# CoCoLoFa: A Dataset of News Comments with Common Logical Fallacies Written by LLM-Assisted Crowds

**Min-Hsuan Yeh**[1] **Ruyuan Wan**[2] **Ting-Hao 'Kenneth' Huang**[2]
[1]University of Wisconsin-Madison, Madison, WI, USA.
samuelyeh@cs.wisc.edu
[2]The Pennsylvania State University, University Park, PA, USA.
{rjw6289,txh710}@psu.edu

## Abstract

Detecting logical fallacies in texts can help users spot argument flaws, but automating this detection is not easy. Manually annotating fallacies in large-scale, real-world text data to create datasets for developing and validating detection models is costly. This paper introduces CoCoLoFa, the largest known English logical fallacy dataset, containing 7,706 comments for 648 news articles, with each comment labeled for fallacy presence and type. We recruited 143 crowd workers to write comments embodying specific fallacy types (*e.g.*, slippery slope) in response to news articles. Recognizing the complexity of this writing task, we built an LLM-powered assistant into the workers' interface to aid in drafting and refining their comments. Experts rated the writing quality and labeling validity of CoCoLoFa as high and reliable. BERT-based models fine-tuned using CoCoLoFa achieved the highest fallacy detection (F1=0.86) and classification (F1=0.87) performance on its test set, outperforming the state-of-the-art LLMs. Our work shows that combining crowdsourcing and LLMs enables us to more effectively construct datasets for complex linguistic phenomena that crowd workers find challenging to produce on their own. CoCoLoFa is public at CoCoLoFa.org/.

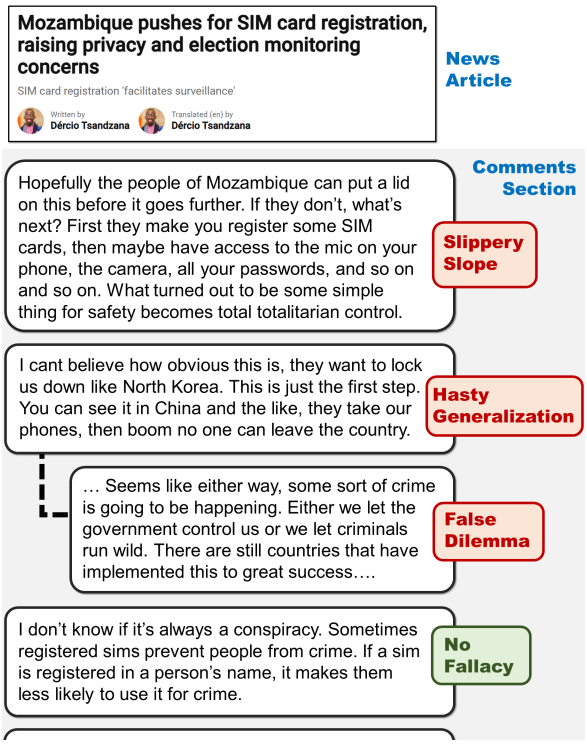



Figure 1: Examples from CoCoLoFa. For each news article, we hired crowd workers to form a thread of comment. Each worker was assigned to write a comment with a specific type of logical fallacy (or a neutral argument) in response to the article.

## 1 Introduction

Logical fallacies are reasoning errors that undermine an argument's validity (Walton, 1987). Common fallacies like slippery slope or false dilemma degrade online discussions (Sahai et al., 2021) and make arguments seem more dubious, fostering misinformation (Jin et al., 2022). Automatically detecting logical fallacies in texts will help users identify argument flaws. However, automatically identifying these fallacies in the wild is not easy. Fallacies are often buried inside arguments that sound convincing (Powers, 1995); over 100 types of logical fallacies exist (Arp et al., 2018). The nature of the problem makes it expensive to build large-scale labeled datasets needed for developing fallacy detection models. Prior works have created datasets for logical fallacies (Table 1): Logic dataset collected examples from online educational materials (Jin et al., 2022); LogicClimate dataset collected instances from news articles, specifically targeting a particular topic range and identifying common fallacious arguments related to those topics (Jin et al., 2022); Argotario dataset was collected using a gamified crowdsourcing approach (Habernal et al., 2017); and the dataset proposed by Sahai et al. (2021) leveraged existing community labels from Reddit users. These previous efforts are in-



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 660–677
November 12-16, 2024 

| Dataset | Genre | # Topics | # Fallacies | Total # Item | # Neg. Item | # Sent. per Item | # Tokens per Item | Vocab. Size |
|---|---|---|---|---|---|---|---|---|
| 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 |
| Argotario (Habernal et al., 2017) | Dialogue | N/A | 5 | 1,338 | 429 | 1.56 | 18.86 | 3,730 |
| 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 | **7,706** | **3,130** | **4.28** | **71.35** | **16,995** |

Table 1: Comparison of datasets of logical fallacies. COCOLOFA is the largest and has the longest text units.

spiring, but they often did not focus on enabling fallacy detection *in the wild*, as each made significant trade-offs to ease the challenges of labeling fallacies: focusing on smaller scales (1,000+ instances; no negative samples), specific topics like climate change rather than a broader range, or clear educational examples instead of complex web discussions. One exception is the Reddit dataset (Sahai et al., 2021), which is relatively large and includes messy Reddit comments. However, it isolates comments from their original threads, limiting the use of context to boost detection and understanding of how fallacies unfold in online discussions.

This paper presents **COCOLOFA**, a dataset containing 7,706 comments for 648 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 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. Our data collection approach replaces the data annotation process with data generation, reducing the need of hiring workers to filter out a large amount of non-fallacious instances at first and making the data collection more scalable. In addition, it increases the ability to control targeted fallacy types for researchers. Compared to previous work (Table 1), COCOLOFA is the largest NLP dataset of logical fallacies, featuring the highest average sentence and word counts per instance. Two experts rated the writing quality and labeling validity of COCOLOFA as high and reliable. The experiments show that COCOLOFA can be used to effectively develop fallacy detection and classification models. As a broader implication, our work shows how crowdsourcing can be integrated with large language models (LLMs) to construct datasets for complex linguistic phenomena that are challenging for crowd workers to produce on their own. This opens up new possibilities for future NLP datasets.

## 2 Related Work

**Logical Fallacy Datasets.** We discussed the major logical fallacy datasets in the Introduction (Section 1); this section focuses on extra studies not previously covered. A follow-up of Argotario (Habernal et al., 2017) collected data on 6 types of logical fallacies and labeled 430 arguments (Habernal et al., 2018). Similarly, Bonial et al. (2022) used the same annotation schema to identify logical fallacies in 226 COVID-19 articles across various mediums. Other research has specifically aimed at detecting logical fallacies in news articles. For example, Da San Martino et al. (2019) annotated 451 news articles with 18 propaganda techniques, 12 of which qualify as logical fallacies. Additionally, Helwe et al. (2024) annotated 200 samples from merged existing datasets with a unified taxonomy and justifications. These datasets are relatively small, highlighting the challenges of annotating large-scale texts for logical fallacies. Emerging research is also exploring the synthesis of logical fallacy datasets using LLMs (Li et al., 2024).

**LLM-Assisted Crowdsourced Data Creation.** Veselovsky et al. (2023) found that many crowd worker's submitted summaries were created using LLMs. We saw it as an interesting opportunity rather than a threat. Integrating LLM assistance directly into the worker's interface offers benefits

for both workers and requesters. For workers, built-in LLMs can aid in complex writing tasks that might otherwise be too challenging and eliminate the need to switch between browser tabs to use external LLMs. For requesters, having a built-in LLM allows for storing all prompts used and texts produced by the LLM, ensuring a more transparent understanding of how LLMs' outputs are woven into the final data. Previous work has integrated AI models into worker interfaces to help produce examples that trigger specific model behaviors, such as model-fooling examples (Bartolo et al., 2022). In this paper, we advocate using LLMs to help workers generate complex examples.

## 3 CoCoLoFa Dataset Construction

We constructed CoCoLoFa, a dataset that contains 7,706 comments in the online comment sections of 648 news articles. Each comment is tagged for the presence of logical fallacies and, where applicable, the specific type of fallacy. 143 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. This section overviews the data construction steps.

### 3.1 Selecting News Articles

We crawled news articles from Global Voices, an online news platform where all of their news articles are under the CC-BY 3.0 license.[1] 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*. 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 irrelevant news articles. 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.[2] See Appendix A for details.

### 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**. Appendix B shows the definitions and examples of these eight fallacies.[3]

### 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 has a simulated news comment section (left) and instructions and questions (right). 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.

[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+*, and *refugees*.

[2] We only selected news published after 2018 because we did not want the news to be too old, so that workers may remember the events in those news and could include their personal feelings and opinions in the comments, making the comments more realistic.

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

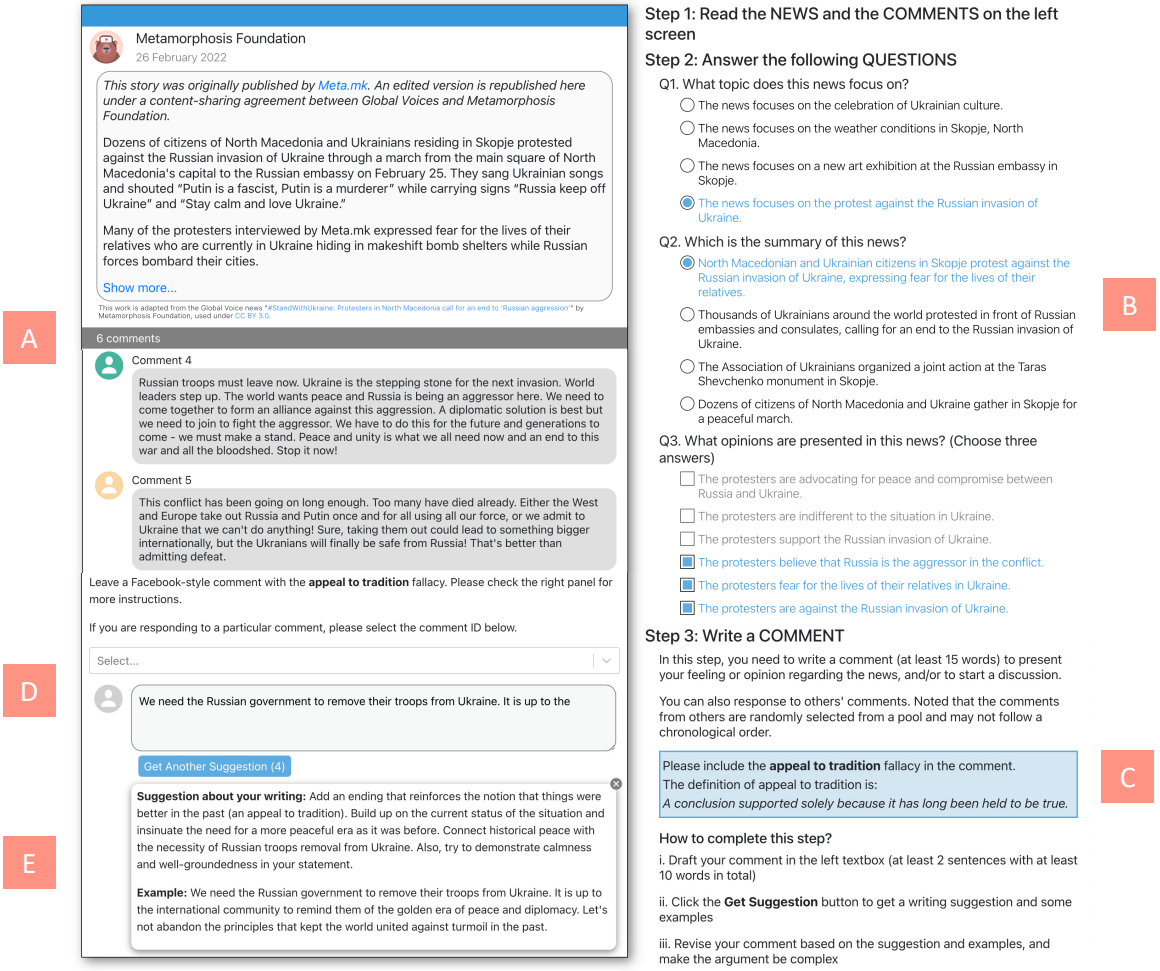



Figure 2: Different components in the task interface: A) The news article and comments, B) Questions for sanity check, C) Instruction of writing fallacious comments, D) Text box and the drop-down list for choosing the responded comment, E) GPT-4 generated guideline and example.

**Step 2: Answer Attention-Check Questions about the News.** As an attention check, the worker will then be asked to answer three multiple-choice questions related to the news (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 (see Appendix C) was empirically tested and was 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), 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 could see comments from other workers on the same article and had the option to either comment directly on the article or reply to existing comments. Each worker was exposed to an article only once. We assigned the fallacy for each task (see Section 3.4). The fallacy definitions we provided on the interface were a shorten version so that the instruction can be concise and easy to follow. The shorten version of fallacy definitions is detailed in Appendix B. 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 C) to generate the suggestion and example automatically based on *(i)* the news article, *(ii)* the comment draft, and *(iii)* the target fallacy. Workers can re-

| | # news | # comments | w/ fallacy | w/o fallacy |
|---|---|---|---|---|
| All | 648 | 7,706 | 4,576 | 3,130 |
| Train | 452 | 5,370 | 3,168 | 2,202 |
| Dev | 129 | 1,538 | 927 | 611 |
| Test | 67 | 798 | 481 | 317 |

Table 2: Statistics of the CoCoLoFa dataset. We divided CoCoLoFa into Train, Dev, and Test sets at ratios of 0.7, 0.2, and 0.1 respectively.

vise 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 (5) times. Copy-and-paste was disabled in the interface, so workers had to type their comments.

**Rationale for the Workflow Design.** This workflow used 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. Notably, our approach—having workers write comments embodying a particular logical fallacy— is conceptually similar to Argotario (Habernal et al., 2017). Our method differs in two ways: First, we provided real-world news as context, requiring workers to base their fallacious arguments on these articles. Second, we conducted multiple rounds of comment collection for each article, allowing workers to respond to others' comments. These two factors allowed CoCoLoFa to more accurately simulate the comment sections of real-world news websites.

### 3.4 Implementation Details

**Four Rounds of Data Collection.** Our data collection process had four iterations. For each iteration, we added the comments collected from previous iterations underneath the article section on the interface. Workers in the 2nd to 4th iterations can respond to previous comments by selecting the comment ID from a drop-down list (Figure 2D). Each worker only interacted with an article once.

**Probability of Each Fallacy Type.** We collected our data on Amazon Mechanical Turk (MTurk) using Mephisto, an open-sourced tool for crowdsourcing task management.[4] For each news article, we recruited 12 workers (3 per iteration) across 12 Human Intelligence Tasks (HITs) to write comments.[5] In the first three iterations, each task randomly received one of eight logical fallacy types with a 10% probability, or a 20% chance to comment without fallacious logic. In the fourth iteration, we increased the probability to 60% for comments without fallacious logic and reduced it to 5% for each fallacy type to gather more negative samples. Workers were paid by $2 USD for each HIT, which takes about 10 minutes on average, leading to an estimated hourly wage of $12.

**Resulting Dataset.** We posted HITs in small batches, closely monitoring data quality daily and manually removing low-quality responses, *i.e.*, those that are (1) obviously off-topic (*e.g.*, saying this task is interesting), (2) writing exactly the same comment for multiple articles, or (3) repeating the same word for the whole comment. Completing 50 news articles typically took about one week, likely due to our exclusive use of workers with Masters Qualifications. 143 workers contributed to the dataset. After removing articles with fewer than 6 comments, the final dataset contained 648 news articles and 7,706 comments. Table 2 shows the statistics of CoCoLoFa.

**Worker-LLM Interactions.** Within our study, each worker asked LLM an average of 1.39 times (SD=0.81) when writing a comment. Workers completely followed the LLM's suggestions in only 3% of comments. The average Levenshtein ratio between the worker's comment and the LLM's last suggestion is 0.35 (1 means the sentences are identical), indicating a significant difference. We observed that most workers either paraphrased the suggestions or added details to their comments.

## 4 Data Quality Assessments

We hired two experts from UpWork.com to assess the data quality. We specified that the experts should have abilities of identifying logical fallacies and writing the explanation to justify their annotations in our job description. Both experts we hired are PhD in Linguistics. One has over 25

[4]Mephisto: https://github.com/facebookresearch/Mephisto

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

| Fallacy | CoCoLoFa | | | Reddit | | |
|---|---|---|---|---|---|---|
| | Exp.1 & Lb. | Exp.2 & Lb. | Betw. Exp. | Exp.1 & Lb. | Exp.2 & Lb. | Betw. Exp. |
| Authority | 0.62 | 0.62 | 0.46 | 0.66 | 0.48 | 0.36 |
| Majority | 0.83 | 0.69 | 0.63 | 0.76 | 0.51 | 0.48 |
| Nature | 0.67 | 0.55 | 0.43 | 0.71 | 0.54 | 0.62 |
| Tradition | 0.52 | 0.39 | 0.56 | 0.64 | 0.53 | 0.49 |
| Worse prob. | 0.67 | 0.58 | 0.74 | 0.53 | 0.56 | 0.52 |
| False dilemma | 0.27 | 0.24 | 0.27 | 0.56 | 0.41 | 0.36 |
| Hasty general. | 0.56 | 0.23 | 0.21 | 0.46 | 0.20 | -0.03 |
| Slippery slope | 0.58 | 0.64 | 0.68 | 0.54 | 0.61 | 0.49 |
| None | 0.40 | 0.23 | 0.28 | 0.18 | 0.11 | 0.14 |
| Average | 0.57 | 0.46 | 0.47 | 0.56 | 0.44 | 0.38 |

Table 3: Cohen's $\kappa$ agreement between experts and labels, as well as the agreement between two experts. CoCoLoFa yielded slightly higher agreements.

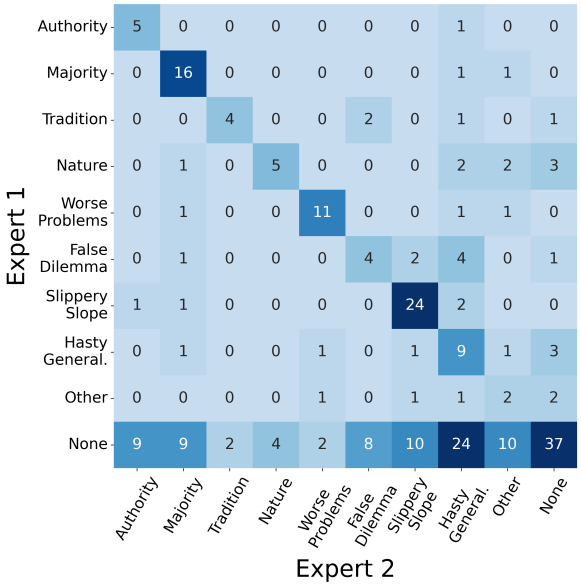



Figure 3: The confusion matrix of the annotation between two experts. Most of the disagreement happened when determining if a comment is fallacious or not.

years of experience in the fields of English composition and rhetoric, and another has over 20 years of experience in translation. Both of them also have rich experience in editing academic articles and volumes. They were compensated $50-$60 per hour. We randomly selected 20 news articles and asked the experts to annotate fallacies in all comments (237 comments in total). For each fallacy type, we converted labels into binary Yes/No (indicating the presence of the fallacy) and calculated the Cohen's kappa ($\kappa$) agreement between experts' and CoCoLoFa's labels, as well as the agreement between two experts. We also sampled 25 instances for each fallacy type plus none (*i.e.*, $25 \times (8+1) = 255$ instances in total) from the Reddit dataset (Sahai et al., 2021) and asked the same experts to annotate them as a comparison. Table 3 shows the results.

**CoCoLoFa yielded slightly higher inter-annotator agreements, while experts often disagreed with each other.** Table 3 shows that experts generally agreed more on the CoCoLoFa's label than on the Reddit dataset. However, Expert 2 consistently showed more disagreement with the labels in both datasets for most fallacy types. Table 3 also shows low agreement between experts on both datasets, particularly for hasty generalization. As shown in Sahai et al. (2021) and Alhindi et al. (2022), this level of $\kappa$ value is normal in annotating logical fallacy data. We computed confusion matrices for experts' annotations and labels in both datasets. The confusion matrix comparing the two experts on CoCoLoFa is shown in Figure 3, and the others are in Appendix E. Figure 3 shows that most disagreements occur in determining the presence of a fallacy rather than its type. We discuss the possible reasons for high disagreement in labeling logical fallacies further in Discussion (Section 6).

**CoCoLoFa was rated more fluent and grammatically correct.** 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) *"Disregarding any logical fallacies, this comment is **grammatically correct and fluently written**."* (Q2) *"This comment appears to have been **written by a person** rather than by a language model such as ChatGPT."* (Q3) *"I feel **confident** about my annotation."* (Q4) *"I need some **additional context** to annotate the comment."* For Q1, CoCoLoFa scored an average of 4.38 (SD=0.91), compared to 4.21 (SD=1.04) for Reddit, suggesting that texts in CoCoLoFa were generally considered more fluent and grammatically correct. For Q2, CoCoLoFa scored 4.39 (SD=0.79), and Reddit scored higher at 4.58 (SD=0.59), indicating that experts found Reddit's texts more human-like. This echoes the findings in Table 3, which shows a lower inter-annotator agreement for Reddit, likely due to its messier, real-world internet text. Although humans sometimes struggle to distinguish LLM-generated texts, the purpose of Q2 was to ensure that CoCoLoFa's text did not obviously appear machine-generated, such as through identifiable errors like repetition, which humans can recog-

nize (Dugan et al., 2023). There was no clear difference between CoCoLoFa and Reddit for Q3 (4.53, 4.57) and Q4 (1.59, 1.60).

**Concerns over argumentation scheme.** During the annotation process, experts identified that some workers did not include fallacies in their comments. Instead, they used argumentation schemes to make their argument "fallacy-like" but valid. To address such an issue, some previous work, such as Ruiz-Dolz and Lawrence (2023), suggested using a series of critical questions of the corresponding argumentation scheme to assess the validity of an argument. However, having annotators or comment writers go through these questions for each comment will significantly limit the scalability of our approach. Given that experts only identified 12 out of 237 comments to be "fallacy-like," we considered our approach a reasonable trade-off.

## 5 Experimental Results

We evaluated three types of baseline models with both detection and classification tasks on Logic, LogicClimate, Reddit, and CoCoLoFa dataset (Table 1). We additionally tested the models using a collection of annotated New York Times news comments. We define the two tasks as follows:

**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`.

**Fallacy Classification.** Given a known fallacious comment, the model classifies it into one of the eight fallacy types. In this task, we removed all negative samples. We only evaluated baselines on Reddit and CoCoLoFa because Logic and LogicClimate used different fallacy type schemes.

### 5.1 Baseline Models

**BERT.** We fine-tuned 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 fine-tuned 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]`."

**LLMs.** We prompted two commonly used LLMs, GPT-4o and Llama3(8B), for detecting and classifying logical fallacy.[6] We designed different prompts (see Appendix C), including both zero-shot and few-shot, as well as Chain-of-Thought (COT) prompting (Wei et al., 2022).

**Use of Context.** For Reddit and CoCoLoFa, which provide context such as news titles or parent comments, we incorporated this context into models' inputs. For BERT and NLI models, we appended the context to the target comment. For LLMs, we used placeholders in the prompt to include this information. Further implementation details are available in Appendix F.

### 5.2 Results of Fallacy Detection

**BERT-based models fine-tuned on CoCoLoFa had better generalizability than when fine-tuned on Reddit.** Table 4 shows the detection task results. BERT fine-tuned on CoCoLoFa achieved the highest F1 score (0.86) on its test set and showed better generalizability compared to when fine-tuned on Reddit. It surpassed BERT fine-tuned on Reddit in Logic and LogicClimate. On the Reddit dataset, it scored only 0.05 `F1` points lower than BERT fine-tuned on Reddit (0.63 vs. 0.68), but on CoCoLoFa, BERT fine-tuned on Reddit scored 0.13 `F1` points lower (0.73 vs. 0.86).

State-of-the-art LLMs still showed strong performance, achieving the best F1 on Reddit and the best recall on Logic. Notably, LLMs performed poorly on LogicClimate, where fallacious sentences were extracted from context. This might suggest that contextual understanding is crucial for LLM predictions, indicating a need for further research.

### 5.3 Results of Fallacy Classification

**BERT-based models fine-tuned on CoCoLoFa had better generalizability, with classification seeming to be easier than detection.** Table 5 shows the classification results, which are similar to those of the detection task. The NLI model—a BERT-based model—fine-tuned on CoCoLoFa, achieved the highest F1 score (0.87) on its test set. Both BERT and NLI models fine-tuned on CoCoLoFa exhibited better generalizability than those fine-tuned on Reddit. When tested on the Reddit dataset, BERT and NLI models fine-tuned on CoCoLoFa scored only 0.19 and 0.09 F1 points lower, respectively, than their Reddit-tuned

[6] We excluded Gemma(7B) due to its poor performance.

| Model | Train On / Prompt | LOGIC | CLIMATE | Reddit | | | CoCoLoFa | | |
|---|---|---|---|---|---|---|---|---|---|
| | | R | R | P | R | F | P | R | F |
| BERT | Reddit | 51 | 83 | **66** | 69 | 68 | 62 | 89 | 73 |
| | CoCoLoFa | 64 | 83 | 61 | 64 | 63 | **83** | 89 | **86** |
| NLI | Reddit | 67 | **91** | 63 | 80 | 70 | 62 | **96** | 75 |
| | CoCoLoFa | 52 | 52 | 63 | 50 | 56 | 82 | 86 | 84 |
| GPT-4o | zero-shot | 86 | 37 | 59 | **90** | 71 | 72 | 88 | 79 |
| | few-shot | 64 | 25 | 63 | 87 | **73** | 72 | 79 | 75 |
| | COT | **88** | 56 | 64 | 81 | 72 | 76 | 82 | 79 |
| Llama3 | zero-shot | 41 | 8 | 53 | 27 | 36 | 76 | 43 | 55 |
| | few-shot | 79 | 65 | 51 | 89 | 65 | 62 | 95 | 75 |
| | COT | 65 | 28 | 61 | 53 | 56 | 77 | 56 | 65 |

Table 4: The result of fallacy detection task. For LOGIC and LOGICCLIMATE (CLIMATE), we reported the `Recall` rate as they only have positive samples. While for others, we reported `Precision`, `Recall`, and `F1` score. The highest (second-highest) scores are set in **bold** (underlined).

| Model | Train On / Prompt | Reddit | | | CoCoLoFa | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F | P | R | F |
| BERT | Reddit | 71 | 70 | 70 | 65 | 64 | 62 |
| | CoCoLoFa | 65 | 51 | 51 | 85 | 86 | 86 |
| NLI | Reddit | 70 | 72 | 70 | 70 | 67 | 66 |
| | CoCoLoFa | 66 | 62 | 61 | **87** | **87** | **87** |
| GPT-4o | zero-shot | 80 | 76 | 76 | 82 | 80 | 79 |
| | few-shot | 78 | 75 | 75 | 84 | 84 | 83 |
| | COT | **81** | **81** | **81** | 85 | 85 | 85 |
| Llama3 | zero-shot | 58 | 41 | 40 | 57 | 42 | 41 |
| | few-shot | 52 | 33 | 32 | 57 | 50 | 48 |
| | COT | 56 | 48 | 47 | 63 | 58 | 58 |

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. Noted that the `F1` scores we reported were `macro F1` across all fallacy types. The highest (second-highest) scores are set in **bold** (underlined).

counterparts. Conversely, on CoCoLoFa, Reddit-tuned BERT and NLI models scored 0.24 and 0.21 F1 points lower, respectively, than those tuned on CoCoLoFa. Additionally, LLMs, particularly GPT-4o, performed best on the Reddit dataset. We also observed that classification tasks generally performed better than detection tasks, indicating that determining the type of fallacy in a comment might be easier than deciding whether a fallacy exists.

### 5.4 Results of Fallacy Detection in the Wild

The primary motivation for this project is to identify logical fallacies *in the wild* (Ruiz-Dolz and Lawrence, 2023). Therefore, we additionally tested our models on the New York Times Comments Dataset (Kesarwani, 2018). We sampled 500 comments and hired an expert (one in Section 4) to label the fallacies. Table 6 shows the results of fallacy detection on this dataset. The expert annotating the NYT comments identified several fallacies beyond the eight predefined types, so we report two sets of results for each model: one where comments with additional fallacy types are treated as fallacious (positive samples), and another where they are considered non-fallacious (negative samples).

**Detecting fallacies in real-world settings is still challenging.** Although LLMs outperformed all fine-tuned models, their low `F1` score of 0.34 in the second setting (*i.e.*, negative) indicates that LLMs are still unreliable in precisely identifying logical fallacies, motivating the need for further research.

The results also show that BERT-based models fine-tuned on CoCoLoFa outperformed those fine-tuned on Reddit in most cases except for the Recall on NLI models, suggesting CoCoLoFa's potential in training more generalizable models. Additional experimental results on the NYT dataset can be found in Appendix G.

## 6 Discussion

Throughout the project, we learned that annotators often disagree when labeling logical fallacies, as consistently shown by the low inter-annotator agreement reported in all related literature (Sahai et al., 2021; Alhindi et al., 2022), including our own. This section outlines the three main sources of disagreement we identified and offers design suggestions for mitigating (or retaining) them.

### 6.1 Sources of Disagreement

**Complexity in Defining Logical Fallacies.** Many fallacies are similar or overlap, with a single text potentially presenting multiple fallacies. Furthermore, different datasets can provide inconsistent definitions for the same fallacy name. For example, "appeal to authority" might be defined as either "mention of false authority" or "referral to a valid authority without supporting evidence", adding to the confusion (Alhindi et al., 2022). Additionally, when asking experts to annotate the NYT dataset, they identified many comments that embodied other types of fallacy, such as ad hominem, even though they were outside the eight types of

| Model | Train On / Prompt | P | R | F |
|---|---|---|---|---|
| BERT | Reddit | 39 / 15 | 65 / 58 | 49 / 23 |
| | CoCoLoFa | 45 / 18 | 65 / 64 | 53 / 28 |
| NLI | Reddit | 41 / 16 | 82 / 79 | 55 / 27 |
| | CoCoLoFa | 49 / 18 | 62 / 57 | 55 / 28 |
| GPT-4o | zero-shot | 52 / 21 | 75 / 84 | **61** / **34** |
| | few-shot | **54** / 21 | 47 / 48 | 50 / 29 |
| | COT | 47 / 19 | 84 / **87** | 61 / 31 |
| Llama3 | zero-shot | 45 / **22** | **91** / 64 | 60 / 33 |
| | few-shot | 43 / 16 | 87 / 87 | 58 / 28 |
| | COT | 48 / 20 | 80 / 68 | 60 / 30 |

Table 6: The result of fallacy detection on 500 NYT samples. The left/right numbers are the scores where other types of fallacy were considered as positive/negative. Models trained on CoCoLoFa outperform those trained on Reddit. The highest (second-highest) scores are set in **bold** (underlined).

fallacies we predefined in our annotation interface. These fallacies have inherently vague boundaries. For example, ad hominem fallacies are difficult to classify as they require distinguishing between personal attacks aimed at undermining an argument and simple insults. These complexities suggest that fallacy labeling efforts can benefit from standardized definitions and allowing multiple labels per item to capture nuanced perspectives.

**Variability in Annotators' Judgments of Fallacies.** In our study, one expert consistently identified more fallacies than the other, highlighting that annotators can differ significantly in their interpretations of rhetorical strategies. For instance, both experts identified an "appeal to authority" in a comment on abortion legality, which stated: *"The majority's voice should be the guiding light for lawmakers. That's what democracy is about."* However, one expert considered this a valid rhetorical usage, not a fallacy, explaining that it was used to define "democracy" within the text, while the other expert simply labeled it as a fallacy. Requiring annotators to provide rationales may clarify their reasoning for classifying texts as fallacious.

**Divergence Between Writer Intent and Reader Perception.** Despite instructions for workers to write comments with a specific fallacy, annotators sometimes identified different fallacies. This highlights the challenge of aligning readers' interpretations with writers' intentions. It also raises a question: who determines whether a text contains a fallacy and what type of fallacy it represents—the writer, the reader, or an external party? These discrepancies may stem from the nature of fallacies, which can be based on words, sentences, or complex reasoning within the broader context (Bonial et al., 2022), as readers and writers may focus on different elements within the same comments.

### 6.2 Design Suggestions

We propose three design suggestions for future projects involving human labeling of logical fallacies in text: *(i)* provide **clear, operationalized instructions**, *(ii)* implement a **multi-class labeling scheme** that allows a text instance to contain multiple fallacies, and *(iii)* collect **rationales for each fallacy label**, ensuring that if an instance is labeled with multiple fallacies, each one is supported by a distinct rationale. Prior works have adopted some of these approaches. For *(i)*, Ruiz-Dolz and Lawrence suggested using critical questions, such as *"How well supported by evidence is the allegation made in the character attack premise?"*, to validate whether a text contains a fallacy. For *(ii)*, the Climate dataset employed multi-label annotation (Jin et al., 2022). For *(iii)*, Sahai et al. had annotators answer specific questions for each fallacy label. While these approaches have been individually explored in prior studies, we recommend combining *all* three to create a more comprehensive and robust annotated dataset. The project that most closely aligns with this approach is by Helwe et al., which annotated 200 text instances using a unified multi-label scheme. They noted, however, that such detailed annotation is very resource-intensive, as some annotators took four hours to label 20 items. We suspect some of our suggestions may also be costly to scale. More research is needed to explore the trade-offs between data quality and scalability.

## 7 Conclusion and Future Work

This paper presents CoCoLoFa, the largest known logical fallacy dataset, curated through a collaboration between LLMs and crowd workers. BERT-based models fine-tuned using CoCoLoFa achieved good performances in fallacy detection and classification tasks. In the future, we plan to develop models that use context and reasoning to identify fallacies, especially on out-of-distribution data. Additionally, while CoCoLoFa includes eight fallacy types, over a hundred exist. We aim to expand it to cover more.

## 8 Limitations

Like most crowdsourced datasets, CoCoLoFa inherits the biases of using online crowdsourcing platforms to collect data. For example, the crowd workers on Amazon Mechanical Turk are not necessarily representative of the user populations on social media and news platforms; they may prioritize different topics and hold opinions that differ from those of typical online users. In addition, the writing style of commenting in the crowdsourcing task may also differ from that of 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. Apart from the content, the master's qualification we required crowd workers to have may lower the demographic diversity (Loepp and Kelly, 2020), leading to a further risk of bias.

Besides, we integrated GPT-4 into our platform to assist crowd workers in writing high-quality comments. However, we acknowledge that GPT-4 may have a preferred stance (*e.g.*, North American attitudes) when generating example arguments. Although we forced workers to provide input and included that input in the prompt to guide the generation, the biases in GPT-4 may still exist and negatively affect the human written comments.

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.

## Acknowledgement

We thank Meta Research for their support of this work, and Jack Urbanek and Pratik Ringshia for their technical assistance with the Mephisto framework. We are also grateful to the two experts recruited via UpWork for data labeling and the crowd workers from Amazon Mechanical Turk for dataset creation. Special thanks to the anonymous reviewers for their valuable feedback and to Phakphum Artkaew and Reuben Woojin Lee for their help during the early stages of the project.

## References

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

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*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3.

Claire Bonial, Austin Blodgett, Taylor Hudson, Stephanie M. Lukin, Jeffrey Micher, Douglas Summers-Stay, Peter Sutor, and Clare Voss. 2022. The search for agreement on logical fallacy annotation of an infodemic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

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)*.

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*.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. *Proceedings of the AAAI Conference on Artificial Intelligence*.

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*.

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)*.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. MAFALDA: A benchmark and comprehensive study of fallacy detection and classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

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*.

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*, Abu Dhabi, United Arab Emirates.

Aashita Kesarwani. 2018. New york times comments dataset.

Yanda Li, Dixuan Wang, Jiaqing Liang, Guochao Jiang, Qianyu He, Yanghua Xiao, and Deqing Yang. 2024. Reason from fallacy: Enhancing large language models' logical reasoning through logical fallacy understanding. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3053–3066, Mexico City, Mexico. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*.

Eric Loepp and Jarrod T. Kelly. 2020. Distinction without a difference? an assessment of mturk worker types. *Research & Politics*.

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

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

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*.

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)*, Online.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27.

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*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*.

## 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 7.

## B Details of Fallacy Types

### B.1 Eight Chosen Fallacies

We draw the definition and example of the chosen fallacies from Logically Fallacious.[7]

**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.

[7]Logically Fallacious: https://www.logicallyfallacious.com/

| 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 7: Top 10 words of the selected topics

**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.

**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.

### B.2 Shorten Version of Fallacy Definitions

- Appeal to authority: Using an expert of dubious credentials or using only one opinion to promote a product or idea.
- Appeal to majority: A proposition is claimed to be true or good solely because a majority or many people believe it to be so.
- Appeal to tradition: A conclusion supported solely because it has long been held to be true.
- Appeal to nature: Judgment is based solely on whether the subject of judgment is "natural" or "unnatural."
- Appeal to worse problems: Dismissing an argument or complaint due to what are perceived to be more important problems.
- False dilemma: Two alternative statements are given as the only possible options when, in reality, there are more.
- Hasty generalization: Basing a broad conclusion on a small or unrepresentative sample.
- Slippery slope: Asserting that a proposed, relatively small, first action will inevitably lead to a chain of related events resulting in a significant and negative event and, therefore, should not be permitted.

## C GPT-4 Prompts

For the few-shot prompt, we manually select 4 samples from the Reddit and CoCoLoFa dataset as the example data and write the explanation for them as the demonstrative output. For the Chain-of-Thought prompt, we ask LLMs to first answer several questions w.r.t. logical fallacy, then use the answers to determine the presence and the type of a logical fallacy in the input.

**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 Detection (Zero-shot).**

> 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]: `[comment]`.

**Prompt for Detection (Few-shot).**

> 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].
>
> Here are some examples:
>
> [TITLE]: `[title 1]` [PARENT COMMENT]: `[parent comment 1]` [COMMENT]: `[comment 1]` [OUTPUT]: `[label 1]` [EXPLANATIONS]: `[explanation 1]`
>
> [...]
>
> [TITLE]: `[title 4]` [PARENT COMMENT]: `[parent comment 4]` [COMMENT]: `[comment 4]` [OUTPUT]: `[label 4]` [EXPLANATIONS]: `[explanation 4]`
>
> [TITLE]: `[title]` [PARENT COMMENT]: `[parent comment]` [COMMENT]: `[comment]`

**Prompt for Detection (COT).**

> 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.'
>
> Let's think step by step. First, answer these questions:
>
> - What are the key indicators of a logical fallacy?
> - How is reasoning affected by a logical fallacy?
> - In sentences with logical fallacies, are there any common patterns?
> - How does the context of a sentence affect the presence of a logical fallacy?
>
> Then, use the answers to these questions to determine the presence of a logical fallacy in the given [COMMENT]. The output format should

be [YES/NO/UNKNOWN] [EXPLANATIONS]

[TITLE]: `[title]` [PARENT COMMENT]: `[parent comment]` [COMMENT]: `[comment]`

**Prompt for Classification (Zero-shot).**

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]`

**Prompt for Classification (Few-shot).**

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.

Here are some examples:

[TITLE]: `[title 1]` [PARENT COMMENT]: `[parent comment 1]` [COMMENT]: `[comment 1]` [OUTPUT]: `[label 1]` [EXPLANATIONS]: `[explanation 1]`

[...]

[TITLE]: `[title 6]` [PARENT COMMENT]: `[parent comment 6]` [COMMENT]: `[comment 6]` [OUTPUT]: `[label 6]` [EXPLANATIONS]: `[explanation 6]`

[COMMENT]: `[comment]`

[LOGICAL_FALLACY]" `[fallacy]`

[TITLE]: `[title]`

[PARENT_COMMENT]: `[parent]`

**Prompt for Classification (COT).**

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.

Let's think step by step. First, answer these questions:

- What are the differences between fallacies in the [LOGICAL_FALLACY] list?
- For each fallacy type, are there any common patterns in the fallacious sentence?

Then, use the answers to these questions to determine the type of logical fallacy in the given [COMMENT].

[COMMENT]: `[comment]`

[LOGICAL_FALLACY]" `[fallacy]`

[TITLE]: `[title]`

[PARENT_COMMENT]: `[parent]`

## D Data Diversity

**CoCoLoFa covers diverse topics.** Table 8 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).

| Topic | Train | Dev | Test |
|---|---|---|---|
| Protest | 2.9% | 3.1% | 3.0% |
| International Relations | 11.5% | 12.4% | 11.9% |
| Race Issue | 4.9% | 4.7% | 4.5% |
| Women Rights | 9.3% | 10.1% | 10.4% |
| Russo-Ukrainian War | 7.7% | 9.3% | 6.0% |
| Environmental Issue | 8.8% | 10.1% | 7.5% |
| Gender Issue | 3.8% | 3.1% | 4.5% |
| Human Rights | 1.8% | 1.6% | 3.0% |
| Drug Issue | 0.2% | 0.0% | 0.0% |
| Police Brutality | 16.8% | 14.0% | 14.9% |
| Immigration / Refugees | 7.1% | 5.4% | 6.0% |
| COVID / Health Issue | 12.6% | 13.2% | 9.0% |
| Legislation | 6.2% | 7.0% | 6.0% |
| Freedom of Speech | 14.8% | 11.6% | 14.9% |
| Election | 6.2% | 4.7% | 3.0% |
| Sustainability | 5.1% | 4.7% | 6.0% |
| Religious Conflict | 2.0% | 2.3% | 1.5% |
| Political Debates | 4.0% | 3.9% | 4.5% |
| U.S. Politics | 0.2% | 0.0% | 3.0% |
| Digital Rights | 11.5% | 14.0% | 11.9% |
| East Asian Politics | 9.7% | 7.8% | 9.0% |

Table 8: 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.

- **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 9 shows the statistics for each thread structure type in CoCoLoFa. In total, CoCoLoFa had 347 unique thread structures, most of which were of Single Conversation. The diversity of thread structures was high.

## E Annotation Agreement

Figure 4 shows the confusion matrices between experts annotation and labels for both CoCoLoFa and Reddit datasets, as well as the confusion matrix between two experts annotation on the Reddit datasets.

| Type | # Unique Structures | # Articles | Evenness ($J$) |
|---|---|---|---|
| Flat | 5 | 26 | 0.51 |
| Single Conversation | 134 | 312 | 0.93 |
| Multi Conversation | 149 | 246 | 0.96 |
| Complex | 59 | 64 | 0.99 |
| Total | 347 | 648 | 0.95 |

Table 9: Statistics of the thread structure. The 648 comment threads we collected formed 347 unique structures, with the majority falling under the category of 'Multi Conversation'.

## F Experimental Details

We had two different versions of BERT and NLI models. One was fine-tuned on the Reddit dataset, the other was fine-tuned on CoCoLoFa. We fine-tuned them with default hyperparameters set in the original paper, *i.e.*, Sahai et al. (2021) and Jin et al. (2022), respectively. Both models were fine-tuned on a server with an A100 GPU. The training took less than 2 hours for each settings. We ran Llama3 on the same server with Ollama,[8] a package that allows us to run open-weight LLMs with 4-bits quantization on a local server. The inference took 5 to 20 seconds for each instance, depending on the prompt and the input.

## G Additional Results on NYT

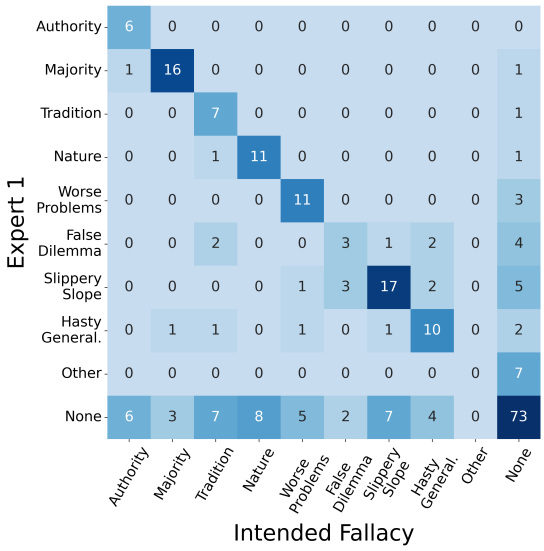
To increase the reliability of the NYT annotation, we hired another expert to annotate 250 NYT comments sampled from the annotation set. The overall Cohen's kappa score between two experts is 0.22, echoing our finding in Sec 4 that it is hard to obtain high IAA in logical fallacy annotation, and that logical fallacy detection in the wild is hard.

Table 10 shows the performance of different models on the 250 samples. We considered both union and intersection labels, where the former one considered a borderline case as fallacy while the latter one considered it as non-fallacy. The result suggests that models fine-tuned on CoCoLoFa generally outperform those trained on Reddit, aligning with the result we showed in Sec 5.4.

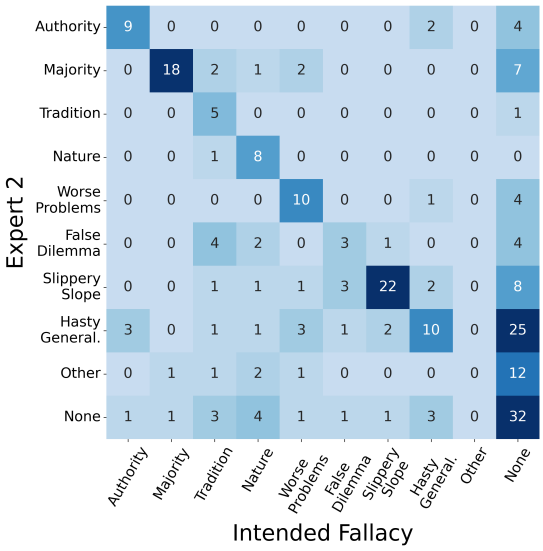
[8] Ollama: https://ollama.com/

| Model | Train On / Prompt | P | R | F |
|---|---|---|---|---|
| BERT | Reddit | 84 / 33 | 66 / 62 | 74 / 43 |
| | CoCoLoFa | 90 / 37 | 58 / 57 | 70 / 45 |
| NLI | Reddit | 81 / 36 | 91 / 95 | 86 / 52 |
| | CoCoLoFa | 88 / 40 | 59 / 63 | 70 / 49 |
| GPT-4o | zero-shot | 92 / 50 | 69 / 95 | 79 / 65 |
| | few-shot | 95 / 53 | 46 / 60 | 62 / 56 |
| | COT | 90 / 40 | 82 / 88 | 86 / 55 |
| Llama3 | zero-shot | 92 / 46 | 53 / 64 | 68 / 54 |
| | few-shot | 83 / 36 | 87 / 89 | 85 / 51 |
| | COT | 86 / 44 | 92 / 72 | 73 / 54 |

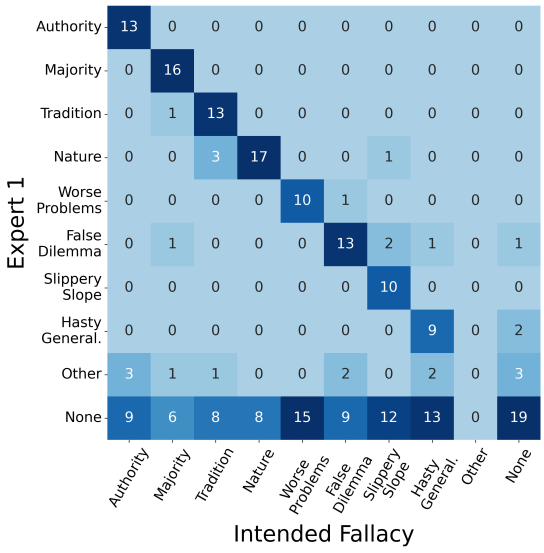
Table 10: The result of fallacy detection on 250 NYT samples labeled by two annotators, aggregated in two ways: union and intersection. The left/right numbers are scores with union/intersection labels, where the former one considered a borderline case as fallacy while the latter one considered it as non-fallacy.



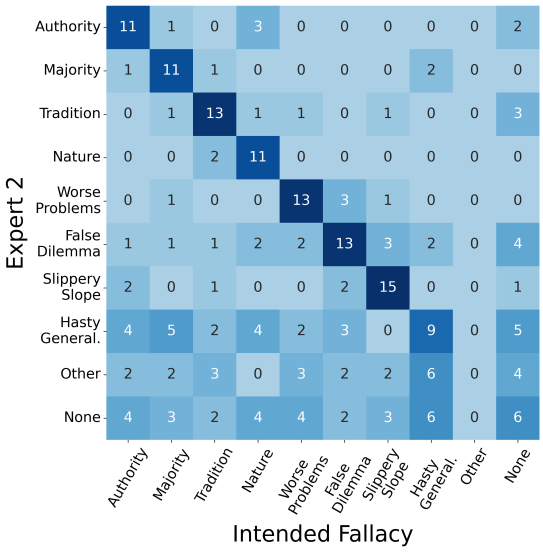

(a) Expert 1 vs. labels (CoCoLoFa).



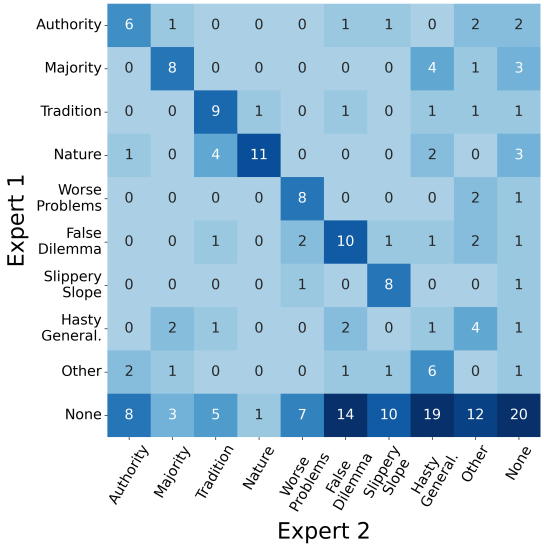

(b) Expert 2 vs. labels (CoCoLoFa).




(c) Expert 1 vs. labels (Reddit).




(d) Expert 2 vs. labels (Reddit).




(e) Expert 1 vs. expert 2 (Reddit).

Figure 4: The confusion matrix of the annotation agreement.