.59	**82.99**	**64.55**	**42.76**
GPT-Neo	67.09	39.53	60.22	62.12	40.88	33.70
+ K-SHOTS	61.22	50.08	44.16	64.48	42.09	35.19
+ CTX	**81.09**	43.36	67.21	39.89	41.12	32.43
+ CTX + K-SHOTS	56.24	44.90	56.23	57.81	40.03	35.54
OPT	68.80	42.12	54.41	64.46	41.12	30.06
+ K-SHOTS	70.09	40.80	60.99	61.08	39.84	33.34
+ CTX	72.06	45.56	65.11	74.54	40.02	31.50
+ CTX + K-SHOTS	68.21	43.96	45.98	68.50	42.99	30.87

Table 5: Results of our experiments with the DETAILED prompt, in terms of accuracy and micro-F1 scores, detailing the performance impact when adding context as discussed in §4. Results in bold indicate the best performance overall, per metric.

parameters, tend to produce more precise results when presented with concise prompts. This is reflected in increased performance of INF/OTH and INF/LEX bias tasks of the concise prompt (Table 4) over the detailed prompt (Table 5).

Overall, it is observed that LLMs benefits the most from informational context examples, i.e., INF bias over non-bias OTH i.e., INF/OTH task. Furthermore, we note that performance improvement of DETAILED prompt over CONCISE prompt is more prominent in instruction-tuned models including FLAN-T5 variants, FLAN-UL2, GPT-3.5-turbo and GPT-4, whereas the CONCISE prompt holds more significance when presented to GPT-Neo and OPT models. We think this is because smaller language models tend to struggle when presented with long contexts, being essentially unable to parse the information provided, even if it is potentially use-

ful for the task at hand. Finally, we also find that GPT-4 outperforms GPT-3.5-turbo.

5.1 Comparison to prior work

We compare our approach against state-of-the-art results using model fine-tuning. To the best of our knowledge, no previous studies have employed LLMs for zero-shot or few shot political bias classification. Table 6 presents best performance outcomes for all selected models, as determined from the findings in Table 5 against state-of-the-art work. In case of OPT and GPT-Neo, we present the best results achieved per model, taken from both Table 4 and Table 5. We see that GPT-4, GPT-3.5-turbo, FLAN-UL2, and FLAN-T5-XXL achieve higher performance showing increased INF F1-score of 69.07, 65.86, 61.08, and 60.71, respectively, in the INF/OTH task, surpassing the best BERT model by Maab et al. (2023b) with INF F1-score of 58.15. We find that LLMs with enhanced capabilities of context utilization hold significance in contributing towards higher performance over baselines including BERT (Chen et al., 2020), RoBERTa (Lei et al., 2022), MultiCTX that uses multi-contrast learning across BASIL articles (Guo and Zhu, 2022a), and target-aware BERT (Maab et al., 2023b).

To detect informational bias from lexical bias sentences, i.e., INF/LEX task, the only baseline available is the BERT model by Maab et al. (2023b). As can be seen, we find that no LLM model is able to outperform the existing prior work in INF/LEX task, i.e., the top two performance in terms of INF F1-score of 75.46 and 74.01 is achieved by Maab et al. (2023b), followed by the INF F1-score of 64.55 with GPT-4 as third, and FLAN-UL2 as fourth with only 62.91. Since distinguishing particular types of bias, i.e., between informational and lexical bias in INF/LEX is a challenging task, and no training is performed while utilizing LLMs, none of our models obtain a performance superior to prior models with dedicated training. See Appendix B for robustness study against baseline BERT models of Maab et al. (2023a,b).

5.2 Role of k

We also explore the impact of k-shot context examples CTX + K-SHOTS to uncover the significance of context and the extent to which it affects model performance. For this purpose, we choose GPT-4, our overal best performing model, and experiment with variations of k using k-shot experiments to analyze the changes attributed by contexts examples in INF/OTH and INF/LEX bias tasks. We present k=2, 4, 6, 8, 10 context augmented examples to GPT-4. For this experiment, we allocate an equal number of examples for k-shot learning in both INF/OTH and INF/LEX bias tasks. For instance, in INF/OTH task, 2-shot experiment entails one example originating from context-augmented INF bias, and the other from OTH. The same is true for INF/LEX, where one example is sourced from context-augmented INF, while the other is from context-augmented LEX bias.

As shown in Figure 2, the F1-score of INF and LEX bias demonstrates a continuous increase as

	INF/ OTH			INF/ LEX		
		F1 Score			F1 Score	
Model	**Acc.**	**INF**	**OTH**	**Acc.**	**INF**	**LEX**
FLAN-T5-Base	73.03%	42.81	**89.28**	61.67%	51.18	33.88
FLAN-T5-XL	71.79%	53.32	82.38	69.29%	57.23	36.76
FLAN-T5-XXL	76.67%	60.71	77.34	72.88%	56.74	38.73
FLAN-UL2	78.87%	61.08	69.40	81.89%	62.91	41.36
GPT-3.5-turbo	75.98%	65.86	75.83	76.08%	60.89	40.58
GPT-4 1.78T	78.88%	**69.07**	73.59	82.99%	64.55	42.76
GPT-Neo 2.7B	61.22%	50.08	44.16	58.27%	42.93	28.26
OPT 2.7B	76.24%	46.11	53.98	69.33%	45.25	33.15
BERT	-	41.46	-	-	-	-
RoBERTa	-	46.47	-	-	-	-
ArtCIM	-	42.80	-	-	-	-
WinSSC	-	37.58	-	-	-	-
EvCIM	-	45.81	-	-	-	-
MultiCTX	-	46.08	-	-	-	-
BERT+ctx	84.90%	56.88	-	83.36%	74.01	66.97
BERT+ctx (**)	**86.40%**	58.15	-	**84.77%**	**75.46**	**71.93**

Table 6: Comparison of our best LLM settings to prior work. (**) denotes a model using extra training data. References: BERT (Chen et al., 2020), RoBERTa (Lei et al., 2022), ArtCIM (van den Berg and Markert, 2020), WinSSC/EvCIM/MultiCTX (Guo and Zhu, 2022a), BERT+ctx (Maab et al., 2023b).

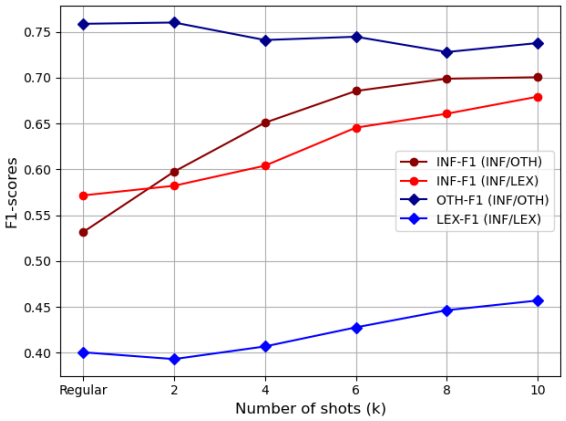


Figure 2: Influence of k in few-shot setting on GPT-4 performance on INF/OTH and INF/LEX bias tasks

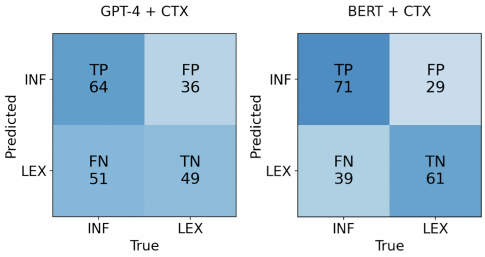


Figure 3: Comparison of INF/LEX results on hundred randomly selected examples of each using GPT-4 + CTX and BERT + CTX

we increment the number of shots provided, uncovering the significance of context in the manner of examples. It is also observed that no significant performance change occurs after k approaches 6, however there is a slight rise in F1-scores of INF and LEX from 6-shot, when 10-shot contextual examples are introduced. These results indicate that having an augmented context as shot as input to the LLM leads to a notable performance improvement, while GPT-4 only needs a few examples for classifying political bias on BASIL.

5.3 Error Analysis

To gain a deeper insight into the zero-shot abilities of LLMs in our tasks, we performed a qualitative analysis of the predictions made by GPT-4 when subject to different prompting techniques based on context. Table 7 shows samples of correct and incorrect contexts extracted from the different sources (article/event) which result in different predictions, deviating the model from its baseline performance. We noted that in those examples where the context aligns with the target sentence, the model is generally able to make the right prediction. However, there are still cases where there are nuanced representations of context, on which the model faces difficulties in understanding the overall input, resulting in incorrect predictions.

We also compare the performance of fine-tuned BERTs (Fan et al., 2019; Chen et al., 2020; Maab et al., 2023a,b) against our approach with GPT-4 on a set of a hundred randomly selected samples from both informational bias and lexical bias categories without context, denoted as "Regular (No context)." These were then contrasted with augmented context samples of the same examples using BANC, ABTA, and EBTA from Maab et al. (2023b). See Appendix B for additional details on this study.

Figure 3 shows confusion matrices summarizing

	Model Input	Pred.
-	At one point he suggested that those who opposed his position were heartless.	X
BANC	*Mr. Perry was widely criticized within his own party for backing a state plan that gives some children of illegal immigrants the same lower-cost in-state college tuition that is enjoyed by American citizens who attend the state's public universities and who have been Texas residents at least three years.* At one point, he suggested that those who opposed his position were heartless.	✓
EBTA	At one point he suggested that those who opposed his position were heartless. *You said I don't want to build a fence, Romney said. You talk about magnets, you put in place a magnet. Perry has also said the federal government should extend work visas permitting undocumented immigrants to move freely between the U.S. and their home countries, but stressed that he still opposes amnesty or a path to citizenship.*	✓
-	On that same day, Obama had lunch with Clinton.	X
BANC	*The report comes one day after President Obama insisted nothing improper happened with Sestak.* On that same day, Obama had lunch with Clinton. *This is punishable by prison.*	✓
EBTA	On that same day, Obama had lunch with Clinton. *As the Huffington Post reported on Thursday, various political historians and ethics lawyers have approached the Sestak news with yawns, noting that quid pro quos and backroom job offers are fairly common in administrations.*	✓
-	I'm very frustrated.	✓
BANC	*I think this is overdue, and I think other states should jump on board, Overman said of the lawsuit.* I'm very frustrated. *I take an oath of office, as does every other police officer in this country.*	X
EBTA	I'm very frustrated. *Colorado's attorney general, John Suthers, a Republican, said in a statement that the challenge from Nebraska and Oklahoma was without merit. Like many elected officials in Colorado, Mr. Suthers had opposed Amendment 64, which legalized marijuana. But on Thursday, he said we will vigorously defend against the lawsuit attempting to undo it.*	✓
-	A leader only starts a fight he's prepared to finish.	X
BANC	*But any drunken redneck can walk into a bar and start a fight.* A leader only starts a fight he's prepared to finish. *The field of confirmed and potential GOP presidential candidates includes more than a dozen people.*	X
EBTA	A leader only starts a fight he's prepared to finish. *And if someone can capture both the blue-collar, working-class Republicans, the conservatives, many of them even union members, as well as evangelicals, there's a real pathway to the nomination. Huckabee, an ordained Southern Baptist minister, is a celebrated figure among evangelical Christians. He was the longest-serving Arkansas governor, from 1996 to 2007.*	X

Table 7: Examples of zero-shot DETAILED prompt performance of GPT-4 on the INF/OTH bias task when using our context-augmented prompting techniques. In the table, italic portions of the model input denote context obtained using the corresponding technique.

our obtained results, where GPT-4 exhibits difficulties in predicting lexical bias. The reasons for the weak performance of LLMs in detecting INF/LEX bias, in comparison to pre-trained BERT models, can be summarized by various factors. Firstly, the INF/LEX bias detection task is more challenging due to the potential scarcity of lexical bias sentences even after context-augmentation, and their limited span within bias sentences. Notably, informational bias spans predominantly encapsulate entire sentences, while lexical bias spans are ob-

served in concise words or phrases (Fan et al., 2019; Maab et al., 2023a). Similarly, according to Chen et al. (2020), distinguishing informational and lexical bias is more difficult compared to detecting any type of bias. Secondly, pre-trained BERT models (Maab et al., 2023a), benefit from fine-tuning on tasks with ample training data, i.e., when combined with additional augmented context using backtranslation (BT), the detection of both informational and lexical bias becomes more apparent and fluent. Consequently, this leads to superior performance compared to language models (LLMs) that in this work are utilized without any training or fine-tuning.

5.4 CONCISE vs. DETAILED Prompt

Finally, we also compare the performance of zero-shot CONCISE versus zero-shot DETAILED prompting. Although DETAILED prompting enhances overall performance, OPT demonstrates superior performance when CONCISE prompting is employed. We examined bias samples to showcase how OPT benefits from context information with a CONCISE prompt. As found by Zhang et al. (2022), OPT models such as OPT-175B struggle with declarative instructions or straightforward interrogatives. Our DETAILED prompt aligns with this observation when compared with the CONCISE prompt, which affords more flexibility. Similarly, in line with our findings, OPT benefits from the addition of context information. See Table 10 in Appendix for more details.

To explore OPT, we conducted a comparison between Zero-shot-CONCISE and Zero-shot-DETAILED prompts specifically for the task of INF/OTH i.e., detecting informational bias only. This analysis also highlights the distinctions in contextual information over regular non-context samples. In Table 10, the examples selected include text with and without contextual information, where Incorrect predictions are attributed to the absence of a sufficient amount of context for the target sentence. We observe that OPT performs well when context makes sense, particularly with CONCISE prompts. For instance, in Table 10, the third biased sentence sample "He also praised the campaign that Mr. Sanders ran." is integrated with a context involving Obama endorsing Clinton and discussing the candidacy. The inclusion of this context fails to provide an explanation for the target 'Bernie Sanders,' leading to an incorrect model prediction for both CONCISE and DETAILED prompts, contrary to the reasonable context in the second example, "He called it a very good night, and said Holder's visit had let people know their voices had been heard," where the context also describes Holder's visit as receptive to finding solutions and offering assistance to the target 'Eric Holder'. Thus the model makes the correct prediction even with only the CONCISE prompt. However, OPT capability is limited in detecting irregular context when only a CONCISE prompt is provided. This analysis of samples reveals that certain contexts do not contribute significantly to correct predictions. The primary trend suggests that the zero-shot approach with CTX tends to be more effective.

6 Conclusion

This paper aims at advancing our understanding of prompt-based identification of media bias by LLMs, a crucial and challenging task to the research community, news media, and social media companies. We establish a framework for examining prompt-based models in a zero-shot and few-shot configuration to classify sentences that manifest bias. Our approach demonstrates the utility of using context-augmented informational and lexical bias from prior work, an area that we think has received inadequate attention. We provide insights into the dynamics between LLMs, tasks, and prompts, discerning their individual capacities in the detection of bias. Notably, we find considerable performance improvements when adding context-augmented information to both small and large-parameter LLM models, indicating that in-context learning can approach the performance levels of traditional models in classifying media bias. While our study demonstrates that LLMs are applicable to the area of misinformation detection, they do not always represent a huge improvement. Formulating prompts that strike a balance between the simple and the (over)-elaborate remains a challenge, as is particularly evident in GPT-Neo and OPT.

We hope our work inspires further research in bias detection as a means to gain insight into enhancing LLMs. Future work includes exploring additional aspects of bias, e.g., bias related to culture, race, or age, as well as exploring the understanding and limitations of LLMs. We also think other similar bias-detection tasks could be grouped together in our prompting setting, following Lee et al. (2021).

Limitations

The topic of bias detection using zero and few-shot prompts is relatively young and not addressed properly. Therefore, the amount of existing probable directions to research is immense. As there is a scarcity of bias representations and annotated media coverage in other languages, our work is exclusively founded on bias representations of English news articles, and BASIL stands alone as the sole annotated dataset containing informational bias. Also, the state-of-the-art models used for comparison are not competitors to the models used in this paper. Hence, another direction of future work is to implement similar kind of models for understanding bias and discovering knowledge limits of large language models.

Ethical Considerations

Political entities are subjects of continuous debate in news media, and it is crucial to necessitate the development of fast and reliable methods to disseminate accurate information. This becomes especially challenging because conventional bias detection models encounter rigorous training. To address this, we must adopt a more flexible and tolerant approach, especially when utilizing large language models. However, it is essential to strike a balance between bias detection, model's adaptability, and the preservation of diverse viewpoints.

Finally, we think that one major potential risk of our work is that given the simplicity and effectiveness of our prompt-based techniques, companies may choose to deployed in online products and services that are available to the masses. However, since in this paper we have not tested nor analyzed the predictions from a qualitative point of view, it is impossible for us to foresee potential impacts that this may have in the future at this point. For example, it is plausible that the predictions of our models may contain other kinds of biases, which may be amplified when deployed at a massive scale. We urge for extreme caution in utilizing our techniques in production-level environments, and call researchers and practitioners to help by conducting studies to further understand the nature and potential repercussions of using the outputs of our models in broader contexts.

Acknowledgements

The authors wish to express gratitude to the funding organization as this work has been supported by the Mohammed bin Salman Center for Future Science and Technology for Saudi-Japan Vision 2030 at The University of Tokyo (MbSC2030).

References

Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed Abdelmajeed, Atif Mehmood, and Muhammad Tariq Sadiq. 2020. Document-level text classification using single-layer multisize filters convolutional neural network. *IEEE Access*, 8:42689–42707.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. 2022. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35:38546–38556.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. *arXiv preprint arXiv:1904.00542*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Iz Beltagy, Arman Cohan, Robert Logan IV, Sewon Min, and Sameer Singh. 2022. Zero-and few-shot nlp with pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 32–37.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. Technical report. If you use this software, please cite it using these metadata.

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.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al-Khatib, Benno Stein, and Henning Wachsmuth. 2020. Detecting media bias in

news articles using gaussian bias distributions. *arXiv preprint arXiv:2010.10649*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. 2019. Pretrained language models for sequential sentence classification. *arXiv preprint arXiv:1909.04054*.

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

Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

Hugging Face. 2021. The ai community building the future. *URL: https://huggingface. co*.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Richard Fletcher and Sora Park. 2017. The impact of trust in the news media on online news consumption and participation. *Digital journalism*, 5(10):1281–1299.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020a. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020b. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Matthew Gentzkow, Jesse M Shapiro, and Daniel F Stone. 2015. Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1, pages 623–645. Elsevier.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511, Boulder, Colorado. Association for Computational Linguistics.

Shijia Guo and Kenny Q Zhu. 2022a. Modeling multi-level context for informational bias detection by contrastive learning and sentential graph network. *arXiv preprint arXiv:2201.10376*.

Shijia Guo and Kenny Q. Zhu. 2022b. Modeling multi-level context for informational bias detection by contrastive learning and sentential graph network.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*.

Ari Holtzman, Peter West, and Luke Zettlemoyer. 2023. Generative models as a complex systems science: How can we make sense of large language model behavior? *arXiv preprint arXiv:2308.00189*.

Christoph Hube and Besnik Fetahu. 2019. Neural based statement classification for biased language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Pascale Fung, Hao Ma, Wen-tau Yih, and Madian Khabsa. 2021. On unifying misinformation detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5479–5485, Online. Association for Computational Linguistics.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander G Hauptmann. 2006. Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 109–116.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes

good incontext examples for gpt-3. In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114.

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*.

Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. 2023a. An effective approach for informational and lexical bias detection. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 66–77.

Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. 2023b. Target-aware contextual political bias detection in news. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 782–792, Bali, Indonesia.

Nikhil Mehta, María Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jeff Z Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I 17*, pages 669–683. Springer.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Andrea Prat and David Strömberg. 2013. The political economy of mass media. *Advances in economics and econometrics*, 2:135.

Alec Radford and Tomas Mikolov. 2018. Improving language understanding by generative pretraining. Technical report, OpenAI.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Jesse M Shapiro. 2016. Special interests and the media: Theory and an application to climate change. *Journal of public economics*, 144:91–108.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Esther van den Berg and Katja Markert. 2020. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6315–6326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. 2022. What gpt knows about who is who. *arXiv preprint arXiv:2205.07407*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Parsing the output of the LLMs

Table 8 shows two examples of LLM answers generated by OPT and GPT-Neo that require additional processing to map onto expected answers ("answer cleansing"). Table 9 provides additional details on the algorithm we apply.

B Robustness study

Table 11 provides a detailed ablation analysis of the contribution of the different contextualization strategies to various base LLM models on the INF/LEX bias classification task. We re-implemented the BERT models of (Maab et al., 2023a,b) using HuggingFace (Face, 2021), utilizing BERT-base (Devlin et al., 2018), with a batch size of 32, a learning rate of 5×10^{-5}, and 15 epochs. Finally, Table 12 illustrates these findings with concrete examples.

INF/OTH	Model Input: INF with CTX (Bias)	Model Input: OTH (No bias)
CONCISE Prompt	Given the following text, classify it as either 'bias' or 'no bias' Text:"The move is almost certain to be challenged in court. Any crisis on our border is of President Trumps own making: family separations, child detention, turning our backs on asylum seekers, and more.There is no national emergency, California Attorney General Xavier Becerra, who has sued the Trump administration multiple times, said in a statement.", Classification:	Given the following text, classify it as either 'bias' or 'no bias' Text:"The stalemate caused a record-breaking partial government shutdown that forced an estimated 800,000 government employees to go without pay for 35 days." Classification:
OPT (2.7B) Output	B ✓pick first letter "B" for Bias	'No bias' ✓ pick first letter "N" for No bias According to the government, this is the first time a shutdown has lasted longer than a week since 1975. Classification: 'N/A' Text: The President said that the Democrats are not trying to fund his $5 billion wall. Text: It is unclear if the President actually said this. Classification: 'N/AW
GPT-Neo (2.7B) Output	'Bias' or 'No bias'. ✓pick first letter i.e., "B" of Bias 'Tremendous damage' done by President Trump 'by his own words'. When he tweeted that he was going to force the government to pay for the wall, he was making a prediction, not a promise. Nancy Pelosi, D-Calif., a staunch Democrat who is up for reelection in 2017	bias. X The government shutdown was the longest in US history and left 800,000 people to go without pay. As a result, this article was written. A shutdown is a kind of an act of war. President Trump January 11, 2017

Table 8: Answer cleansing examples

Prompt Format	Cleansing Approach	Pseudocode
Bias or No bias	Pickup the first letter 'B' or 'N' (uppercase or lowercase) encountered in the text and removal of unnecessary expressions	prediction = re.sub(r'["/' / n / . / s]', '', pred) prediction = prediction.replace(",", "") prediction = re.findall(r'[A-Za-z]', prediction) foundletter = next((letter for letter in prediction if letter.upper() in ('B', 'N')), None)
Informational or Lexical	Pickup the first letter 'I' or 'L' (uppercase or lowercase) encountered in the text	foundletter = next((letter for letter in prediction if letter.upper() in ('I', 'L')), None)

Table 9: Details of answer cleansing approach

INF/OTH	Model Input and Prompt Answer
Model Input (No context)	The NRA said it needs time to digest the findings, but it commended Hutchinson and his team for contributions that will go a long way to making America's schools safer.
CONCISE	'No bias'. X According to the NRA, the study "finds that handguns do not increase crime."
DETAILED	'No bias'. X According to the NRA, the report "will serve as a catalyst for continued progress toward universal background checks, mental health reform and school safety reforms, including improved law enforcement training and the adoption of common sense firearms laws."
Model Input + CTX	The NRA said it needs time to digest the findings, but it commended Hutchinson and his team for contributions that will go a long way to making America's schools safer. *The report marks the culmination of a controversial process that began in December, when NRA CEO Wayne LaPierre held a press conference the week after the shooting in Newtown, Conn. The 225-page report also called for states to require schools to conduct a safety assessment and for a pilot program to be created assessing mental health as a pre-indicator for violence. Mark Mattioli, the father of one Newtown shooting victim, also spoke after Hutchinson's announcement to applaud the review effort. The review comes as Congress and a number of state legislatures pursue gun control legislation. This is a recommendation for solutions, real solutions that will make our kids safer. That's what we need. At least one state, Indiana, is considering the idea of armed officers at schools.*
CONCISE	'bias' ✓ revert
DETAILED	'bias' (a). ✓ It does not address the root causes of violence, nor does it suggest the specific actions needed to combat it; rather, it identifies a number of potential solutions. It does so with an airy disregard for individual rights and without a clear understanding of how many of the proposals would affect gun ownership and Second Amendment rights in practice."
Model Input (No context)	He called it a very good night, and said Holder's visit had let people know their voices had been heard.
CONCISE	'No bias'. X He was the first to admit that the visit to the local community had not met with universal approval, but he said the fact it had brought people together was "extraordinary".
DETAILED	'No bias'. X By the next day, the story had fallen entirely silent. Classification: 'Bias' Text: One of the first things Holder said Sunday, he added, was 'let me say we need a justice system that really represents the African-American community. A bias detector would give an output of true (this is neither bias nor no bias) if the text is either. A bias detector would have a false positive (false positive) if it's output is false;
Model Input + CTX	*Johnson said there were six arrests Wednesday, compared to 47 the previous night.* He called it a very good night, and said Holder's visit had let people know their voices had been heard. *In nearby Clayton, a grand jury began hearing evidence to determine whether Wilson should be charged in Brown's death. We didnt know it was Eric Holder [coming], Rev. John Paul Hopping of Our Lady of Guadalupe, told HuffPost. They were really nice and they were really listening. They are looking for solutions and they want to help.*
CONCISE	'bias'. ✓ According to the article, the riot that ensued erupted over a rumor that a protester had been shot, CNN reported. The city of Cleveland issued a curfew at 9 p.m., and police arrested a 30-year-old man on curfew violation.
DETAILED	'No bias'. X Obama will have to wait until the summer before trying to restart his push for immigration reform. In remarks at the National Press Club, Obama said its an issue that will have to be addressed legislatively.
Model Input (No context)	He also praised the campaign that Mr. Sanders ran.
CONCISE	How X he got so many votes remains a mystery. However, there is nothing unusual about the campaign.
DETAILED	'No bias'. X My friend and colleague, the Vice President of the United States, said one of the reasons he thought Mr. Trump would be so easy to beat is that voters would have no interest in voting for an authoritarian.
Model Input + CTX	*Shortly after their meeting, Mr. Obama endorsed Mrs. Clinton in a video.* He also praised the campaign that Mr. Sanders ran. *The visit came a day after the senator huddled with his team at his headquarters in Vermont to discuss the fate of his candidacy.*
CONCISE	'No bias' X
DETAILED	'No bias'. X Shortly after their meeting, Mr. Obama endorsed Mrs. Clinton in a video. He also praised the campaign that Mr. Sanders ran.

Table 10: Contrasting contextual information vs. regular non-context samples: A comparison of zero-shot-CONCISE and zero-shot-DETAILED with OPT, illustrated with few sample examples.

	INF/ LEX						
	OPT	**FLAN-UL2**	**GPT-3.5-turbo**	**GPT-4**	**BERT + BT** (Maab et al., 2023a)	**BERT (*)**	**BERT (**)**
Regular (No Context)	32.8	37.9	45.8	48.7	**51.4**	-	-
+ BANC	33.5	41.0	52.5	54.6	-	58.6	**62.3**
+ ABTA + EBTA	32.9	43.1	51.7	55.1	-	63.3	**65.5**
+ CTX	35.8	50.5	56.8	59.5	-	67.6	**71.3**
+ K-SHOTS	33.4	46.1	52.9	**53.7**	-	-	-
+ CTX + K-SHOTS	38.6	58.2	62.4	**65.5**	-	-	-

Table 11: Robustness study of zero-shot DETAILED prompts using Regular (No Context), CTX (combined context of BANC, ABTA, and EBTA (Maab et al., 2023b)), K-SHOTS (no context few-shots), and CTX + K-SHOTS (with context few-shots) using various LLM models and reimplementation of BERT in using BERT + BT from Maab et al. (2023a), BERT (*) = BERT + BANC + ABTA + EBTA Maab et al. (2023b), and BERT (**) = BERT + BANC + ABTA + EBTA + BT Maab et al. (2023b). We randomly picked up 100 samples from informational bias and 100 samples from lexical bias from BASIL without context named as Regular (No Context). Using the target-aware techniques, BANC, ABTA, and EBTA from (Maab et al., 2023b), we contextualized the picked samples termed CTX. For Regular (No Context), BANC, ABTA+ EBTA, and CTX, LLM models, such as OPT, FLAN-UL2, GPT-3.5-turbo, and GPT-4, use DETAILED prompt. In case of fine-tuned BERT models, no context-augmented data is provided during testing, however out of the remaining picked-up samples the relevant context-augmentation technique is applied for training data with non-overlapping samples. Please note that BERT+BT and BERT(**) model requires additional data using BT over context-augmented data for training (Maab et al., 2023a,b), where BT stands for backtranslation.

INF/LEX	**Model Input**	**GPT-4**	**BERT (*)**
Informational	His advisers have complained that Democrats are slowing the process, and resisting the kind of swift confirmation the Senate gave many of Obama's nominees in 2009.	✓	✓
	AARP's new and welcome position is a positive step towards the type of reforms I've championed, and I look forward to working with the organization to shape the changes in a way that makes the least detrimental impact to present and future retirees, she said in a statement.	X	✓
	At one point he suggested that those who opposed his position were heartless.	✓	X
	Cordova is just one of 198 people that Weiner follows on Twitter, though he has nearly 50,000 followers.	X	X
Lexical	On Saturday, the mayor of Hoboken, Dawn Zimmer, said Mr. Christie's lieutenant governor and another senior administration official had **threatened** in May to withhold federal recovery aid for Hurricane Sandy unless she supported a development favored by the Christie administration.	✓	✓
	Gay and civil rights groups **praised** the ruling.	X	✓
	WASHINGTON – The U.S. Attorney's office in New Jersey has subpoenaed documents from the reelection campaign of New Jersey Gov. Chris Christie (R) and the New Jersey Republican State Committee, as part of its investigation into the "Bridgegate" **scandal.**	X	✓
	The Christie administration closed down two of the three George Washington Bridge access lanes in Fort Lee, N.J., in September, in what appeared to be a **political retribution scheme** aimed at the borough's Democratic mayor.	X	X

Table 12: Classification results produced by GPT-4 + CTX and BERT(*) + CTX using Zero-shot-DETAILED prompt. The lexical bias spans are highlighted in bold.