CoCoLoFa: News Comment Sections with Common Logical Fallacies

Anonymous ACL submission

Abstract

Detecting logical fallacies in texts could improve online discussion quality by helping users spot argument flaws and construct better arguments. However, automatically identifying logical fallacies in the wild is not easy. Fallacies are often buried inside arguments that sound convincing; over 100 types of logical fallacies exist. Building large labeled datasets needed for developing automatic fallacy detection models can be expensive. This paper introduces CoCoLoFa, the largest logical fallacy dataset, containing 5,772 comments for 647 news articles, with each comment labeled for fallacy presence and type. To collect data, we first specified a fallacy type (*e.g.*, slippery slope) and a news article to crowd workers, then asked them to write comments that embody the fallacy in response to the article. We built an LLM-powered assistant in the interface to help workers draft and refine comments. Experts rated the writing quality and labeling validity of CoCoLoFa as high and reliable. Models trained on CoCoLoFa achieved the highest fallacy detection performance (F1=0.65) on real-world news comments from the New York Times, surpassing those trained on other datasets and even GPT-4.

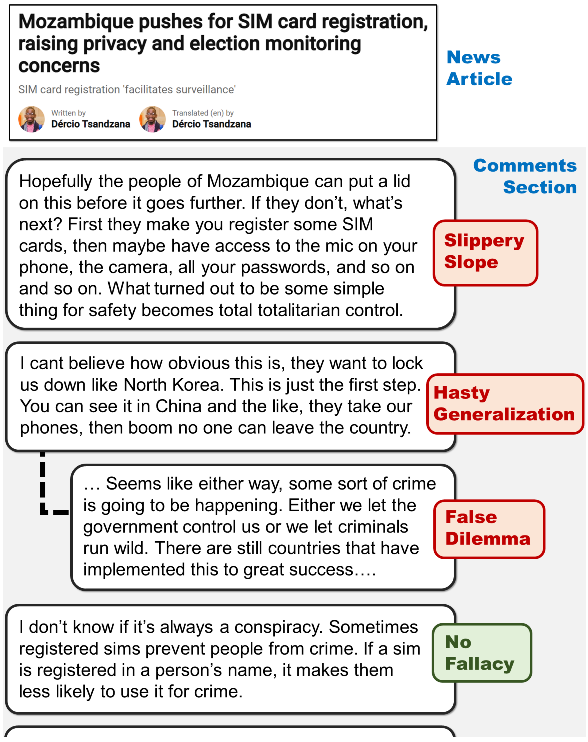

Figure 1: Examples from CoCoLoFa. For each news article, we hired crowdworkers to form a thread of comment. Each worker was asigned to write a comment with either a specific type of logical fallacy or a neutral argument. Everything in CoCoLoFa is CC-licensed and releasable.

1 Introduction

Logical fallacies are reasoning errors that undermine an argument's validity (Walton, 1987). Common fallacies in online conversations like slippery slope, appeal to nature, or false dilemma not only lead to poor-quality discussions (Sahai et al., 2021) but also make arguments appear more dubious, promoting misinformation (Jin et al., 2022). Being able to automatically detect logical fallacies in texts will help users to more easily identify problems in arguments and to compose their own arguments more effectively. However, automatically identifying logical fallacies *in the wild* is challenging. Fallacies are often buried inside arguments that sound convincing but are, in fact, flawed (Powers, 1995). Furthermore, over 100 types of logical fallacies exist (Arp et al., 2018). The nature of the problem makes it extremely expensive to build large-scale labeled datasets needed for developing automatic fallacy detection models.

Prior work has attempted to create datasets for logical fallacies, each addressing the great challenge of labeling in unique ways (Table 1). The Logic dataset collected examples from textbooks (Jin et al., 2022); the LogicClimate dataset gathered instances from news articles, focusing on a narrow topic range to simplify the iden-

Dataset	Genre	# Topics	# Fallacies	# Item	# Neg. Item.	# Sentences per Item	# Tokens per Item	Vocab.
LOGIC (Jin et al., 2022)	Quiz questions	N/A	13	2,449	0	1.92	31.20	7,624
LOGICCLIMATE (Jin et al., 2022)	Sentences in news article	1	13	1,079	0	1.43	39.90	6,419
Reddit (Sahai et al., 2021)	Online discussion	N/A	8	3,358	1,650	2.98	57.01	15,814
COCOLOFA (Ours)	Online discussion	20+	8	5,772	1,918	4.19	70.00	14,894

Table 1: Comparison with other datasets. COCOLOFA contains the largest amount of items spanning diverse topics. Moreover, it boasts the highest average number of sentences and tokens per item among all datasets.

tification of common fallacious arguments related to those topics (Jin et al., 2022); the dataset proposed by Sahai et al. (2021) leveraged existing community labels from Reddit users. However, these datasets cannot effectively train models to detect logical fallacies in real-world scenarios: Textbook examples, being educational, make fallacies obvious, short, and lack subtle or ambiguous cases. Narrow topic focuses, like climate change, miss the wide range of online discussion topics. Moreover, Reddit's community-labeled data often removed crucial context by isolating comments from their original discussion threads, hindering effective detection. Some datasets' absence of negative examples suggests they were not intended for developing detection models.

This paper introduces COCOLOFA, a dataset containing comment sections from 647 news articles, with each comment labeled for fallacy presence and type (Figure 1). The intuition of our data collection approach is first to specify a fallacy type (*e.g.*, slippery slope) and also present a news article (*e.g.*, on abortion laws) to crowd workers, and then ask them to write comments that embody the fallacy in response to the article (*e.g.*, "Abortion legalization leads to normalization of killing"). Recognizing the difficulty of this writing task, we built an LLM-powered assistant in the interface to help workers draft and refine comments with detailed editing suggestions and examples from LLMs. 114 workers contributed to COCOLOFA, which contained 5,772 comments. Compared to previous datasets, COCOLOFA is the largest collection of text units labeled with logical fallacies, spanning the broadest array of topics, and featuring the longest text units on average (Table 1). Two professional editors rated the writing quality and labeling validity of COCOLOFA as high and reliable. Our experiments show that models trained on COCOLOFA achieved the highest fallacy detection performance (F1=0.65) on online news comments from the New York Times, surpassing those trained on other datasets and even GPT-4.

This paper's contribution is threefold. First, we constructed COCOLOFA, the largest dataset of logical fallacies featuring the longest texts across the broadest range of topics. Second, we highlighted the power of combining crowdsourcing with LLMs, allowing researchers to generate data that naturally would be difficult to produce. Finally, through extensive experiments, we illustrated methods to benchmark a model's capability in detecting and classifying logical fallacies in real-world scenarios, including situations where slight contextual changes affect the identification of fallacies.

2 Related Work

2.1 Logical Fallacy Data Collection

As discussed in the Introduction (Section 1), several studies have tried to collect logical fallacies data. Habernal et al. (2017) created a game-based system enabling players to write and label fallacious arguments. A follow-up study later collected 6 types of logical fallacies data and ended up labeling 430 arguments (Habernal et al., 2018). Some studies collected logical fallacies within news articles. For instance, Da San Martino et al. (2019) annotated 7,485 instances from 451 news articles with 18 propaganda techniques, out of which 12 techniques are logical fallacies. Jin et al. (2022) collected 2,449 logical fallacies examples from student quiz websites, and annotated 1,079 fallacious sentences with 13 fallacy types from news articles related to climate change. It is noteworthy that these datasets provided only positive samples for classification, not for identifying logical fallacies.

For identifying logical fallacies in online discussions, Sahai et al. (2021) proposed a strategy to collect fallacious and non-fallacious comments from Reddit by identifying the keywords of fallacies in the response of each comment (*i.e.*, community labels). They used this approach to collect 1,708 fallacious comments, corresponding with 1,650 non-fallacious comments. The writing style in this dataset closely matches that of COCOLOFA, but its limitation is that the highlighted fallacious comments are sometimes obvious and also removed from their original context.

2.2 Human-LLMs Collaboration in Crowd Work

Veselovsky et al. (2023) found that 33-46% of crowd worker's submitted summaries were created using LLMs. Rather than viewing this as an issue, we saw it as an opportunity. By integrating LLMs directly into the worker's interface, we eliminated the need for workers to switch between pages and gain control over the prompts and generation process. Through careful design, LLMs can assist crowd workers in performing complex tasks efficiently, enhancing performance. For instance, Bartolo et al. (2022) introduced Generative Annotation Assistants (GAAs), which provide suggestions to annotators in a Dynamic Adversarial Data Collection task, helping them identify model-fooling examples more easily by accepting, modifying, or rejecting these suggestions. This approach not only accelerated the annotation process by over 30% but also increased model fooling rates by more than 5x. GAAs succeed because humans alone struggle to create model-fooling examples. Similarly, we found it challenging to craft comments with logical fallacies and coherent arguments, highlighting the utility of such assistance in our work.

3 COCOLOFA Dataset Construction

We constructed COCOLOFA, a dataset that contains 5,772 comments in the online comment sections of 647 news articles. Each comment is tagged for the presence of logical fallacies and, where applicable, the specific type of fallacy. Online crowd workers, aided by GPT-4 integrated into their interface, wrote these comments. COCOLOFA also includes the titles and contents of the news articles, all of which are CC-BY 3.0 licensed. We split the dataset into train (70%), development (20%), and test (10%) sets by article, ensuring a balanced representation of 21 topics across the splits. The dataset creation process is as follows.

3.1 Selecting News Articles

We crawled news articles from Global Voices,[1] an online news platform where all of their news articles are under the CC-BY 3.0 license.

To simulate heated online discussions, we took a data-driven approach to select news articles on topics that often provoke disagreements and numerous opinions. We first selected a set of article tags, provided by Global Voices, that are traditionally more "controversial", such as *politics*, *women-gender*, *migration-immigration*, and, *freedom-of-speech*. The full list was in Appendix A Second, we crawled all the 25,370 articles published from Jan. 1st, 2005, to Jun. 28th, 2023, that have these tags. Third, we trained an LDA model (Blei et al., 2003) to discover 70 topics within these news articles. Finally, according to the top 40 words of each topic, we manually selected 21 interested topics and filtered out the news articles that are irrelevant to the interested topics. Appendix A shows all the topics and the top 10 words. Using top frequent words to select representative events was also used in constructing other datasets that sampled real-world events (Huang et al., 2016). As a result, a total of 15,334 news articles were selected, of which 650 published after 2018 were randomly selected to construct the COCOLOFA dataset.

3.2 Fallacy Types Included in COCOLOFA

Over 100 informal logical fallacies exist (Arp et al., 2018), making it impractical to cover all in a dataset. We reviewed how past studies, such as Sahai et al. (2021), Jin et al. (2022), Habernal et al. (2017), and Da San Martino et al. (2019), selected fallacy types. Following Sahai et al. (2021), we chose eight common logical fallacies in online discussions: **(1) Appeal to Authority, (2) Appeal to Majority, (3) Appeal to Nature, (4) Appeal to Tradition, (5) Appeal to Worse Problems, (6) False Dilemma, (7) Hasty Generalization, and (8) Slippery Slope.** These eight logical fallacies have been proved to be frequently used and identified in online discussion threads (Sahai et al., 2021). The definitions and examples of these logical fallacies can be found in Appendix D.

[1]Global Voices: https://globalvoices.org/. Besides common news topics like *economics* and *international relations*, Global Voices also focuses on topics related to human rights, such as *censorship*, *LGBTQ+*, *freedom of speech*, and *refugees*.

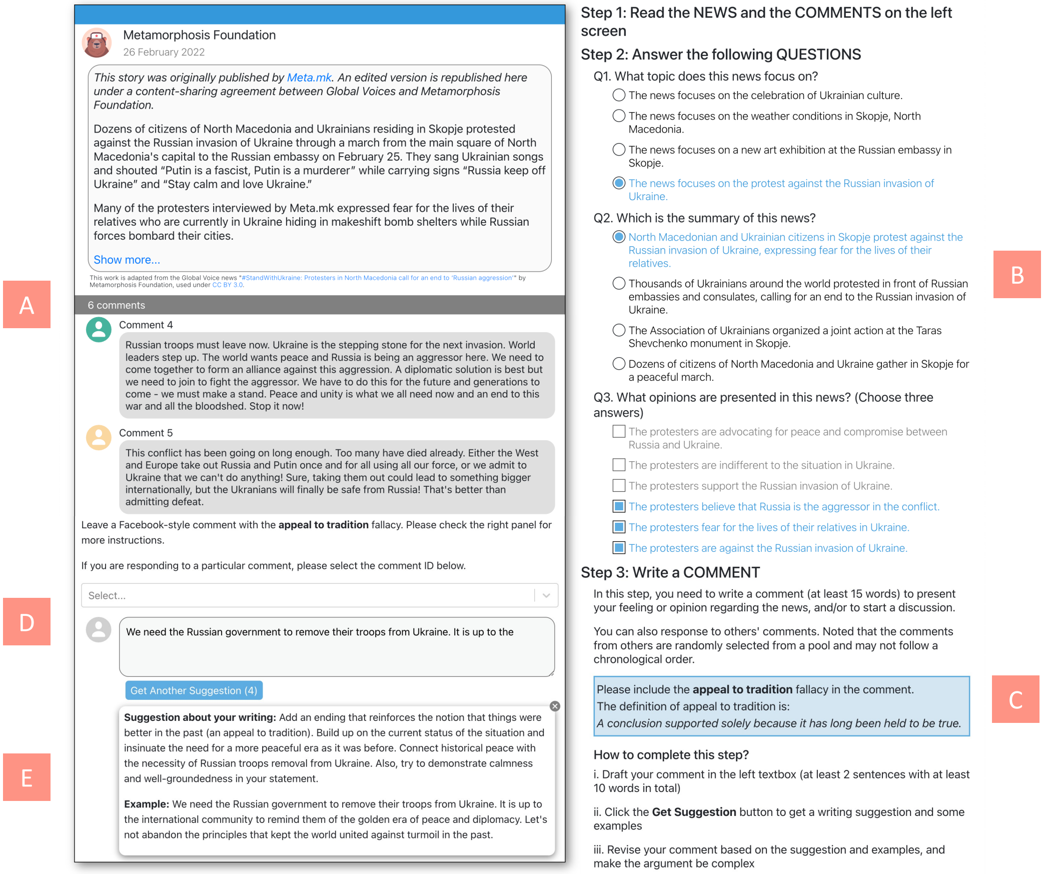

Figure 2: Different components in the task interface: A) The news article and comments, B) Questions for sanity check, C) Instruction of writing fallacious comments, D) Text box and the drop down list for choosing the responded comment, E) GPT-4 generated guideline and example.

3.3 Collecting Comments with Specified Logical Fallacies from Crowd Workers Assisted by LLMs

We designed a crowdsourcing task instructing crowd workers to write comments containing specific logical fallacies. The intuition is that showing an often controversial topic (*e.g.*, abortion) alongside a logical fallacy definition (*e.g.*, slippery slope) allows workers to easily come up with relevant commentary ideas with the fallacy (*e.g.*, "Abortion legalization leads to normalization of killing."). After drafting their idea quickly, LLMs like GPT-4 can be employed to elaborate and refine the comment with the worker. Figure 2 shows the worker interface, which contains two panels: the left is a simulated news comment section; the right contains the instructions and questions. The workflow of crowd workers is as follows.

Step 1: Read the News Article. Upon reaching the task, the worker will be first asked to read the shown news article (Figure 2A). The article was selected by the procedure described in Section 3.1.

Step 2: Answer Attention-Check Questions about the News. For quality control, the worker will then be asked to answer three multiple-choice questions related to the news as an attention check (Figure 2B). These questions are: (1) "What topic does this news focus on?", (2) "Which is the summary of this news?", and (3) "What opinions are presented in this news? (Choose three answers)". We prompted GPT-4 to generate correct and incorrect options for these questions. The prompt used, as shown in Appendix B, was empirically tested and was shown to be effective in filtering out underperforming workers. The workers whose answering accuracy was lower than 0.6 were disallowed to enter our system for 24 hours.

Step 3: Draft a Comment Containing the Specified Logical Fallacy and Revise with LLMs. We divided the writing task into two smaller steps: drafting and revising.

First, workers were presented with a logical fallacy definition, such as "Appeal to Tradition" (Figure 2C),[2] and then tasked with writing a response to a news article, requiring at least two sentences or a minimum of 10 words (Figure 2D). They had access to comments from other workers on the same article and could either comment on the article directly or reply to existing comments. Each worker was exposed to an article only once. The requester assigned the fallacy for each task; the process is described in Section 3.4).

Second, after drafting, workers were instructed to click the "Get (Another) Suggestion" button for a detailed revision suggestion and example embodying the fallacy (Figure 2E). We prompted GPT-4 (see Appendix B) to generate the suggestion and example automatically based on *(i)* the news article, *(ii)* the comment draft, and *(iii)* the target fallacy. Workers can revise their comments and click the button again for new suggestions based on the revised comment. Within each task, they can click the button up to five times. Copy-and-paste was disabled in the interface, so workers had to type their comments.

This workflow employed LLMs to assist workers, making a hard writing task easier. Meanwhile, it forced workers to provide their insights as input for LLMs, ensuring data diversity and a human touch. The built-in LLM assistance decreased the likelihood of workers turning to external LLMs, allowing researchers to provide a prompt that fully considered the context, including news content, the specific fallacy, and workers' opinions.

3.4 Crowdsourced Data Collection Process

Our data collection process allowed workers to not only comment on news but also to reply to others' comments. To achieve this, we used a data-collecting process with three iterations. For each iteration, we added the comments collected from previous iterations underneath the article section on the interface. Workers in the 2nd and 3rd iterations can respond to previous comments. Above the comment's text box (Figure 2D), we provided a drop-down list for workers to choose the comment they wanted to reply to.

We collected our data on Amazon Mechanical Turk (MTurk) using Mephisto, an open-source platform designed to launch, monitor, and review crowdsourcing tasks. For each news article, we recruited 9 workers (3 per iteration) across 9 Human Intelligence Tasks (HITs) to write comments.[3] Each HIT was randomly assigned a logical fallacy from the eight types, each with a 10% chance, or a 20% chance to comment without fallacious logic. Workers were restricted to commenting on each article only once, with each task priced at $2 USD. One HIT generally takes about 10 minutes, leading to an estimated hourly wage of $12. The study received approval from the leading researcher's institute's IRB office.

We posted HITs in small batches, closely monitoring data quality daily and manually removing low-quality responses as necessary. Completing 50 news articles typically took about one week, likely due to our exclusive use of workers with Masters Qualifications. 114 workers contributed to the dataset. As each worker can only see each article once, we decided to exclude worker ID from data release. After removing articles with fewer than 6 comments, the final dataset contained 647 news articles and 5,772 comments. Table 2 shows the basic statistics of COCOLOFA.

	# news	# comments	w/ fallacy	w/o fallacy
All	647	5,772	3,854	1,918
Train	452	4,029	2,689	1,340
Dev	129	1,155	758	397
Test	66	588	407	181

Table 2: Statistics of the COCOLOFA dataset. We devided COCOLOFA into Train, Dev, and Test sets at ratios of 0.8, 0.2, and 0.1 respectively.

Fallacy	Expert 1	Expert 2	Avg.
Appeal to authority	0.73	0.82	0.78
Appeal to majority	0.72	0.88	0.80
Appeal to nature	0.61	0.75	0.68
Appeal to tradition	0.53	0.61	0.57
Appeal to worse problems	0.78	0.66	0.72
False dilemma	0.46	0.55	0.51
Hasty generalization	0.46	0.38	0.42
Slippery slope	0.76	0.68	0.72

Table 3: Cohen's κ agreement between experts and our labels. Experts agreed with our labels at a substantial level ($\kappa \in [0.6, 0.8]$) across most fallacy types.

4 Data Quality Assessment

To assess the text quality of COCOLOFA, we hired two professional editors from UpWork.[4] Both ed-

[2]We used the definitions from Logically Fallacious: https://www.logicallyfallacious.com/

[3]Four MTurk's built-in worker qualifications were used: Masters Qualification, Adult Content Qualification, and Locale (US, CA, AU, GB, and NZ Only) Qualification.

[4]UpWork: https://www.upwork.com

itors had over 20 years of editing experience and PhDs in Linguistics. They were paid $50-$60 per hour, and they typically spent 30 to 45 minutes reviewing each article, which included 9 comments.

We randomly selected 20 new articles and asked the editors to annotate fallacies in all comments. For each fallacy type, we converted labels into binary Yes/No (indicating the presence of the fallacy) and calculated the Cohen's kappa (κ) agreement between experts' and CoCoLoFa's labels (see Table 3). Most fallacy types show substantial agreement levels (0.6-0.8), indicating that the workers accurately included the requested fallacies in their comments. By comparison, the average κ for each fallacy type in the Reddit dataset was just 0.51 (Sahai et al., 2021).

We also asked the experts to respond to the following questions for each comment using a 5-point Likert scale, from 1 (Strongly Disagree) to 5 (Strongly Agree):

- Q1: I feel confident about my annotation. **(Confidence)**
- Q2: I need some additional context to annotate the comment. **(Context Dependent)**
- Q3: This comment appears to have been written by a person rather than by a language model such as ChatGPT. **(Written by Human)**
- Q4: Disregarding any logical fallacies, this comment is grammatically correct and fluently written. **(Text Quality)**

The average scores of Q1 and Q2 were 4.64 (SD=0.62) and 1.42 (SD=0.68), respectively, suggesting that the comments are self-content and have enough information for identifying fallacies. The average scores of Q3 and Q4 were 4.40 (SD=0.82) and 4.13 (SD=1.17), respectively, suggesting that the comments we collected have great quality and are mostly written by workers themselves.[5]

5 Experimental Results

We evaluated three baseline models with both detection and classification tasks on CoCoLoFa and other logical fallacies datasets shown in Table 1.

5.1 Three NLP Tasks

Fallacy Detection. Given a comment, the model predicts whether the comment is fallacious or not. Logic and LogicClimate only have positive examples, so we only reported Recalls.

[5] Other analyses, such as topic distribution and the diversity of thread structure, are shown in Appendix C.

Trained On	Model	Logic	LogicClimate	Reddit			CoCoLoFa		
		R	R	P	R	F	P	R	F
Reddit	BERT	51	83	66	69	68	71	91	80
	NLI	59	72	66	68	67	71	93	80
CoCoLoFa	BERT	54	77	65	44	53	86	76	81
	NLI	53	66	58	43	50	81	83	82
	GPT-4	80	31	62	57	60	88	37	52

Table 4: The result of fallacy detection task. We trained models on Reddit and CoCoLoFa datasets, and tested them on Logic, LogicClimate, Reddit, and CoCoLoFa. For Logic and LogicClimate, we reported the `Recall` rate as they only have positive samples. While for others, we reported `Precision`, `Recall`, and `F1` score.

Trained On	Model	Reddit			CoCoLoFa		
		P	R	F	P	R	F
Reddit	BERT	71	70	70	73	71	68
	NLI	70	72	70	71	76	72
CoCoLoFa	BERT	60	53	52	87	87	87
	NLI	54	64	54	89	89	89
	GPT-4	84	80	80	88	86	86

Table 5: The result of fallacy classification task. The high performance for most models suggests that once the fallacies are detected, it is easy for model to discern their types.

Fallacy Classification. Given a fallacious comment, a model predicts the fallacy type that the comment has. In this task, we removed all negative samples. We only evaluated baselines on Reddit and CoCoLoFa because Logic and LogicClimate considered different fallacy types.

Detection and Classification Under Context Attack. Concerns exist about fallacy detection models relying on word patterns instead of grasping argument logic. To test this, we used GPT-4 to add a sentence (*i.e.*, the *attack*) to comments to correct logical fallacies without altering the stance. For instance, the comment "Your friend should not be refusing her doctor's treatment plan" shows an "Appeal to Authority" fallacy. The added sentence, "considering she has repeatedly expressed her trust in her doctor's expertise and acknowledged the potential positive outcome of the treatment," neutralizes the fallacy. Models understanding argument logic would struggle, while those focusing on word patterns would be less affected, as the added context matches the original stance. In this task, we

Trained On	Model	LOGIC	LOGICCLIMATE	Reddit			COCOLOFA		
		R	R	P	R	F	P	R	F
Reddit	BERT	53_{+2}	86_{+3}	64_{-2}	75_{+6}	69_{+1}	70_{-1}	92_{+1}	79_{-1}
	NLI	63_{+4}	84_{+12}	60_{-6}	74_{+6}	66_{-1}	70_{-1}	96_{+3}	81_{+1}
COCOLOFA	BERT	55_{+1}	83_{+6}	63_{-2}	56_{+12}	59_{+6}	83_{-3}	80_{+4}	82_{+1}
	NLI	67_{+14}	87_{+11}	54_{-4}	58_{+15}	56_{+5}	77_{-4}	90_{+7}	83_{+1}
	GPT-4	53_{-27}	9_{-21}	66_{+4}	35_{-22}	46_{-14}	88_{+0}	25_{-12}	39_{-13}

Table 6: The result of context attack on the fallacy detection task. We reported the models' performance after the input was attacked, and calculated the discrepancy between the attacked and original performances, denoted by a subscript. GPT-4 exhibited contrasting behavior compared to finetuned models, indicating differences in their inference strategies.

Trained On	Model	Reddit			COCOLOFA		
		P	R	F	P	R	F
Reddit	BERT	69_{-2}	68_{-2}	68_{-2}	71_{-2}	70_{-2}	67_{-1}
	NLI	46_{-24}	57_{-15}	65_{-5}	44_{-27}	59_{-17}	69_{-3}
COCO-LOFA	BERT	55_{-5}	44_{-9}	53_{+1}	86_{-1}	76_{-9}	81_{-6}
	NLI	58_{+4}	52_{-12}	52_{-2}	84_{-5}	85_{-4}	85_{-4}
	GPT-4	73_{-11}	69_{-11}	69_{-11}	85_{-3}	84_{-2}	84_{-2}

Table 7: The result of context attack on the classification attack. All models have smaller performance decrease on COCOLOFA, indicating its greater resilience to context attacks compared to the Reddit dataset.

run the detection and classification models as the expanded (attacked) comments.

5.2 Baseline Models

BERT. We finetuned BERT (Devlin et al., 2019) and used the encoded embedding of the [CLS] token to predict the label.

NLI. Inspired by Jin et al. (2022), we finetuned an NLI model with a RoBERTa (Liu et al., 2019) as the backbone. We treated the input comment as the premise and the label as the hypothesis. For the detection task, the hypothesis template was "The text [has/does not have] logical fallacy." For the classification task, the template was "The text has the logical fallacy of [label name]."

GPT-4 (Zero-shot). We prompt GPT-4 for zero-shot prediction. (See prompts in Appendix B.) For Reddit and COCOLOFA that provides context information (thread/news title and parent comment) to each instance, the baseline models took the context information as input as well. For BERT and NLI models, the context information is appended to the target comment. For GPT-4, we designed placeholders for the information in the prompt.

5.3 Fallacy Detection Results

We trained the BERT and the NLI models on both Reddit and COCOLOFA datasets, and tested all models on all four datasets. Table 4 shows the results of the detection task. Two key observations emerge. Firstly, based on the numbers, fallacy detection seems tougher in the Reddit dataset than in COCOLOFA. This is likely due to lower inner-annotator agreement in Reddit's labels ($\kappa = 0.51$) compared to COCOLOFA ($\kappa = 0.65$), making Reddit's labels less reliable. Additionally, Reddit's label balance contrasts with CoCoLoFa's positive label skew. Secondly, despite GPT-4's prowess in many NLP tasks, it underperformed in this task, particularly against simpler finetuned models. However, GPT-4 excelled in the LOGIC dataset, the only one that contains the arguments' logic forms. A possible explanation is that GPT-4 excels at grasping the logic behind the words, unlike other models that primarily depend on the text itself for predictions. We explore this idea more thoroughly in Section 5.5.

5.4 Fallacy Classification Results

Table 5 shows the results of the classification task. We only tested models on Reddit and COCOLOFA datasets as they considered the same fallacy types. It is noteworthy that the classification task assumes that a logical fallacy is present, focusing exclusively on instances where gold-standard labels indicate the presence of logical fallacies.

The result shows that most models achieve a high F1 score on both Reddit and COCOLOFA datasets, suggesting that it is easy to distinguish their types once the fallacies are detected. The practical implication is that in efforts to both detect and classify fallacies, the performance of the detection task is more important.

Trained On	Model	P	R	F
Reddit	BERT	44	66	52
	NLI	47	82	60
CoCoLoFa	BERT	55	63	59
	NLI	52	86	65
	GPT-4	67	54	60

Table 8: The result of fallacy detection on the New York Times Comments Dataset. Models trained on CoCoLoFa outperform those trained on Reddit.

5.5 Fallacy Detection and Classification Results Under Context Attack

We show the result of context attack on detection and classification tasks in Table 6 and 7. For each setting, we report the attacked `Precision`, `Recall`, and `F1` score and their differences compared with the original score, denoted using subscript text.

Results in Table 6 show that adding a neutralizing sentence (*i.e.*, the context attack) significantly reduced GPT-4's performance, while the performances of BERT and NLI models showed only minimal changes. This result echos our hypothesis in Section 5.3 that GPT-4 excels in understanding the logic behind words, in contrast to other models (BERT and NLI) that rely more on textual content to make predictions.

Another observation from Tables 6 and 7 is that GPT-4's performance decreased less in CoCoLoFa compared to other datasets. This could be due to CoCoLoFa having the longest average text length per item and being highly self-contained, as experts noted the context was not necessary for predicting labels (Section 4), minimizing the attack's impact.

5.6 Fallacy Detection Results on NYT Dataset

A primary motivation for this work is to facilitate automatic logical fallacy detection *in the wild*. Therefore, the ultimate test for CoCoLoFa should be developing a model using the dataset and applying it to comments from actual news websites. To this end, we tested models using the New York Times Comments Dataset (Kesarwani, 2018). New York Times Comments Dataset contains over 2 million comments on the news articles published in the New York Times in January-May 2017 and January-April 2018. We sampled 2,000 comments and used our finetuned models as well as GPT-4 to identify the logical fallacies in them. From this collection, we then sampled 250 comments and hired a professional editor (one in Section 4) to label the fallacies. The result in Table 8 shows that the models finetuned on CoCoLoFa significantly outperformed models finetuned on Reddit (with Dependent Samples t-test $p < 0.005$), demonstrating that CoCoLoFa is good for developing models that identify logical fallacies in online discussion.

6 Discussion

On Identifying Ad Hominem Fallacies. The editor who labeled CoCoLoFa and NYT comments observed a high frequency of ad hominem fallacies. These fallacies are hard to classify because they must suggest that the reader disregard someone's argument due to personal attacks, rather than merely insult. The distinction between a targeted insult meant to undermine an argument and a simple derogatory remark is often subtle. When in doubt, the editor labeled such instances as "possible" ad hominem or chose "not sure" for greater ambiguities. This case highlights the difficulty of identifying and classifying fallacies in the wild. By improving how we gather and examine fallacy data, we can better understand and tackle these issues, highlighting the value of our work.

7 Conclusion and Future Work

This paper introduces a new logical fallacy detection dataset, CoCoLoFa, curated through a collaboration between LLM and crowd workers. Comprising 647 news articles paired with 5,772 corresponding fallacious and non-fallacious comments, CoCoLoFa offers a valuable resource for research in this domain. Through empirical evaluation, we have shown the efficacy of models trained on CoCoLoFa in identifying logical fallacies in real-world discourse, outperforming existing datasets. Furthermore, our investigation unveiled limitations in current fine-tuned models for logical fallacy detection: their potential ignorance of context and reasoning process. We showed this issue through a novel context attack, emphasizing the need for future research to address this deficiency.

In the future, we aim to design a model that takes both context and reasoning processing into account for identifying logical fallacies. Moreover, while CoCoLoFa currently has eight types of fallacies, the landscape of logical fallacies is vast, comprising over a hundred recognized types. Recognizing this, we will expand CoCoLoFa to include more fallacy types.

8 Limitations

Like most crowdsourced datasets, CoCoLoFa inherits the common biases of using online crowdsourcing platforms to collect data. For example, the crowd workers on Amazon Mechanical Turk do not represent the user population of social media and news platforms. They may care about different topics and have different opinions toward real online users. In addition, the writing style of commenting in the crowdsourcing task may also be different from debating online. Although we developed a platform that simulated the interface of the online news comment section, the real-time feedback and the vibe of online discussion are still difficult to simulate.

Another limitation is that CoCoLoFa currently considers only eight types of fallacy, as we mentioned in the future work. Given that there are many common fallacy types apart from the fallacies we collected, models trained on our dataset may only have a limited ability to detect fallacies in the wild.

9 Ethics Statement

Although CoCoLoFa is collected for logical fallacy detection, we acknowledge the potential misuse of the dataset for training models to generate fallacious comments. Furthermore, our data collection process has revealed that GPT-4 has the capability to generate such comments, posing risks of propagating misinformation online. Therefore, we advocate for research aimed at LLMs to prevent the generation of harmful and misleading content.

References

Robert Arp, Steven Barbone, and Michael Bruce. 2018. *Bad Arguments: 100 of the Most Important Fallacies in Western Philosophy | Wiley*.

Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. 2022. Models in the loop: Aiding crowdworkers with generative annotation assistants. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3754–3767, Seattle, United States. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018. Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, page 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aashita Kesarwani. 2018. [link].

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.

E.C. Pielou. 1966. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, 13:131–144.

Lawrence H. Powers. 1995. The one fallacy theory. *Informal Logic*, 17(2).

Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, page 644–657, Online. Association for Computational Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. (arXiv:2306.07899). ArXiv:2306.07899 [cs].

Douglas N. Walton. 1987. *Informal Fallacies: Towards a Theory of Argument Criticisms*. Benjamins, John, Philadelphia.

A Selected Global Voices and LDA Topics

The selected Global Voices' tags are *politics*, *health*, *environment*, *protest*, *refugees*, *religion*, *war-conflict*, *women-gender*, *migration-immigration*, *gay-rights-lgbt*, *law*, *labor*, *international-relations*, *indigenous*, *humanitarian-response*, *human-rights*, *governance*, *freedom-of-speech*, *ethnicity-race*, *elections*, *disaster*, and *censorship*.

The selected LDA topics and the top 10 words for each topic are shown in Table 9.

B GPT-4 Prompts

Prompt for Generating Attention Check Questions.

> Create `[n_correct]` correct and `[n_incorrect]` incorrect answers based on the question: `[question]`
>
> Here is the news content: `[news]`
>
> Here is an example output format:
>
> - Correct Answer 1: This is the 1st correct answer
>
> - ...
>
> - Correct Answer n: This is the n-th correct answer
>
> - Wrong Answer 1: This is the 1st wrong answer
>
> - ...
>
> - Wrong Answer n: This is the n-th wrong answer

Prompt for Generating Guideline and Example.

> Users will provide a news and a part of their comment toward the news. Please give a suggestion of writing the remaining comment. Below are some criteria for the comment:
>
> 1. The comment should be in the style of commenting on Facebook posts
>
> 2. The comment should be concise
>
> 3. If there is no `[fallacy_type]` fallacy in the comment, include it in. Otherwise, develop the logic further
>
> 4. The `[fallacy_type]` fallacy should be as subtle as possible.
>
> The definition of `[fallacy_type]` is: `[definition]`
>
> The output should be
>
> <guideline>A guideline of writing the comment. The guideline should be concrete</guideline>
>
> <example>An example of the comment that matches the guidelines. The example should be an extension of the user's draft</example>

Prompt for Context Attack.

> Some people may think the following piece of text, [ORIGINAL STATEMENT], embodies some forms of logical fallacies. This could be caused by the fact that this piece of text is relatively short and presented in isolation without relevant context. Please generate one sentence, [ADDED CONTEXT], that can be attached at the end of this piece of text in order to eliminate the concerns of embodying logical fallacies. Namely, "[ORIGINAL STATEMENT] [ADDED CONTEXT]" will not be considered as having logical fallacies. The added sentence, [ADDED CONTEXT], needs to align with the stance or sentiment of [ORIGINAL STATEMENT]. Do not use any transition words like "but" or "however" in [ADDED CONTEXT] that might reverse the stance or sentiment of it.
>
> [ORIGINAL STATEMENT]: `[comment]`

ID	Topic	Top 10 words
3	Protest	march, protest, movement, social, public, wing, people, protests, right, support
4	International Relations	minister, government, prime, prime_minister, corruption, public, office, state, party, general
10	Race Issue	black, art, white, racism, work, culture, artists, people, cultural, artist
15	Women Rights	women, violence, men, woman, sexual, gender, female, girls, rape, harassment
21	Russo-Ukrainian War	russian, russia, ukraine, soviet, kazakhstan, country, ukrainian, central, kyrgyzstan, state
28	Environmental Issue	indigenous, climate, change, mining, environmental, climate_change, communities, global, region, land
29	Gender Issue	sex, gay, marriage, lgbt, abortion, sexual, same, homosexuality, lgbtq, community
30	Human Rights	rights, human, human_rights, international, activists, people, groups, activist, community, organizations
31	Drug Issue	venezuela, drug, latin, venezuelan, america, latin_america, trafficking, panama, vez, drugs
32	Police Brutality	police, protests, protesters, protest, people, violence, government, security, video, forces
35	Immigration / Refugees	bangladesh, refugees, country, indonesia, sri, immigration, people, refugee, migrants, border
36	COVID / Health Issue	health, medical, people, pandemic, cases, hospital, doctors, hiv, government, virus
45	Legislation	law, court, legal, laws, data, public, protection, constitution, article, legislation
46	Freedom of Speech	government, freedom, expression, speech, state, freedom_expression, public, media, law, free
47	Election	election, elections, vote, presidential, electoral, candidates, candidate, voters, votes, voting
50	Sustainability	water, food, energy, farmers, power, electricity, waste, plant, rice, river
51	Religious Conflict	religious, muslim, muslims, islam, religion, islamic, hate, ethnic, group, anti
55	Political Debates	political, party, government, opposition, people, country, politics, parties, democracy, power
62	U.S. Politics	united, states, united_states, american, obama, america, president, york, visit, trump
66	Digital Rights	internet, access, users, online, mobile, content, data, websites, google, service
68	East Asian Politics	hong, kong, hong_kong, taiwan, pro, china, democracy, mainland, taiwanese, chinese

Table 9: Top 10 words of the selected topics

Prompt for Detection.

> Determine the presence of a logical fallacy in the given [COMMENT] through the logic and reasoning of the content. If the available information is insufficient for detection, output “unknown.” Utilize the [TITLE] and [PARENT_COMMENT] as context to support your decision, and provide an explanation of the reasoning behind your determination. The output format should be [YES/NO/UNKNOWN] [EXPLANATIONS]
>
> [TITLE]: `[title]`
>
> [PARENT_COMMENT]: `[parent]`
>
> [COMMENT]: `[comment]`

Prompt for Classification.

> Determine the type of fallacy in the given [COMMENT]. The fallacy would be one of in the [LOGICAL_FALLACY] list. Utilize the [TITLE] and [PARENT_COMMENT] as context to support your decision, and provide an explanation of the reasoning behind your determination.
>
> [COMMENT]: `[comment]`
>
> [LOGICAL_FALLACY]" `[fallacy]`
>
> [TITLE]: `[title]`
>
> [PARENT_COMMENT]: `[parent]`

Topic	Train	Dev	Test
Protest	2.9%	3.1%	1.5%
International Relations	11.9%	10.9%	12.1%
Race Issue	4.9%	4.7%	4.5%
Women Rights	10.0%	7.8%	10.6%
Russo-Ukrainian War	8.2%	7.8%	6.1%
Environmental Issue	9.3%	8.5%	7.6%
Gender Issue	3.5%	3.1%	4.5%
Human Rights	1.8%	1.6%	3.0%
Drug Issue	0.2%	0.0%	0.0%
Police Brutality	15.9%	14.7%	19.7%
Immigration / Refugees	7.3%	4.7%	6.1%
COVID / Health Issue	11.3%	14.7%	15.2%
Legislation	6.4%	6.2%	6.1%
Freedom of Speech	15.3%	11.6%	12.1%
Election	6.0%	4.7%	4.5%
Sustainability	5.3%	4.7%	4.5%
Religious Conflict	2.0%	2.3%	1.5%
Political Debates	4.2%	3.9%	3.0%
U.S. Politics	0.2%	0.8%	1.5%
Digital Rights	11.9%	13.2%	10.6%
East Asian Politics	9.5%	8.5%	9.1%

Table 10: Proportions of different topics in each split. The distribution of topics remains consistent across all splits, with each topic maintaining a similar proportion regardless of the split.

Type	# Unique Structures	# Articles	Evenness (J)
Flat	4	100	0.29
Single Conversation	79	471	0.81
Multi Conversation	30	51	0.96
Complex	21	25	0.98
Total	134	647	0.79

Table 11: Statistics of the thread structure. The 647 comment threads we collected formed 134 unique structures, with the majority falling under the category of 'Single Conversation'.

C Data Diversity

CoCoLoFa covers diverse topics. Table 10 shows the proportions of each topic in CoCoLoFa. As each news article may have multiple topics, the summation of each column may exceed 100%. The result indicates that most of the news we collected is related to *international relations*, *women rights*, *police brutality*, *COVID/health issue*, *freedom of speech*, *digital rights*, and *East Asian politics*.

CoCoLoFa contains comment sections with diverse thread structures. To analyze the structure of discussion threads in CoCoLoFa, we categorized the structures into four types:

- **Flat:** Every comment directly responds to the news article.
- **Single Conversation:** Only one comment received one or more replies.
- **Multiple Conversations:** Several comments received replies, but none of these replies received their own responses (no second-layer responses).
- **Complex:** Any structure that does not fit into the above categories.

We calculated the diversity of structures using the evenness index J, proposed by Pielou (1966):

$$J = H/\log S \tag{1}$$

where

$$H = -\sum_i p_i \log p_i \tag{2}$$

H is the Shannon Diversity Index (Shannon, 1948), S is the total number of unique structures, and p_i is the proportion of a unique structure within its category. The value of J ranges from 0 to 1, with higher values indicating greater evenness in structure diversity. Table 11 shows the statistics for each thread structure type in CoCoLoFa. In total, CoCoLoFa had 134 unique thread structures, most of which were of Single Conversation. The diversity of thread structures was high.

D Details of Fallacy Types

We draw the definition and example of the chosen fallacies from Logically Fallacious[6].

Appeal to authority. *Definition:* Insisting that a claim is true simply because a valid authority or expert on the issue said it was true, without any other supporting evidence offered. *Example:* Richard Dawkins, an evolutionary biologist and perhaps the foremost expert in the field, says that evolution is true. Therefore, it's true.

Appeal to majority. *Definition:* When the claim that most or many people in general or of a particular group accept a belief as true is presented as evidence for the claim. Accepting another person's belief, or many people's beliefs, without demanding evidence as to why that person accepts the belief, is lazy thinking and a dangerous way to accept information. *Example:* Up until the late 16th century, most people believed that the earth was the center of the universe. This was seen as enough of a reason back then to accept this as true.

[6] https://www.logicallyfallacious.com/

Appeal to nature. *Definition:* When used as a fallacy, the belief or suggestion that "natural" is better than "unnatural" based on its naturalness. Many people adopt this as a default belief. It is the belief that is what is natural must be good (or any other positive, evaluative judgment) and that which is unnatural must be bad (or any other negative, evaluative judgment). *Example:* I shop at Natural Happy Sunshine Store (NHSS), which is much better than your grocery store because at NHSS everything is natural including the 38-year-old store manager's long gray hair and saggy breasts.

Appeal to tradition. *Definition:* Using historical preferences of the people (tradition), either in general or as specific as the historical preferences of a single individual, as evidence that the historical preference is correct. Traditions are often passed from generation to generation with no other explanation besides, "this is the way it has always been done"—which is not a reason, it is an absence of a reason. *Example:* Marriage has traditionally been between a man and a woman; therefore, gay marriage should not be allowed.

Appeal to worse problems. *Definition:* Trying to make a scenario appear better or worse by comparing it to the best or worst case scenario. *Example:* Be happy with the 1972 Chevy Nova you drive. There are many people in this country who don't even have a car.

False dilemma. *Definition:* When only two choices are presented yet more exist, or a spectrum of possible choices exists between two extremes. False dilemmas are usually characterized by "either this or that" language, but can also be characterized by omissions of choices. *Example:* You are either with God or against him.

Hasty generalization. *Definition:* Drawing a conclusion based on a small sample size, rather than looking at statistics that are much more in line with the typical or average situation. *Example:* My father smoked four packs of cigarettes a day since age fourteen and lived until age sixty-nine. Therefore, smoking really can't be that bad for you.

Slippery slope. *Definition:* When a relatively insignificant first event is suggested to lead to a more significant event, which in turn leads to a more significant event, and so on, until some ultimate, significant event is reached, where the connection of each event is not only unwarranted but with each step it becomes more and more improbable. *Example:* We cannot unlock our child from the closet because if we do, she will want to roam the house. If we let her roam the house, she will want to roam the neighborhood. If she roams the neighborhood, she will get picked up by a stranger in a van, who will sell her in a sex slavery ring in some other country. Therefore, we should keep her locked up in the closet.