Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Bike Frames: Understanding the Implicit Portrayal of Cyclists in the News

Xingmeng Zhao[1], Dan Schumacher[1], Sashank Nalluri[1], Suahana Shrestha[1], Xavier Walton[2], and Anthony Rios[1]

[1]The University of Texas at San Antonio
[2]Alamo Colleges
{xingmeng.zhao, anthony.rios}@utsa.edu

Abstract

Increasing cycling for transportation or recreation can boost public health and reduce the environmental impacts of vehicles. However, news agencies' ideologies and reporting styles often influence public perception of cycling. For example, if news agencies overly report cycling accidents, it may make people perceive cyclists as "dangerous," reducing the number of opting to cycle. Additionally, a decline in cycling can result in less government funding for safe infrastructure. In this paper, we develop a novel prompting method to detect the perceived perception of cyclists within news headlines. To support this, we introduce a new dataset called "Bike Frames", which contains 31,480 news headlines and 1,500 human annotations. Our analysis focuses on 11,385 headlines from the United States. We also propose the BikeFrame Chain-of-Code (CoC) framework, which predicts cyclist perception, identifies accident-related headlines, and determines fault. This framework uses structured pseudocode to represent logical reasoning steps and incorporates news agency bias to enhance prediction accuracy, outperforming traditional chain-of-thought methods used in large language models. Most importantly, we find that incorporating news bias information significantly impacts performance, improving the average F1 score from .739 to .815. Finally, we conduct a comprehensive case study on U.S. news headlines, revealing differences in reporting between mainstream news agencies and cycling-specific websites, as well as variations in coverage based on the gender of cyclists. **WARNING: This paper contains descriptions of accidents and death.**

Introduction

Bicycling is a means of transportation and exercise that can reduce greenhouse gas emissions and improve public health. For instance, increasing pedestrian and bicycling trips, with a corresponding average decrease in automobile trip lengths by as little as one to three miles, can significantly affect emissions and fuel consumption (Gotschi and Mills 2008). Moreover, bicycling could reduce CO2 by six to fourteen million tons and reduce fuel consumption by 700 million to 1.6 billion gallons (Gotschi and Mills 2008). Bicycling can also provide health benefits, such as reducing cardiovascular risk factors, the likelihood of coronary heart disease, general morbidity, mortality risk, cancer risk, and obesity (Oja et al. 2011). Unfortunately, the public's perception of cycling—which can be influenced by how the news and social media portray cyclists—might impact both the number of cyclists and investment in cycling infrastructure by the community (Berke et al. 2019). Individual reactions to the news, illustrating a broader interplay of personal values and actions, often guide their decisions and behaviors (Chen et al. 2014). Hence, in this paper, we develop a method to analyze people's perception of cyclists and the language used in portraying cyclists in news headlines by introducing a new dataset "Bike Frames"[1] and applying novel methods trained on this dataset to conduct a case study. This case study delves into extracting explicit and implicit information from headlines about how bicyclists—and motorcyclists as a comparison group—are portrayed.

Macmillan et al. (2016) introduce a framework for understanding the media's impact on public perceptions of cycling. Their analysis showed that while cycling trips in London doubled from 1992 to 2012, media coverage of cyclist fatalities increased 13-fold, a trend not observed in motorcyclist-related media coverage during the same period. This disproportionate coverage might create complex feedback loops affecting cycling growth, with the impact likely varying between cities. Therefore, analyzing public perception through news sources, like local agencies, is essential for understanding and addressing how bicyclists are portrayed, which is significant for the broader transportation and urban science research community. For instance, Macmillan and Woodcock (2017) found that increasing cycling in cities is good for health and addressing climate change. Likewise, Aldred et al. (2019) found that media and public opposition were not reported as major issues for communities not considering new bicycling infrastructure, but they can substantially affect the release of funding for new cities that begin considering new infrastructure. However, Aasvik and Bjørnskau (2021) point out that prior work has mainly relied on surveys and qualitative analysis, making it a challenge to measure perceptions at scale, especially when comparing diverse communities.

Previous research indicates that gender can influence factors like safety perception and home responsibilities, affect-

[1]The dataset and code will be available on GitHub at https://github.com/Zephyr1022/BikeFrames.

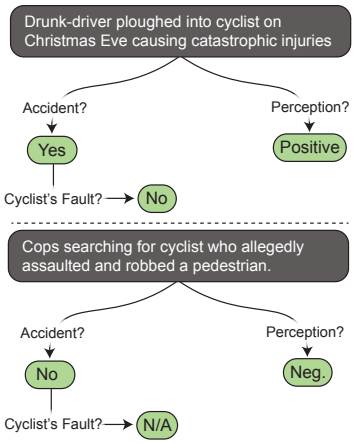

Figure 1: Understanding and explaining implicit views of cyclists requires reasoning. Hence, our "Bike Frames" aim to (1.) detect accident-related headlines, (2.) identify implicit suggestions of "who was at fault", and (3.) measure perception towards cyclists.

ing cycling behaviors. For example, female cyclists [2] may be more concerned about road safety and more likely to be influenced by news broadcasts than males, especially if they perceive a heightened risk from such coverage (Emond, Tang, and Handy 2009). Additionally, Harris and Jenkins (2006) discovered that women tend to be more risk-averse and feel more negative consequences of sharing the road with automobiles or other vehicles than men. These findings suggest that female cyclists may be more likely to favor off-road bike lanes separated from traffic (Garrard, Handy, and Dill 2012). Understanding the interaction of gender on perception can help planners and policymakers develop strategies to promote cycling, particularly among women.

Overall, to better understand the perception of cyclists in news headlines, we introduce a novel Bike Frames Corpus, methodology, and analysis approaches that can extract the perception portrayed by news headlines using machine learning algorithms. We summarize the task our dataset addresses in Figure 1. Specifically, we measure two aspects: the perceived view of the writer's feelings about the bicyclist and information about whether or not the bicyclist was in an accident. In the example "*Drunk-driver plowed into the cyclist*," we can understand that the headline is related to a driver-bicyclist accident and that the driver was probably at fault. We may perceive that the writer expresses sympathy towards the cyclist (i.e., sees the cyclist in a positive light) given their explicit remark about the driver's initial condition (intoxicated) and through the strong verb "*plowed*." In contrast, if the headline said "*Cyclist allegedly hit by a driver*," the text would still discuss an accident. However, we may perceive the writer as having a more positive perception of the driver, reducing the positive perception of the bicyclist. Therefore, we developed the dataset to measure the public perceptions of bicyclists and motorcyclists from news headlines for direct comparison. Specifically, the approach uncovers readers' sentiment patterns toward cycling-related news articles.

We explore using large language models (LLMs) to predict perception and fault in news headlines. Substantial research uses in-context examples and chain-of-thought reasoning to improve prediction (Wei et al. 2022; Chen et al. 2023). However, traditional methods like Chain-of-Thought (Wei et al. 2022) and Program-of-Thought (Chen et al. 2023) process logic in a single operation, which limits their expressiveness and efficiency. Inspired by Li et al. (2024) and Chae et al. (2024), we introduce the BikeFrame Chain-of-Code (CoC) framework to enhance the reasoning of LLMs by using structured pseudocode for precise logic management (Yang et al. 2025). Our approach breaks down complex problems into manageable intermediate reasoning steps. It also integrates news agency bias analysis to predict accident occurrences, identify fault, and assess cyclists' perceptions. We use code representation prompts to guide LLMs without actually executing code, and use intermediate `print()` statements to clarify decision rationales. Previous studies show that using structured pseudocode improves the coherence and interpretability of a model's reasoning, making it better at tackling ambiguous tasks like commonsense and social reasoning (Weir et al. 2024). The use of news agency bias is highly informative. Research has shown that news agencies often produce biased headlines (Weatherly et al. 2007). Therefore, we investigate whether prompting LLMs to consider these biases can improve predictive performance when classifying cyclist perception.

In summary, our contribution is three-fold. First, we introduce a new dataset for understanding the perception of cyclists in news headlines that may influence readers' perceptions of bicyclists. Second, we developed a novel prompting method called BikeFrame Chain-of-Code, which incorporates structured pseudocode logic with news agency bias reasoning tasks, substantially improving over traditional prompting methods. Third, we perform a comprehensive case study on important aspects of transportation—the portrayal of bicyclists in the media—which has not been the main focus of prior research. Our work expands on prior media framing biases (Levin, Schneider, and Gaeth 1998; Rho, Mark, and Mazmanian 2018). We investigate gender biases in reporting cycling accidents, comparing the portrayals of male versus female cyclists.

Related Work

In this section, we discuss four main areas of research related to our study: bicycle and travel infrastructure, gender differences in cycling, linguistic analysis, and semantic analysis to understand the news. We detail the methodologies used to develop accurate classifiers and explore the broader social implications related to cycling.

Bicycling, Transportation, and Infrastructure Research. There has been substantial research into bicycling infrastructure and the public's perception of bicyclists using various methods (e.g., surveys, crowd-sourcing, and content analysis). Macmillan et al. (2014) and Macmillan and Wood-

[2]Our discussion on gender, particularly the terms "male" and "female," is rooted in the appearance of pronouns in cycling news headlines, not self-disclosed gender.

cock (2017) used dynamics modeling to illustrate a feedback loop between political will, environmental views, and bicycling growth. Aldred et al. (2019) found that media and public opposition impact funding for new bicycling infrastructure in cities considering it. Ferster et al. (2021) showed cyclical changes in attitudes toward bicycling infrastructure, while Barajas (2021) noted the lack of bike infrastructure in Black and Latino neighborhoods, suggesting that analyzing local news perceptions can help advocate for improved infrastructure in these communities. Boufous, Aboss, and Montgomery (2016) highlighted that newspaper reports of cycling accidents often focus on dramatic, unusual incidents, contrasting with public health reporting and potentially misleading about cycling risks. Macmillan et al. (2016) found a significant increase in media coverage of cyclist fatalities over two decades, indicating a specific media bias towards cyclists, unlike the steady coverage of motorbike accidents.

Gender and Bicycling. Previous research (Emond, Tang, and Handy 2009) shows that female and male cyclists perceive safety and needs differently. This means that the impact of social media on their attitudes and behavior may also differ. Female cyclists are often more risk-averse and influenced by content addressing their specific concerns, such as the preference for off-road paths away from traffic (Garrard, Rose, and Lo 2008). In contrast, male cyclists may respond more to news that aligns with their attitudes and perceptions, including higher exposure to severe crashes. Singleton and Goddard (2016) found that gender-based differences in attitudes, preferences, and social norms, including financial or logistical barriers, affect women's likelihood to cycle, suggesting a need for policy reevaluation to support female cyclists. Prati et al. (2019) and Bouaoun, Haddak, and Amoros (2015) further report that women experience higher discomfort in mixed traffic and have a lower risk of crashes, whereas men, though facing higher injury death rates, have a slightly lower risk per trip. AitBihiOuali and Klingen (2022) showed that dedicated cycling infrastructure can increase female cycling participation by 4-6%. These findings highlight the importance of understanding the demographic determinants of cyclists' safety perceptions and the need for tailored approaches to bridge gender gaps in cycling (Emond, Tang, and Handy 2009; Garrard, Rose, and Lo 2008).

Linguistic Analysis. Drawing from established methods in prior studies (Chen et al. 2014; Rho, Mark, and Mazmanian 2018), we identified both content features like n-grams and stylistic features using the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al. 2015). This LIWC analysis measured the frequency of tokens related to cognition, perception, and various linguistic styles, including emotional and temporal language. Our goal was to understand how these different features in headlines might affect reader perceptions, in line with research showing that the linguistic style and sentiment of writing significantly influence audience reactions on news platforms and social media (Beasley and Mason 2015). Additionally, Media agents often employ different headlines with partisan biases for the same event, even when the article content is almost identical (Silverman et al. 2016; Rho, Mark, and Mazmanian 2018). Studies have also shown that people often share articles based on the headline alone, without fully reading the story (Silverman et al. 2016). This indicates a common behavior of circulating news based on headlines, which suggests that the way headlines are perceived can significantly influence individuals' decisions and opinions.

Semantic Analysis and Understanding the News. Semantic analysis in NLP has led to extensive research on text understanding, such as Frame Semantics in FrameNet (Baker, Fillmore, and Lowe 1998; Fillmore, Johnson, and Petruck 2003), which interprets different perspectives of the same event. While Card et al. (2015) developed a media frames corpus for news article annotation, it lacked insights into specific entity perceptions. Recent work by Sap et al. (2020) has delved into social bias using frame semantics, combining pragmatic inference with commonsense reasoning. Alongside this, new tools have emerged, such as Liu et al. (2021) multi-modal feature fusion method for image captioning, Spinde et al. (2021) dataset for detecting biased news, and Ang and Lim (2022) model for analyzing inter-organization relationships from news content, providing enhanced support for news writers. Most similar to our work is the study by Gabriel et al. (2022) that creates "perceived" news frames from headlines regarding how people think about misinformation, focusing on how readers interpret the intent behind the news. This approach parallels Tourni et al. (2021) development of the Gun Violence Frame Corpus (GVFC), where different news articles are analyzed for varied perspectives on the same topic. However, our work differs in two direct ways: first, it focuses specifically on the portrayal of bicycling in news headlines, analyzing implicit sentiments toward cyclists; and second, unlike previous studies that mainly examined direct textual categories (Tourni et al. 2021), our approach delves into implicit sentiment towards the cyclist mentioned in the headline from the writer's perspective, offering a deeper insight into media representation of cycling.

Data Collection and Annotation

Data Collection. We use Google News to collect cyclists' and accident-related news headlines from 2001 to 2021. Google News is a popular news aggregation service that has been available on Google for a long time. It aggregates the latest news across multiple sources and categorizes them based on trends, regions, and topics. We use Google News XML RSS Feed API [3]. The API limits the number of articles we can receive each month. We use the keywords "cycling", "cyclist", and "bike" to pull headlines. We scraped 32,980 for cyclist-related headlines. For annotation, we randomly sample 50% of the headlines with the keywords "crash", "death", "killed", and "dead" and 50% without these keywords. Annotators are uniformly provided a random sample generated from the entire dataset and the sub-sampled subset. We use the second set of keywords to guarantee that accident-related posts are annotated. However, it is important to note that accident-related posts can also appear without the sub-sampled keywords. To identify US-based headlines, we manually reviewed the URLs of news agencies

[3] https://news.google.com/news/rss.

from the data. This ensured that our selection was based on the news agency's location rather than the headline content.

Annotation Guidelines. We annotate two major categories for each news headline: Perception and Accident categories. Each category helps understand how people interpret the writer's perception of the cyclists in the headline, whether the article is about an accident, and the readers perceived opinion about who caused the accident.

Perception Categories. Perception is used to analyze and measure how readers perceive how readers interpret the writer's perception towards the cyclists in the headline. Specifically, we ask annotators to label perception into Negative, Positive, and Neutral categories.

***Negative*.** The negative class is used when the reader perceives the writer's attitude toward the cyclist as negative. Such negative perceptions can be expressed in different ways. For example, in the headline:

> ***Example****: Cyclist harasses motorists at Serangoon roundabout, smacks vehicles while hurling vulgarities—The Independent*

provides an example where a cyclist is claimed to have caused damage to vehicles. Likewise, negative perception can be caused because of explicit mentions of breaking traffic laws, for example, the headline

> ***Example:*** *Dashcam sees bus & cyclist ignore red light in Norwich - Norwich Evening News*

discusses a cyclist running a red light. Generally, annotators agreed that it is negative when a cyclist is portrayed as committing a traffic crime. Upon inspection of the article, someone submitted a video because of the danger that can be caused by a bus running a red light, which can be perceived as a much greater danger than a cyclist running a red light. However, the writer found it equally important to mention the cyclists because their appearance in the video was interesting and relevant to the writer's implicit viewpoint. Thus, we argue that the writer is implicitly reporting the cyclist negatively by equating their actions with the more severe actions of the bus driver.

***Positive*.** A positive perception reflects the reader's interpretation of the writer's headline as being positive about the cyclist. This can manifest itself in multiple ways. For example, the headline

> ***Example:*** *A Plover man donated a kidney to save a stranger. Now he's cycling cross-country to raise awareness. — Stevens Point Journal*

discusses how a cyclist is raising awareness for organ donation. On the other hand, there are many examples where an accident is discussed where the cyclist describes direct mention of severe injuries. For instance, the headline

> ***Example:*** *Speeding taxi drags cyclist to his death, leaves another injured - DispatchLIVE*

describes a cyclist who was killed in an accident. In such gruesome (i.e., descriptive) descriptions, annotators assumed that the writer expressed a negative sentiment towards the driver. Yet, the writer's perception of the cyclist is positive, i.e., the writer would not write such a descriptive narrative of the accident if they thought negatively about the cyclist. Hence, even though the writer reports a gruesome event, it is perceived to raise awareness about the driver's actions. Less descriptive headlines were labeled neutral. This is the case in previous work by Joye (2015) where establishing an emotional bond (e.g. "panic and chaos everywhere") can be a tactic to get the audience to care.

	Category	Cyclists
Related	Yes	623
	No	877
Fault	Cyclist	36
	Unknown	1163
	Other	301
Perception	Negative	94
	Neutral	236
	Positive	1170

Table 1: Dataset Statistics for the labeled data.

***Neutral*.** There are many headlines where it is unclear whether the writer views the cyclists with positive or negative perceptions. We label such headlines as Neutral. For example, the headline

> ***Example:*** *French cyclist Brunel signs two-year deal with UAE Team Emirates - Gulf Today*

provides a simple fact about an athlete. The text has little emotional expression, and it is impossible to determine whether the writer is a fan of the athlete. On the other hand, the headline

> ***Example:*** *SH 183 reopened in Cedar Park after cyclist-vehicle crash - KXAN.com*

reports a crash similar to the positive headline discussed above. However, the description of this headline is not directly about the accident—it is focused on the reopening of a highway. Hence, it is difficult to determine how the writer feels about the cyclist.

Accident-Related Categories. We annotate headlines into two major classes: Related to an Accident (Yes/No) and Accident Fault (Cyclists, Unknown, Driver).

***Related to Accident (Yes/No)*.** A headline is related to an accident if it is directly discussed in the headline. For instance, the headline

> ***Example:*** *Caught On Camera: Passenger Opens SUV Door, Seriously Injuring Cyclist in Lincoln Park*

directly mentions a crash involving cyclists (we use the same criteria for motorcyclists). Likewise, we annotate headlines that involve an "attack" of some kind (e.g., a fight, assault, etc.) as being related to an accident. For example, the headline

> ***Example:*** *Cyclist attacked by two unknown tracksuited men on Boothferry Road — Hull Live*

mentions cyclists actually getting attacked by two tracksuited men.

	Related	Fault	Perception
Cohen's Kappa	.90	.71	.63

Table 2: Cohen's Kappa between the two annotators for each category for all of our labeled data.

Geographic Region	Cyclists
Non US-Based	20,095
US-Based	11,385

Table 3: Dataset Statistics for the Non-labeled Data.

The "Not Related" category is used for headlines that do not mention anything related to an accident. For instance, the headline

> ***Example:*** *Gallery: A Paris-Roubaix for the ages — Cyclist*

discusses a race, not an accident or attack.

Fault. If a headline is related to an accident, we annotate it with who is perceived to be at fault by readers, i.e, who caused the accident. Specifically, annotators will label each headline as "Cyclists", "Unknown" or "Other". For instance, the "**Cyclists**" fault headline

> ***Example:*** *Blame entirely on the cyclist. Cyclist collided with truck*

directly puts the blame *entirely* on the cyclist; hence it is labeled as being the cyclist's fault. As an example of the "**Unknown**" class, the headline

> ***Example:*** *SH 183 reopened in Cedar Park after cyclist-vehicle crash - KXAN.com*

shows a fact and does not provide details about who caused the accident. Finally, for the "**Other**" class, in the headline

> ***Example:*** *Alludes to hit and run on the vehicle owner's part.*

we see an example where a driver was involved in a hit-and-run situation. In this case, someone "Other" (i.e., the driver) is at fault.

Agreement and Dataset Statistics. We followed a two-stage annotation procedure. First, two annotators completed the process independently of one another. One-on-one interviews were conducted with annotators to identify dataset issues versus annotator errors. The information gained from the discussions was then used to improve the guidelines. Next, the two annotators completed 1,500 headlines related to cyclists. The agreement results are shown in Table 2. The Cohen's Kappa agreement scores ranged from .62 to .90, representing "substantial agreement" and "almost perfect agreement," respectively (Landis and Koch 1977). To improve the accuracy and consistency of perception annotations, we introduced a third annotator to help adjudicate areas with lower agreement scores. This step significantly improved the final dataset, as evidenced by a Cohen's Kappa of 0.63 for Perception, indicating a substantial agreement (where 1 indicates perfect agreement). Second, to improve data annotation further, we again met with the annotators and reviewed and discussed disagreements among the annotated headlines to form the final gold standard dataset. The final dataset statistics can be found in Table 1.

US-Based Website Dataset Statistics. In our dataset, we have a total of 31,480 non-labeled news headlines and 1,500 labeled ones. These headlines are collected from a total of 5,827 unique website links worldwide. Among these, 2,168 are based in the US, and within the US-based websites, there are 125 domain-specific links and 2,043 general news website links. In our work, we manually identified all of these website links by reviewing each link one-by-one. Geographical factors play a significant role in cycling activities. Literature suggests that geographical contexts significantly influence cycling activities (Relia et al. 2018). For instance, factors such as infrastructure, climate, and traffic regulations in the US may differ from those in other regions (Chan and Wichman 2020), potentially affecting cycling prevalence and the nature of cycling-related incidents reported in the media. Hence, our analysis is narrowed down to US-based websites. Based on these websites, we refine our data by filtering news headlines based on pronouns such as "he" or "she" to categorize them into male and female-related headlines. The statistics from this categorization are shown in Table 3 and Table 5 in the appendix.

FAIR Requirements. Our dataset adheres to FAIR principles (Findable, Accessible, Interoperable, and Re-usable). The dataset *will be* Findable and Accessible through Github. Moreover, the data will be licensed under the Creative Commons Attribution License (CC BY 4.0). Finally, the data is shared as a text file format along with the complete annotation guidelines, which are shared as Word documents. Thus, the data is reusable and interoperable.

Methodology

We use large language models and prompting technique to classify cyclist perception. Previous studies on enhancing large language models' reasoning have mainly focused on two methods. One method involves generating rationales in natural language, such as Chain-of-Thought (Wei et al. 2022), or code, like Program-of-Thought (Chen et al. 2023). These approaches execute reasoning step-by-step in real-time without a separate planning phase. As a result, models must process and apply logic in a single operation, limiting their expressiveness. Moreover, these models cannot reuse previously understood logic when encountering similar problems, impacting their efficiency (Chae et al. 2024).

The second strategy to improve LLM reasoning involves generating detailed plans in natural language, which are then broken down into specific reasoning steps, as demonstrated in methods like Least-to-Most (Zhou et al. 2022) and Plan-and-Solve (Wang et al. 2023). However, prior research suggests that natural language may not be the most effective way to express task logic. Instead, code representations (i.e., pseudocode) can enhance reasoning and planning by providing structured patterns that support logical thinking (Hu et al. 2023). For example, Li et al. (2024) introduced Chain-

of-Code, which uses pseudocode to handle undefined behaviors more effectively. This helps models simulate expected outputs and better manage errors. Similarly, Puerto et al. (2024) found that code representations can trigger reasoning abilities in LLMs during the inference stage. Most traditional methods focus on single instances, which overlooks the potential to identify and use shared reasoning patterns across diverse tasks (Zhou et al. 2024; Yang et al. 2025). To address this, Chae et al. (2024) proposed THINK-and-EXECUTE, a method that identifies and applies these common reasoning patterns to improve reasoning performance.

Inspired by Li et al. (2024) and Chae et al. (2024), we use programming languages such as Python to formulate the logic required to solve specific tasks, including those of PoT and CoC. This allows us to take advantage of their structured syntax, execution feedback, and modular design. We choose pseudocode over natural language to better guide the reasoning process of LLMs. Our method breaks complex problems into a series of intermediate reasoning steps, helping streamline development by prompting the model step by step. Rather than executing the code, the LLM simulates its logic internally to produce the expected output. Our approach consists of two primary steps (as shown in Figure 2): (1) **DESIGN**, where we develop code to address each subtask associated with the Bike Frame tasks methodically, and (2) **EXECUTE**, where we prompt the model to reason over the code by invoking the "BikeFrame" class. Here, we input test headline titles and news sources as shown in Figure 2 within the LLMs block. We describe each subsection next.

DESIGN: Pseudocode. We developed a Chain-of-Code-like prompt named the "BikeFrame," that jointly predicts three tasks: determining if an accident happened, identifying who is at fault, and assessing perception toward cyclists. The "BikeFrame" class in the prompt includes two main functions for accident detection and perception analysis, shown in Figure 2. First, the accident analysis function checks the headline for mentioning accidents (e.g., collisions or injuries). The class description, highlighted in red in Figure 2, guides the LLMs in performing this analysis. If an accident is detected, the function conducts a behavior analysis to identify actions or failures contributing to the incident, extracting behaviors of all parties involved, and assessing if any traffic laws were violated. The class then evaluates the tone of the headline to determine the perception towards each party. The results from the accident analysis are passed to the cyclist perception function via a return statement. In the perception function, these analyses are combined to identify the most likely party at fault and predict the overall perception towards cyclists. This pseudocode structure makes our prompt flexible, allowing it to incorporate different reasoning components.

Additionally, A novel feature of our BikeFrame class is the ***integration of an analysis of the news agency's reporting style*** by evaluating the tone, language, and themes in their coverage of cyclist-related stories. This analysis helps discern any biases or recurring patterns in the agency's reporting. By combining these insights, the LLMs identify the most likely party at fault and predict the overall perception towards cyclists. This method ensures a comprehensive understanding of how the news agency's style influences public perception, supported by studies that emphasize the importance of media framing in shaping public opinion (Sharofqizi and Ruzimurodovna 2024). We simply add a prompt in the pseudocode to extract the opinion of the news source from LLMs based on their pretrained knowledge. To ensure the LLMs use this news source analysis in subsequent decisions, we use intermediate `print()` statements that explicitly reference the analysis. These statements clarify the rationale behind each decision, illustrating how the model arrives at its conclusions. The sections where we integrate the news source analysis are in yellow in Figure 2.

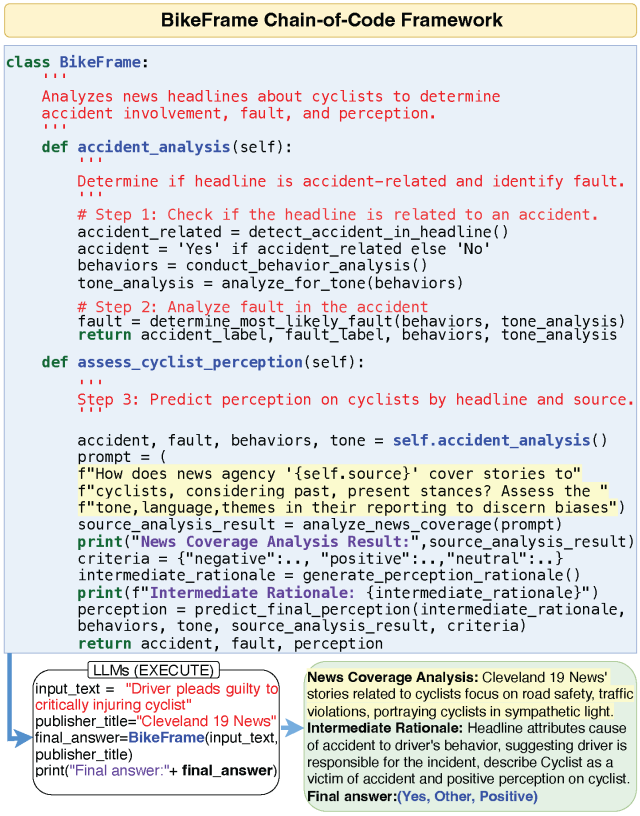

Figure 2: BikeFrame Chain-of-Code prompt. The news information component is highlighted in yellow.

EXECUTE: Simulated Pseudocode Execution. The emulated execution process is visually demonstrated in a green box in Figure 2. Our code prompts are not actually executed; instead, they are read by the LLM and used to generate a natural language response. The green box shows the printed output for news coverage analysis, intermediate reasoning steps, and the final answer for joint classification tasks. This part of the prompt is appended to the design prompt and represents an instantiation of the class before being passed to the LLM. Using the `print()` statement, we guide the LLM to follow the predefined rules of our Bike Frame framework. This structured approach improves the model's consistency with entity tracking and logical constraints, significantly reducing errors and hallucinations, and

	Perception Categories			Accident Categories				
	Negative	**Neutral**	**Positive**	**Not Accident**	**Cyclist**	**Unknown**	**Other**	**Avg**
Uniform	.094	.176	.470	.513	.018	.415	.241	.275
Stratified	.051	.093	.793	.561	.000	.761	.286	.364
Logistic Regression	.338	.488	.829	.911	.154	.930	.811	.637
Logistic Regression + LIWC	.174	.514	.899	.908	.182	.949	.847	.639
RoBERTa	.444	**.714**	.935	.962	.105	.942	.853	.708
MT RoBERTa	.710	.696	.940	.960	.154	.957	.880	.757
MTLPT RoBERTa	.519	.677	.933	**.970**	.133	.959	.898	.727
MT + MTLPT RoBERTa	.583	.646	.936	**.970**	.200	.959	.893	.741
Zero-Shot GPT3.5	.238	.372	.510	.868	.364	.853	.671	.554
Zero-Shot GPT3.5 + CoT	.320	.309	.534	.841	.500	.819	.635	.565
Few-Shot GPT3.5	.253	.344	.433	.918	.421	.898	.809	.582
Few-Shot GPT3.5 + CoT	.282	.438	.723	.898	.471	.897	.655	.624
Our Methods								
Zero-Shot GPT3.5 + CoC + News	.119	.395	.406	.787	.094	.669	.479	.421
Zero-Shot GPT3.5 + CoC + News + Self-Consistency	.151	.516	.394	.778	.094	.652	.480	.438
Few-Shot GPT3.5 + CoC + News + Self-Consistency	.759	.684	**.947**	.941	**.667**	**.956**	**.903**	**.837**
Ablation								
Few-Shot GPT3.5 + CoC + News	**.813**	.667	.938	.931	.545	.943	.872	.815
Few-Shot GPT3.5 + CoC + Self-Consistency	.583	.417	.776	.934	.600	**.956**	**.904**	.739
Few-Shot GPT3.5 + CoC	.296	.402	.759	.927	.545	.947	.886	.680

Table 4: F1 scores for all classes. The best scores are **bolded** for each column.

improving the reliability and accuracy of its outputs. Prior work has shown that displaying intermediate steps is essential for logically reaching the final answer through chain-of-thought reasoning (Wei et al. 2022; Chae et al. 2024). Tracking these intermediate steps also helps the model maintain an accurate understanding of variable changes throughout the simulated code execution. Following the prompt design of Chae et al. (2024), the final output for joint tasks is shown with `print(Final answer: {final_answer})` as the last output of the system.

Results and Discussion

This section describes our experimental setup, performance results, and a use study using our models on all US-based news headlines, both labeled and unlabeled. Specifically, we analyze reporting differences across cyclist articles, general and domain-specific websites (like bicycling.com), and when gender-related names are mentioned. Additionally, we study linguistic factors in headlines, including style and content, to understand their impact on perceptions of cyclists, both positive/negative and being at-fault in accidents.

Experimental Setup. We employ GPT-3.5-Turbo (OpenAI 2023) for its strong reasoning and code generation capabilities due to its balance of performance and cost-effectiveness. We set the temperature parameter T=0.0 to enable greedy decoding, ensuring outputs are precise and deterministic. To further enhance the model's performance and robustness, we adopt the Self-Consistency approach (Wang et al. 2022), generating multiple outputs using the chain-of-code framework and aggregating them to derive the final answer. Following the setting in Wang et al. (2022), we use a top-k sampling set at 0.9 with 20 sampled reasoning paths to refine our decision-making process, maintaining a reasonable cost-to-performance ratio. We trained multiple baseline models on our dataset, splitting it into train, dev, and test sets with a ratio of 7:1:2, using four NVidia GeForce GTX 1080 Ti GPUs. Each model was trained on the training set and tested to calculate the F1 score for each class.

Baselines We evaluate eight baseline models: Logistic Regression with Tfidf ngram features (LR), Logistic Regression with Tfidf ngram and LIWC features, three RoBERTa variant models with multiclass (multiclass, multi-task, multi-task MTLPT (see appendix for details), and two random baselines. We also evaluate additional prompt-based baselines, including (1) Direct prompting, where the model predicts answers without generating rationales (2) Zero-shot CoT (Kojima et al. 2022), where add "Let think step-by-step" to prompt LLM solving problems through sequential reasoning. (3) Few-Shot CoT (Brown et al. 2020), which uses a few examples to enhance its reasoning.

Model Performance. In Table 4 we report the F1 score for each of the Accident-related classes. Our results show that the RoBERTa model significantly outperforms all Linear and Random baselines, as well as our BikeFrame Chain-of-Code model (.970 vs .941) in detecting accident-related content. This is likely because detecting accident patterns is a simple task, making it easier for RoBERTa to identify the relevant features. However, for the "who is at fault" and perception tasks, our BikeFrame method with news source analysis and self-consistency outperforms all RoBERTa variant models in every class except for the "Neutral" classes. When comparing Few-Shot GPT-3.5 + CoC + News + Self-Consistency to the Zero-Shot method, we find that it significantly improves performance across all classes. The largest improvements are seen in the "Negative" class, which involves complex reasoning chains. To accurately decide on perception, it is often necessary to consider comprehensive

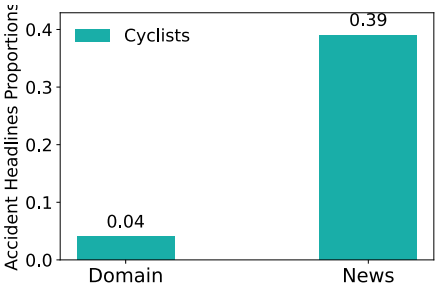

Figure 3: The proportion of accident-related headlines in the domain-specific (e.g., bicycling.com) and general news websites (e.g., nypost.com and chicagotribune.com).

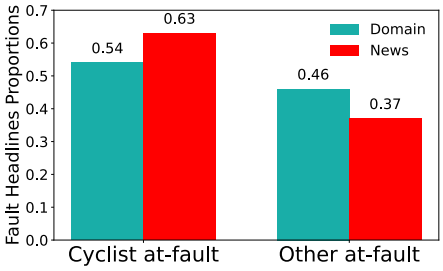

Figure 4: The proportion of headlines in domain-specific websites vs. general news-based websites on who is perceived to be at-fault for the accident, the cyclists, or other entities in the headline (Other).

situations such as whether the cyclist violated traffic laws and how the cyclist is described using sarcasm or passive voice. For the "Cyclist at fault" class prediction, our Few-Shot GPT-3.5 + CoC + News + Self-Consistency model outperforms the best RoBERTa variant, "MT + MTLPT RoBERTa." This class is a low-resource category with only 94 cases, much fewer than other categories. Our model effectively learns patterns from few examples, significantly enhancing performance, which allows the model to better generalize from the small number of available instances.

When comparing our CoC method with regular CoT and direct prompting, we find that for zero-shot learning, both CoT and direct prompting outperform our method. This may be because our prompts are more complex and, without examples, can make it difficult for language models to understand the underlying reasoning chain. However, we also observed that the output of our zero-shot learning approach adheres more closely to the desired output format. Our method results in significantly fewer parsing errors, whereas general CoT and direct prompting often produce outputs that do not fit our desired format or miss some predictions. Despite being zero-shot, our method successfully addresses these issues. Therefore, adding a few examples significantly improves the model's reasoning ability. Overall, our Few-Shot GPT-3.5 + CoC + News + Self-Consistency model outperforms all other models based on average F1 scores.

Ablation Studies We conduct an ablation study on each component of our CoC BikeFrame prompt. To do this, we prepare three types of pseudocode prompts: one without the news source analysis component, one without the self-consistency component, and one without both components. Table 4 shows that the model without the self-consistency component slightly underperforms compared to our best model based on the average F1 score. The most significant performance drop occurs when the news source analysis component is removed, resulting in almost a 10% reduction in the average F1 score. This drop is particularly noticeable in perception prediction, where the model without the news component performs significantly worse across all three classes (for negative .759 vs .583; for neutral .684 vs .417; and for positive .947 vs .776). These results indicate that incorporating comprehensive news source analysis significantly enhances our model's performance, especially in accurately predicting perceptions. This improvement is likely because the news source analysis helps the model understand how different news agents frame cyclists. By examining the language and context used in news articles, the model can better identify biases, sentiments, and the overall narrative presented about cyclists. This deeper understanding enables the model to make more accurate predictions regarding perceptions, as it can account for the subtle nuances and framing techniques used by various news sources.

RQ1: Are there differences in how domain-specific websites and general news websites attribute fault for accidents, specifically to cyclists? We applied our best model to analyze 31,480 US-based news headlines, selected from a larger dataset of 81,746 global headlines. Our study counted accident-related headlines in cycling categories and classified 2,168 US websites as either domain-specific (like bicycling.com) or general news websites (such as nytimes.com). Out of these, 125 were domain-specific and 2,043 were general news sites. We then calculated the proportion of accident-related headlines for both types of sites, as shown in Figure 3. Results showed that general news sites are more likely to report accidents involving cyclists than domain-specific ones, with significant differences (p-value $< .00001$ using Z-test). This pattern suggests that media consumption could influence perceptions of cyclist safety, aligning with studies on how safety perception affects driving behavior (Intravia et al. 2017). These findings raise questions about the impact of media reporting on cycling behavior and suggest that changes in reporting could influence future cycling growth (Alvisyahri, Anggraini, and Sugiarto 2020).

In Figure 4, we report the proportion of headlines in the domain-specific and general news categories that state that whether cyclists or other entities are at fault for an accident. This figure answers the question, which websites are more likely to report negative stories about cyclists? General news websites make the highest proportion of at-fault stories. Specifically, 63% of the headlines in general news websites attribute fault to cyclists, compared to 54% in domain-specific websites. Conversely, 46% of headlines in domain-specific websites attribute fault to other entities, compared to 37% in general news websites. These findings suggest that general news websites are more likely to frame cyclists negatively in accident-related stories.

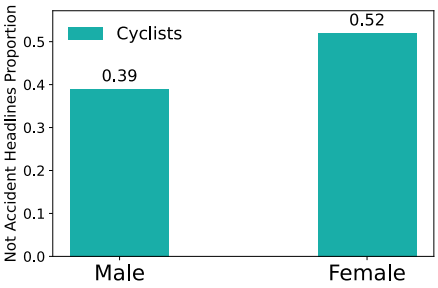

Figure 5: The proportion of male/female headlines unrelated to an accident for cyclists.

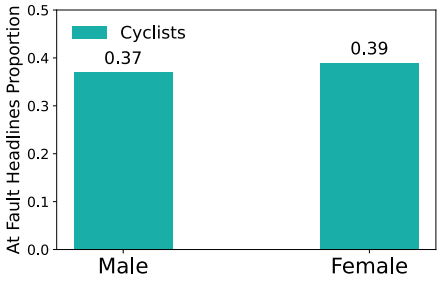

Figure 6: The proportion of at-fault headlines for cyclists in the Male and Female groups.

RQ2: Do news agencies report cycling accidents for males and females in proportion to their respective accident rates and presence in the cycling population? Our study investigates gender disparities in cycling-related news coverage. As mentioned, female cyclists are often influenced by news addressing specific needs, while male cyclists are influenced by news that aligns with their attitudes and perceptions (Emond, Tang, and Handy 2009). In addition, fatalities among women cyclists are often more prominently covered in the news headlines, possibly due to the perception of women as more vulnerable road users, making female-related accidents more newsworthy (Mindell 2012).

Despite male cyclists have a significantly higher risk of accidents, with death rates six times and injury rates five times higher than females (for Disease Control, Prevention et al. 2003), our analysis using gendered pronouns in news headlines reveals uneven media coverage. If accidents were reported equally in the media, we would expect significantly more headlines about male-related accidents. Yet, Figure 5 shows the proportion of non-accident-related headlines is lower for males than females in both cycling (.39 vs. .52), which is statistically significant (p-value $<$ .0001). Interestingly, the actual number of male-related accident reports are only 2.04 times higher than those for females, not aligning with the actual higher risk. If media coverage reflected actual accident rates, we would expect significantly more headlines about male-related accidents. However, our findings suggest that female cyclists receive disproportionately more coverage, indicating potential media bias. This bias highlights the need for policies and interventions to support and encourage women in cycling, considering their different safety perceptions, accident rates, and infrastructure preferences (Prati et al. 2019; AitBihiOuali and Klingen 2022; Singleton and Goddard 2016; Bouaoun, Haddak, and Amoros 2015). This study contributes to a broader understanding of how gender stereotypes in media may not align with actual demographic patterns, emphasizing the importance of equitable representation in urban transportation.

Figure 6 shows that male cyclists are perceived as less at-fault in accidents than female cyclists (.37 vs .39), potentially due to stereotypes about female cyclists being novel riders. This is consistent with research showing that female drivers tend to be more cautious (Al-Balbissi 2003), while male riders are often seen as more skilled but prone to risky behaviors, such as driving under the influence of alcohol. These findings emphasize how gender perceptions affect fault attribution in cycling accidents, underlining the importance of further research into these dynamics and their impact on road behaviors and safety.

RQ3: What linguistic factors in news headlines predict the perception towards cyclists? Psycholinguistic research highlights that language use in social media significantly affects emotions, behaviors, and perceptions (Chen et al. 2014; Ernala et al. 2017; Park et al. 2011). The complexity of readers' reactions to news headlines involves cognitive aspects like inferring intent, emotional responses such as feelings of distrust, and behavioral actions like sharing news. These reactions are influenced by both the content and linguistic style of the headlines (Gabriel et al. 2022). We used a machine learning classifier trained on 3500 manually annotated news headlines to gauge public perceptions of cyclists, incorporating the top 500 n-grams and 93 LIWC features. Our analysis identified the 20 most influential n-grams and LIWC features that shape perceptions of fault and general sentiment towards cyclists and motorcyclists. Specifically, our study delves into those that portray cyclists as at-fault with negative perceptions, as shown in Figure 7 and Figure 8.

Linguistic Style for Cyclists At-Fault. In Figure 7, we can see that headlines shape public perceptions of at-fault cyclists using language that elicits strong emotions and cognitive evaluations, with features like like "prep," "compare," "leisure," "relate," and "affective." These elements suggest both an emotional response and a critical assessment of cyclists' actions, which can bias reader perceptions and contribute to negative sentiments on social media (Liao and Fu 2014). For instance, the use of words like "prep" and "compare" in at-fault headlines can emphasize the cyclists' perceived mistakes or irresponsibility. Such linguistic choices not only amplify the severity of incidents but also frame the cyclist in a vulnerable position, potentially magnifying their fault or recklessness. This highlights the impact of emotive language on public engagement, with headlines potentially altering public reactions and creating echo chambers. These findings emphasize the crucial role of media language in framing cyclists' actions and influencing societal views on cycling safety (Tan, Friggeri, and Adamic 2016).

Linguistic Content for Cyclists At-Fault. In the analysis of linguistic content for at-fault cyclists, n-grams like

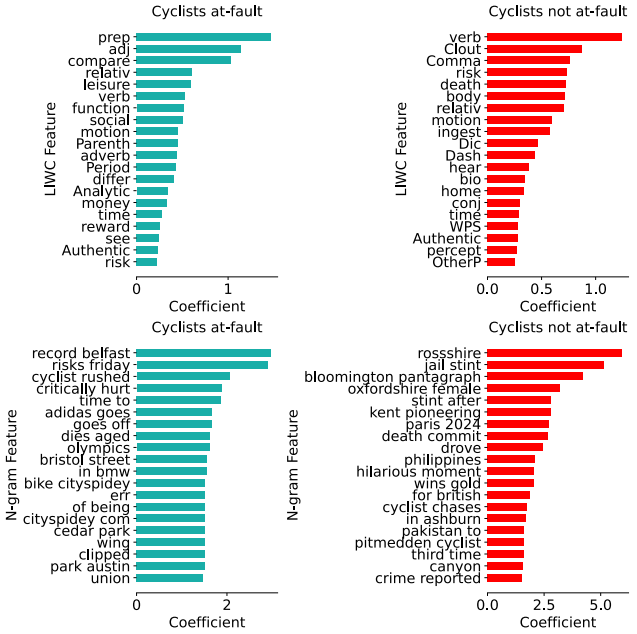

Figure 7: Top 20 Linguistic Factors (LIWC and n-grams (n=1,2)) in news headlines for cyclists at-fault or not.

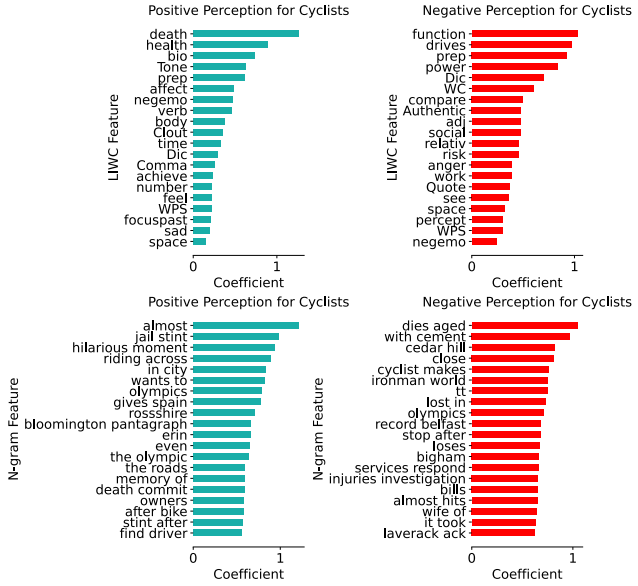

Figure 8: Top 20 Linguistic Factors (LIWC and n-grams (n=1,2)) in news headlines for positive or negative perception towards cyclists.

"record belfast," "risks friday," and "cyclist rushed" indicate cyclists' involvement in accidents, potentially as primary causes. Context-specific n-grams such as "adidas goes" and "olympic medal" relate to competitive events like the Olympics, framing cyclists in a sports context. N-grams like "bike cyclistby," "clipperd," and "park austin" reveal demographic and geographic aspects, indicating that news on cycling accidents often focuses on specific age groups or locations, potentially leading to stereotyping of these groups.

Linguistic Style for Negative Perception of Cyclists. As shown in Figure 8, headlines about cycling incidents often use LIWC features such as "function," "drives," "prep," "power," and "risk." This language emphasizes potential dangers and portrays cycling as hazardous. Additionally, terms related to "compare" and "anger" intensify this negative framing. For instance, words associated with negative emotions and critical assessments can lead readers to perceive cycling as risky and cyclists as reckless.

Linguistic Content for for Negative Perception of Cyclists. Additionally, the use of specific context-related n-grams, such as "dies aged," "cyclist makes," and "ironman world," highlights the severity and dramatic nature of these incidents, which can further skew public perception negatively. These linguistic features collectively convey negativity and biases, significantly impacting public perception and discourse around cycling incidents.

Limitations and Future Work

Our model, primarily analyzing news headlines, faces limitations like not providing full story context and potentially featuring "click-bait" headlines, especially on social media. Extending the analysis to short articles or tweets could offer a more comprehensive view of public perceptions. Its English-centric and US-annotator-based approach may not accurately capture global perceptions, necessitating careful application in diverse contexts. Our research acknowledges the risk of news headline framing, particularly on sensitive topics, being misused to negatively influence public perception, highlighting the need for responsible use of our analysis tools to prevent bias and stereotypes. Future studies will expand beyond our current Google News API dataset to more comprehensively analyze temporal changes, regional differences, and political influences in cyclist news framing, aiming for a broader community impact.

Conclusion

In this paper, we make several key contributions. First, we introduce a new dataset for urban informatics designed to analyze how news headlines portray cyclists. Second, we present BikeFrame, a novel structured prompting approach based on a Chain-of-Code architecture that jointly predicts multitask outcomes and incorporates a news source analysis module, significantly improves model performance. Third, we conducted an in-depth analysis of how cyclists are represented in the news. Key findings include general news websites reporting more accidents than cycling-specific sites, despite the higher fatality rates associated with cycling. Additionally, our study on gender-specific pronouns in news also revealed only a slight difference in accident reporting between male and female mentions (20% relative difference), contrasting with the actual higher accident rates for men (i.e., $>$700% relative difference), which can exaggerate the dangers of cycling for women and potentially limit the number of female cyclists.

References

Aasvik, O.; and Bjørnskau, T. 2021. Cyclists' Perception of Maintenance and Operation of Cycling Infrastructure Results From a Norwegian Survey. *Frontiers in psychology*.

AitBihiOuali, L.; and Klingen, J. 2022. Inclusive roads in NYC: Gender differences in responses to cycling infrastructure. *Cities*, 103719.

Al-Balbissi, A. H. 2003. Role of gender in road accidents. *Traffic injury prevention*, 4(1): 64–73.

Aldred, R.; Watson, T.; Lovelace, R.; et al. 2019. Barriers to investing in cycling: Stakeholder views from England. *Transportation research part A*.

Alvisyahri, A.; Anggraini, R.; and Sugiarto, S. 2020. Motorcyclist perceptions on road safety considering awareness, riding behavior, and risk-taking behavior, as latent variables. IOP Publishing.

Ang, G.; and Lim, E.-P. 2022. Guided Attention Multimodal Multitask Financial Forecasting with Inter-Company Relationships and Global and Local News. In *ACL*.

Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The berkeley framenet project. In *ACL*.

Barajas, J. M. 2021. Biking where Black: Connecting transportation planning and infrastructure to disproportionate policing. *Transportation research part D*.

Beasley, A.; and Mason, W. 2015. Emotional states vs. emotional words in social media. In *Proceedings of the ACM web science conference*, 1–10.

Berke, A.; Sanchez Lengeling, T.; Nawyn, J.; et al. 2019. Bike swarm. In *CSCW*, 1–4.

Bouaoun, L.; Haddak, M. M.; and Amoros, E. 2015. Road crash fatality rates in France: a comparison of road user types, taking account of travel practices. *Accident Analysis & Prevention*, 75: 217–225.

Boufous, S.; Aboss, A.; and Montgomery, V. 2016. Reporting on cyclist crashes in Australian newspapers. *Australian and New Zealand journal of public health*, 40(5): 490–492.

Brown, T.; Mann, B.; Ryder, N.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33.

Card, D.; Boydstun, A.; Gross, J. H.; et al. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of ACL-IJCNLP*.

Chae, H.; Kim, Y.; Kim, S.; Ong, K.; Kwak, B.-w.; Kim, M.; Mac Kim, S.; Kwon, T.; Chung, J.; Yu, Y.; et al. 2024. Language Models as Compilers: Simulating Pseudocode Execution Improves Algorithmic Reasoning in Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22471–22502.

Chan, N. W.; and Wichman, C. J. 2020. Climate change and recreation: evidence from North American cycling. *Environmental and Resource Economics*, 76: 119–151.

Chen, J.; Hsieh, G.; Mahmud, J. U.; et al. 2014. Understanding individuals' personal values from social media word use. In *CSCW*.

Chen, W.; Ma, X.; Wang, X.; et al. 2023. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research*.

Emond, C. R.; Tang, W.; and Handy, S. L. 2009. Explaining gender difference in bicycling behavior. *Transportation Research Record*, 2125(1): 16–25.

Ernala, S. K.; Rizvi, A. F.; Birnbaum, M. L.; et al. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *PACM HCI*, 1: 1–27.

Ferster, C.; Laberee, K.; Nelson, T.; et al. 2021. From advocacy to acceptance: Social media discussions of protected bike lane installations. *Urban Studies*, 58(5): 941–958.

Fillmore, C.; Johnson, C.; and Petruck, M. R. 2003. Background to framenet. *International journal of lexicography*.

for Disease Control, C.; Prevention; et al. 2003. Web-based Injury Statistics Query and Reporting System. *National Center for Injury Prevention and Control, Centers for Disease Control and Prevention (producer)*.

Gabriel, S.; Hallinan, S.; Sap, M.; et al. 2022. Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines. In *ACL*.

Garrard, J.; Handy, S.; and Dill, J. 2012. Women and cycling. *City cycling*, 2012: 211–234.

Garrard, J.; Rose, G.; and Lo, S. K. 2008. Promoting transportation cycling for women: the role of bicycle infrastructure. *Preventive medicine*, 46(1): 55–59.

Gotschi, T.; and Mills, K. 2008. Active transportation for America: The case for increased federal investment in bicycling and walking.

Harris, C. R.; and Jenkins, M. 2006. Gender differences in risk assessment: why do women take fewer risks than men?

Hu, Y.; Yang, H.; Lin, Z.; et al. 2023. Code prompting: a neural symbolic method for complex reasoning in large language models. *arXiv preprint arXiv:2305.18507*.

Intravia, J.; Wolff, K. T.; Paez, R.; et al. 2017. Investigating the relationship between social media consumption and fear of crime: A partial analysis of mostly young adults. *Computers in Human Behavior*, 77: 158–168.

Joye, S. 2015. Domesticating distant suffering: How can news media discursively invite the audience to care? *International Communication Gazette*, 77(7): 682–694.

Kochkina, E.; Liakata, M.; and Zubiaga, A. 2018. All-in-one: Multi-task Learning for Rumour Verification. In *COLING*.

Kojima, T.; Gu, S. S.; Reid, M.; et al. 2022. Large language models are zero-shot reasoners. *NeurIPS*, 35: 22199–22213.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*.

Levin, I. P.; Schneider, S. L.; and Gaeth, G. J. 1998. All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes*, 76(2): 149–188.

Li, C.; Liang, J.; Zeng, A.; et al. 2024. Chain of code: reasoning with a language model-augmented code emulator. In *ICML*, 28259–28277.

Liao, Q. V.; and Fu, W.-T. 2014. Can you hear me now? Mitigating the echo chamber effect by source position indicators. In *CSCW*.

Liu, F.; Wang, Y.; Wang, T.; et al. 2021. Visual News: Benchmark and Challenges in News Image Captioning. In *EMNLP*.

Liu, Y.; Ott, M.; Goyal, N.; et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Macmillan, A.; Connor, J.; Witten, K.; et al. 2014. The societal costs and benefits of commuter bicycling: simulating the effects of specific policies using system dynamics modeling. *Environmental health perspectives*.

Macmillan, A.; Roberts, A.; Woodcock, J.; et al. 2016. Trends in local newspaper reporting of London cyclist fatalities 1992-2012: the role of the media in shaping the systems dynamics of cycling. *Accident Analysis & Prevention*, 86: 137–145.

Macmillan, A.; and Woodcock, J. 2017. Understanding bicycling in cities using system dynamics modelling. *Journal of transport & health*, 7: 269–279.

Mindell, L. D. . W. M., J. S. 2012. Exposure-based,'like-for-like'assessment of road safety by travel mode using routine health data. *PloS one*, 7(12): e50606.

Oja, P.; Titze, S.; Bauman, A.; et al. 2011. Health benefits of cycling: a systematic review. *Scandinavian journal of medicine & science in sports*, 21(4): 496–509.

OpenAI. 2023. ChatGPT. https://openai.com/blog/chatgpt. Accessed: April, 10, 2025.

Park, S.; Ko, M.; Kim, J.; et al. 2011. The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In *CSCW*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; et al. 2015. The development and psychometric properties of LIWC2015. Technical report.

Prati, G.; Fraboni, F.; De Angelis, M.; et al. 2019. Gender differences in cyclists' crashes: an analysis of routinely recorded crash data. *International journal of injury control and safety promotion*, 26(4): 391–398.

Puerto, H.; Tutek, M.; Aditya, S.; et al. 2024. Code Prompting Elicits Conditional Reasoning Abilities in Text+ Code LLMs. In *EMNLP*, 11234–11258.

Read, J. 2008. A pruned problem transformation method for multi-label classification. In *NZCSRS 2008*.

Relia, K.; Akbari, M.; Duncan, D.; et al. 2018. Socio-spatial self-organizing maps: using social media to assess relevant geographies for exposure to social processes. *PACM HCI*.

Rho, E. H. R.; Mark, G.; and Mazmanian, M. 2018. Fostering civil discourse online: Linguistic behavior in comments of# metoo articles across political perspectives. *PACM HCI*.

Sap, M.; Gabriel, S.; Qin, L.; et al. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *ACL*.

Schütze, H.; Manning, C. D.; and Raghavan, P. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Sharof-qizi, O. D.; and Ruzimurodovna, E. U. 2024. EXPLORING THE INFLUENCE OF MEDIA FRAMING ON PUBLIC PERCEPTION. *Journal of Innovation, Creativity and Art*, 3(4): 19–23.

Silverman, C.; Strapagiel, L.; Shaban, H.; et al. 2016. Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate. *Buzzfeed News*, 20(1).

Singleton, P. A.; and Goddard, T. 2016. Cycling by choice or necessity?: Exploring the gender gap in bicycling in Oregon. *Transportation research record*, 2598(1): 110–118.

Spinde, T.; Plank, M.; Krieger, J.-D.; et al. 2021. Neural Media Bias Detection Using Distant Supervision With BABE-Bias Annotations By Experts. In *Findings of EMNLP*.

Tan, C.; Friggeri, A.; and Adamic, L. 2016. Lost in propagation? Unfolding news cycles from the source. In *ICWSM*.

Tourni, I.; Guo, L.; Daryanto, T. H.; et al. 2021. Detecting Frames in News Headlines and Lead Images in US Gun Violence Coverage. In *Findings of EMNLP*.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. 2009. Mining multi-label data. *Data mining and knowledge discovery handbook*, 667–685.

Wang, L.; Xu, W.; Lan, Y.; et al. 2023. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In *ACL*, 2609–2634.

Wang, X.; Wei, J.; Schuurmans, D.; et al. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Weatherly, J. N.; Petros, T. V.; Christopherson, K. M.; et al. 2007. Perceptions of political bias in the headlines of two major news organizations. *Harvard International Journal of Press/Politics*, 12(2): 91–104.

Wei, J.; Wang, X.; Schuurmans, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837.

Weir, N.; Khalifa, M.; Qiu, L.; et al. 2024. Learning to Reason via Program Generation, Emulation, and Search. In *NeurIPS*.

Wolf, T.; Debut, L.; Sanh, V.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yang, D.; Liu, T.; Zhang, D.; et al. 2025. Code to Think, Think to Code: A Survey on Code-Enhanced Reasoning and Reasoning-Driven Code Intelligence in LLMs. *arXiv preprint arXiv:2502.19411*.

Zhou, H.; Nova, A.; Larochelle, H.; et al. 2022. Teaching algorithmic reasoning via in-context learning. *arXiv preprint arXiv:2211.09066*.

Zhou, P.; Pujara, J.; Ren, X.; Chen, X.; Cheng, H.-T.; Le, Q. V.; Chi, E.; Zhou, D.; Mishra, S.; and Zheng, H. S. 2024. Self-discover: Large language models self-compose reasoning structures. *Advances in Neural Information Processing Systems*, 37: 126032–126058.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. However, there could be tools developed that generate headlines that are anti-cycling on purpose, particularly for groups of people using the dataset we developed.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. We develop a novel multi-task learning method to take advantage of class relationships.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. But, not all population information is available. But, we do provide what we can regarding whether the content is US-based.
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see the "LIMITATIONS AND FUTURE WORK" section.
   (g) Did you discuss any potential misuse of your work? Yes, see the "LIMITATIONS AND FUTURE WORK" section.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes for what is related to this project, see the data description and annotation guideline sections.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA. We have research questions to support our modeling and dataset results. But, these are not based on social theories in traditional hypothesis research.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. We included the main model and data in the supplementary material. The rest of the code will be released via GitHub upon acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. See the METHODOLOGY section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes. See the METHODOLOGY section.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes (pre-trained models)
   (b) Did you mention the license of the assets? Yes, it is released under the Creative Commons Attribution License (CC BY 4.0).
   (c) Did you include any new assets in the supplemental material or as a URL? Yes. The data is in the supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR ? Yes. See the DATA COLLECTION AND ANNOTATION section.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? No.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No, because our study was conducted internally using research assistants and was deemed not to require IRB approval.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? All graduate students are paid as research assistants which includes yearly stipend and tuition.

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

A.1. Baseline Model Descriptions

Likewise, we experiment with two novel multi-task learning methods to incorporate co-occurrence information into the model: Multi-Task RoBERTa (MTR) and Multi-Task RoBERTa using the Labeled Powerset Transformation (MTRLP).

Random Baselines. We use two random baselines from the scikit-learn package (Pedregosa et al. 2011): Uniform and Stratified. The Uniform baseline makes predictions for each class with equal proportions. The stratified random baseline makes predictions based on the class proportions in the training dataset.

Linear SVM and Logistic Regression. We trained a Linear SVM and LR model using frequency-inverse document frequency-weighting (TF-IDF) (Schütze, Manning, and Raghavan 2008) of unigrams and bigrams, a technique that assigns importance to words in a text corpus. For Linear SVM, we optimized the model by experimenting with different "C" values and L2 regularization. For LR, we used L1 regularization, 'liblinear' solver, and 'balanced' class weight. Both models were implemented using scikit-learn library (Pedregosa et al. 2011).

RoBERTa. We fine-tuned the RoBERTa (Liu et al. 2019) model from Huggingface (Wolf et al. 2019), averaging the second-to-last layer's token embeddings and passing them to a softmax layer for up to 25 epochs. The best model was selected using validation data. We employed cross-entropy loss, a mini-batch size of 8, a learning rate of 2e-5, and the Adam optimizer with a Cosine linear learning rate scheduler, without warm-up steps.

Multi-Task RoBERTa (MT RoBERTa). We employ a standard method of multi-task learning approach that is successful in similar text classification tasks (Kochkina, Liakata, and Zubiaga 2018). Intuitively, the classes are related to each other. Hence, if we explore methods to take advantage of the implicit relationships between the classes. Specifically, let $\mathbf{h} \in \mathbb{R}^k$ be the representation returned using RoBERTa where k is the size of the hidden layer. Given $\mathbf{h}$, we jointly train the two output layers for Perception and Accident information (Related to an Accident and Fault). The RoBERTa parameters are shared for each task. To train the model, we simply perform a weighted average of the Cross-Entropy losses defined as

$$L = \sum_{i=1}^{T} w_i CE(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

where T is the number of Tasks, $\mathbf{y}_i$ is the ground-truth classes for task i, $\hat{\mathbf{y}}_i$ are the predictions, $CE()$ is the Cross Entropy loss, and w_i is a weight for task i. After experimenting with various weights for $w_i \in [0, 1]$ we found weights of 1 to perform the best on our validation data.

Multi-Task RoBERTa using the Labeled Powerset Transformation (MTLPT RoBERTa). While MT RoBERTa implicitly incorporates label co-occurrence information by jointly training each class, it does not explicitly capture the information which can limit co-occurrence knowledge acquisition for infrequently co-occurring classes. For multi-label classification, methods have been proposed to take advantage of class co-occurrence (Tsoumakas, Katakis, and Vlahavas 2009). The Labeled Powerset (LP) Transformation is a common approach (Read 2008).

Intuitively, instead of training an output layer that predicts each class independently, e.g., "Positive Perception" and "Cyclist-Fault", a new class is created that combines all of the classes assigned to each instance. In this example, if the Perception is Positive and the Fault is assigned to the Cyclist, the new class would be "Positive-Perception_Cyclist-Fault", which consists of transforming a multi-label problem into a single-label multi-class problem. In the transformed problem, each combination of labels presented in the original dataset is transformed into a single class. Despite the disadvantage of being the worst-case computational complexity (involving 2^L classes in the transformed multi-class problem where L is the total classes), the LP transformation is simple, considers label correlations, and after transformation, any multi-class algorithm can be used for classification. However, many of the newly generated classes appear to infrequently make adequate predictions. Hence, we use this task as an additional multi-task regularizer. Specifically, we use the LP transformation as an auxiliary output layer trained in a multi-task setting. The auxiliary output layer is not used for inference.

Formally, we define the new multi-task classification task as

$$L_i = CE(\mathbf{y}_i, \mathbf{y}_i) + \alpha CE(\mathbf{y}_p, \hat{\mathbf{y}}_p)$$

where α is a hyperparameter for the weight of the LP transformed loss, $\mathbf{y}_p$ is the vector of ground-truth for the LP transformed classes, and $\hat{\mathbf{y}}_p$ represents the predictions. We train a model for each class independently (i.e., L_i is a loss function for task i) where the LP transformation outputs are used as a regularizer. We also experiment with combining the LP transformation loss with the MT RoBERTa model described above. Overall, the final loss is defined as

$$L = \alpha CE(\mathbf{y}_p, \hat{\mathbf{y}}_p) + \sum_{i=1}^{T} w_i CE(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

where, again, α is a hyperparameter and w_i is the weight of each of the individual task losses. Empirically, we found setting everything to one (i.e., α and w_i) resulted in the best performance on the validation dataset.

A.2. Summary Statistics for Male vs Female Articles for Unlabeled Corpus

Category	Male	Female
Total Articles	966	603
Yes	588	288
No	378	315
Fault Attribution		
- Cyclist	220	112
- Other	359	172
- Unknown	387	319
Perception		
- Negative	405	217
- Neutral	359	248
- Positive	202	138

Table 5: Summary Statistics for Male vs Female Articles, out of a total of 11,385 headlines from US-based websites

A.3. Prompts Used in Our Experiments

Prompt for Execute BikeFrame Chain-of-Code.[4]

Bike Frame Simulated Execution

```
input_text = "{title}"
publisher_title = "{publisher_title}"
bike_frame = BikeFrame(input_text, publisher_title)
final_answer = bike_frame.analyze_headline()
print("Final answer:"+ final_answer)
```

Instruction: Generate the expected execution output (output from all print() functions) of the code. You don't have to actually run the code and do not care about "not implemented error".

[4]The DESIGN prompt is available at https://github.com/Zephyr1022/BikeFrames along with the dataset.

Prompt for Direct Prompting

Direct Prompting

System: Reads a given input text and analyzes news headlines about cycling by determining whether an accident is mentioned, identifying the party at fault, and assessing the readers' perception towards cyclists. The final answer should be formatted as follows: "Final answer: Accident: [Yes/No], Fault: [Cyclist/Other/Unknown], Perception: [Negative/Positive/Neutral]."
User: text for the task: {title} Final answer should be at the end of your answer and its format should be like "Final answer: your_answer". Generate output following the task description above. Output:

Prompt for Zero-shot CoT

Zero-shot CoT

System: Reads a given input text and analyzes news headlines about cycling by determining whether an accident is mentioned, identifying the party at fault, and assessing the readers' perception towards cyclists. The final answer should be formatted as follows: "Final answer: Accident: [Yes/No], Fault: [Cyclist/Other/Unknown], Perception: [Negative/Positive/Neutral]."
User: text for the task: {title} Final answer should be at the end of your answer and its format should be like "Final answer: your_answer". Generate output following the task description above. Output: Let's think step by step.