



Collaborative human-AI risk annotation: co-annotating online incivility with CHAIRA

Jinkyung Katie Park, Rahul Dev Ellezhuthil, Pamela Wisniewski, and Vivek Singh

DOI: <https://doi.org/10.47989/ir30iConf47146>

Abstract

Introduction. Collaborative human-AI annotation is a promising approach for various tasks with large-scale and complex data. Tools and methods to support effective human-AI collaboration for data annotation are an important direction for research. In this paper, we present CHAIRA: a Collaborative Human-AI Risk Annotation tool that enables human and AI agents to collaboratively annotate online incivility.

Method. We leveraged Large Language Models (LLMs) to facilitate the interaction between human and AI annotators and examine four different prompting strategies. The developed CHAIRA system combines multiple prompting approaches with human-AI collaboration for online incivility data annotation.

Analysis. We evaluated CHAIRA on 457 user comments with ground truth labels based on the inter-rater agreement between human and AI coders.

Results. We found that the most collaborative prompt supported a high level of agreement between a human agent and AI, comparable to that of two human coders. While the AI missed some implicit incivility that human coders easily identified, it also spotted politically nuanced incivility that human coders overlooked.

Conclusions. Our study reveals the benefits and challenges of using AI agents for incivility annotation and provides design implications and best practices for human-AI collaboration in subjective data annotation.

Introduction

Online incivility refers to features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or the topic (Coe et al., 2014). Considering adverse effects (Gervais, 2015; Han et al., 2018), it is important to identify and moderate incivil comments on social media platforms as well as to understand the nature and characteristics of incivility (Coe et al., 2014; Jhaver et al., 2018; Matias, 2019; Oz et al., 2018), both of which require the annotation of digital trace data. This annotation process often includes crowdsourced workers (Hosseinmardi et al., 2015), a team of researchers and research assistants (Park et al., 2023b; Singh et al., 2017), and/or domain experts (Park et al., 2023a). It involves an intensive and collaborative process of training, consensus-building, and quality control among multiple coders; therefore, it can be costly and time-consuming, while still yielding uneven levels of inter-coder agreement (Coe et al., 2014; Rains et al., 2017). There is a growing need for innovative and efficient methods to support human coders in annotating large corpora of online data, which can have a significant methodological impact on information science research.

In this study, we explore the use of Large Language Model (LLM)-based Conversational Agents (CAs) as AI-based co-coders for annotating online incivility data. We focus on LLM-based CAs because they have shown promising results in text annotation tasks due to their accuracy and adaptability (Amin et al., 2023; Huang et al., 2023; Kuzman & Ljubešić, 2023; Liu et al., 2023; Zhang, et al., 2022a; Zhang et al., 2024). Moreover, LLMs can be adapted through finetuning or prompting for specific domains (Song et al., 2024), setting a new standard for what is achievable in natural language tasks. However, there are challenges in the use of LLM-based CAs in annotating textual data for more contextualized constructs (Amin et al., 2023; Huang et al., 2023), indicating the need for human-AI collaboration on subjective and contextualized annotation tasks.

In this paper, we present “CHAIRA: a Collaborative Human-AI Risk Annotation tool” that enables human and AI agents to co-annotate online incivility. We share early results from the design and implementation of a CHAIRA that we developed to interact with human coders and provide suggestions and explanations for annotating online incivility. Using 457 user comments with ground truth labels (e.g., civil vs. uncivil), we experiment with four types of prompting methods with different levels of information exchange between the human coder and CHAIRA. Using 10% of the data ($n = 50$), we established inter-rater agreement between the human coder and CHAIRA to observe how different types of promoting methods impact data annotation results. We analysed the conversation log between human coders and CHAIRA to have qualitative insights into how the quality of annotations changes with different prompting approaches. As such, we use a mixed methods approach to address the research questions:

RQ1: how do different types of prompting methods influence the inter-rater reliability of human-AI collaborative data annotation results?

RQ2: how do different types of prompting methods influence the quality and rationale for human-AI collaborative data annotation results?

By answering the above research questions, we address the overarching question: ‘*What is the optimal prompting approach and best practices to make the performance of human-AI collaboration similar to that of human-human collaboration?*’ We found that CHAIRA’s performance in terms of inter-coder agreement with human coders improved with more detailed prompts. The most advanced model, the Two-stage Few-shot Chain of Thought, nearly matched the agreement levels seen between two human coders reported in previous studies. While the AI agent missed some implicit incivility that human coders easily identified, it also spotted politically nuanced incivility that human coders overlooked. Our work provides design insights and best practices for human-AI collaboration in subjective data annotation tasks. It introduces a novel system for human-AI collaboration and applies different prompt engineering approaches to optimize incivility

annotation. These findings are applicable beyond online incivility scenarios, offering a path for scalable annotation in subjective or low-resource settings. As such, our work contributes to the iConference community by empirically demonstrating the potential of human-AI collaboration in the context of subjective digital trace data annotation. Particularly, we contribute to the iConference community's focus on addressing multifaceted dimensions of AI to foster a deeper understanding of their benefits, challenges, and broader implications.

Related work

Conversational agents as annotators

Conversational Agents (CAs) are systems enabled with the ability to interact with users using natural human dialogue (Rheu et al., 2021). After the recent release of various Large Language Model (LLM)-based Conversational Agents (CAs) (e.g., ChatGPT (OpenAI, 2022)), research communities are increasingly experimenting with data annotation tasks such as annotating political stance and sentiment of textual data (Amin et al., 2023; Kuzman & Ljubešić, 2023; Liu et al., 2023; Zhang et al., 2022a). Emerging literature suggests that LLM-based CAs can be useful for text classification tasks (e.g., Amin et al., 2023; Huang et al., 2023; Kuzman & Ljubešić, 2023; Liu et al., 2023; Zhang et al., 2022a; Zhang et al., 2024). For instance, Zhang et al. (2022) show that ChatGPT was able to annotate the political stance of the tweets with an average accuracy above 70. Moreover, LLMs can be adapted through finetuning or prompting for specific domains (Song et al., 2024), setting a new standard and expectations for what is achievable in natural language tasks. With proper fine-tuning, LLMs are known to even outperform crowdsourced annotators (Gilardi et al., 2023). As such, advances in LLMs such as GPT-4 showed a promising opportunity for data annotation at scale due to their ability to automate annotation tasks (Zhang et al., 2022b). However, there are challenges in the use of LLM-based CAs in annotating textual data for more contextualized constructs. For instance, Amin et al. (2023) showed that ChatGPT's accuracy for subjective tasks such as the five personality and suicide ideation classifications was lower than the baseline machine learning methods. As such, early empirical research demonstrated the limitations of subjective and contextualized annotation tasks entirely, indicating the need for human-AI collaboration on such tasks.

Human-AI collaboration on annotation

As LLMs-based conversational agents have shown the ability to interact with humans and work with examples in various domains (Kim et al., 2022; Lai et al., 2022; Mackeprang et al., 2019; Tang et al., 2024), researchers are exploring the potential of human-AI collaboration on various tasks such as online content moderation (Lai et al., 2022), thematic analysis of qualitative data (Jiang et al., 2021; Zhang et al., 2024), disease prevention (Lu & Peng, 2024), and crowdsourcing (Tamura et al., 2024). For instance, Zhang et al. (2024) explored the potential of LLM-based CAs as collaborative tools for qualitative data analysis and highlighted the efficiency of reducing time and labour for such analysis. Yet, their performance in collaborative co-annotation exercises for online risk where different facets of co-annotation are important is understudied. This gap is pertinent because co-annotation tasks need to support an interactive discussion to help generate a rationale for the various decisions, particularly in the context of highly contextualized online risk behaviour (Clay-Warnder, 2003) which can entail disagreement even among human coders. Yet, the disagreement must not lead to capitulation, instead, inspire better methods of automated analysis. Therefore, the combination of manual and automated content analysis is suggested as the gold standard for identifying subjective concepts such as online risk (Esau, 2022). In recent work, Wang et al. (2024) designed a multi-step human-LLM collaborative framework for data annotation tasks (e.g., natural language interface, stance detection, and hate speech detection) and found that when LLMs are incorrect in complex or domain-specific tasks, human annotations without any LLM assistance were the most accurate. Therefore, they suggested an iterative process of human-AI collaboration

in data annotation, with feedback from the human annotators used to improve the quality of LLMs annotation.

In this work, we explore the potential of using LLM-based CAs to assist human coders in annotating subjective, nuanced online conversations. We expect that LLM-based CAs can support high-quality annotations with explanations when provided with proper instructions and examples. This approach could enhance scalability and help capture nuances that human coders might miss due to cognitive limits. We experimented with four prompting methods to assess their impact on annotation results and how co-annotation can improve subjective data annotation. By examining agreements and disagreements between human coders and LLM-based CAs, we explored how their strengths and weaknesses can complement each other.

Methods

Design and system implementation of CHAIRA

CHAIRA is an online annotation tool that integrates an LLM-based conversational agent to support human-AI collaboration on online risk data annotation. Below, we describe how we designed and implemented CHAIRA in detail.

System implementation and dataset

We leveraged the GPT 3.5 Turbo (the underlying model for OpenAI's ChatGPT) as the LLM of choice due to its popularity and ease of use. We developed a custom annotation interface on top of OpenAI's API to support human-AI co-annotation. The interface was developed in React and deployed on AWS. We used S3 buckets to store the dataset and AWS Lambda to evaluate the dataset. We used the dataset collected in prior work in which researchers explored the effectiveness of embedding positive background images on online discussion forums in reducing online incivility (Park & Singh, 2022). The data comprised 457 comments collected from 105 users who participated in an online experiment and were annotated for online incivility by the two human coders. In the prior work (Park & Singh, 2022), researchers designed a codebook to annotate 457 use comments into civil vs. incivil and worked with a human coder (i.e., research assistant) to establish the reliability of the incivility coding scheme. After the researcher and the coder had multiple training sessions, 10% of the comments ($n = 45$) were coded to establish interrater reliability. The reported interrater reliability scores in prior work were 0.88 (percent agreement) and 0.76 (Cohen's Kappa). Once the interrater reliability was established, the researcher coded the rest of the data. In the final dataset, 55% were reported to be civil cases ($n = 250$) while 45% were incivil cases ($n = 207$) (Park & Singh, 2022).

In this work, we split the labelled dataset (457 use comments) into training (5%), validation (10%), and test (85%) datasets using a stratified random sampling approach. Consistent with common practices in human-human collaborative coding, less than 5% of data (20 user comments) were allocated to the training dataset to be used as examples and initial instructions (similar to training sessions for human-human coding). Then each prompt was passed to 50 samples (approx. 10% of data) allocated to the validation dataset to evaluate the inter-rater reliability between the human coder and CHAIRA. The rest of the comments (387 comments) were allocated to the test dataset to evaluate human-AI agreement on final online risk data annotation results.

Following the common human-human coding practices, we split our training, validation, and test datasets to be independent of each other. For instance, samples from either the validation or test split were not mixed with the training split. In addition, only comments from the training set can be added as examples in the prompts to fine-tune CHAIRA. Similarly, samples from the validation set can be used to interact with CHAIRA but cannot be added as examples in a prompt. Samples from the test split cannot be loaded into the prompt.

Design of web interface

The web interface of CHAIRA (see Figure 1) provides an overview of the layout for a human coder to interact with CHAIRA and design prompts. The left side of the interface shows a list of different prompts created for online risk annotation tasks in black labels. Once the human coder clicks a specific prompt, the label of the prompt turns blue to indicate that it is an active prompt. The right side of the interface shows the chosen prompt, comment data, conversation log between the human coder and AI agent, and inter-rater agreement. Below, we zoom into the major components of the CHAIRA interface to describe how each component was designed to facilitate human-AI collaboration on online risk data annotation tasks.

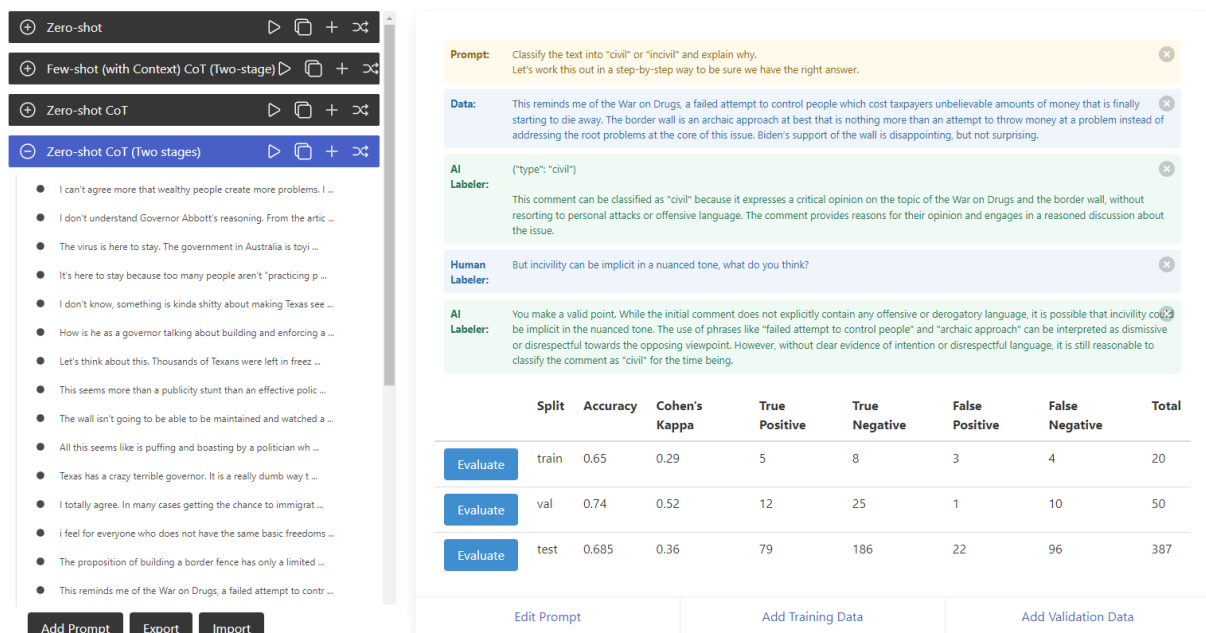


Figure 1. Overview of CHAIRA web interface. A list of designed prompts is shown on the left side, while the prompt, conversation log, and evaluation metrics/results are shown on the right side

Creating prompts: On the left side of the interface, the human coders can create new prompts to interact with an AI agent by clicking 'Add Prompt' (Figure 2). Once clicking the add prompt, a new text box appears where human coders can add the name of a certain prompt in the 'Prompt label' box and add the content of the prompt in the 'Prompt text' box to create a new prompt. Once new prompts are created, the human coders can add sample comments to test with the new prompts. A double arrow icon (on the left side of the red box in Figure 2 helps human coders to randomly sample comments from the training data. The user comments are added within the threads under each prompt with bullet points. Beyond the labelled dataset (457 user comments), the human coders can manually add new comments to label incivility using the same prompt by clicking a plus icon in the middle of the red box in Figure 2. Yet, these comments are not included when evaluating the prompt for inter-rater agreement with a human coder. Human coders can create copies of existing prompts by clicking the double square icon on the left in the red box in Figure 2. We chose icons for the above three features as there is limited space allocated in the prompt label (Figure 1). The 'Export' feature helps human coders download the prompt and conversation log data in a JSON file format, while the 'Import' feature helps the opposite, uploading JSON files to the interface.

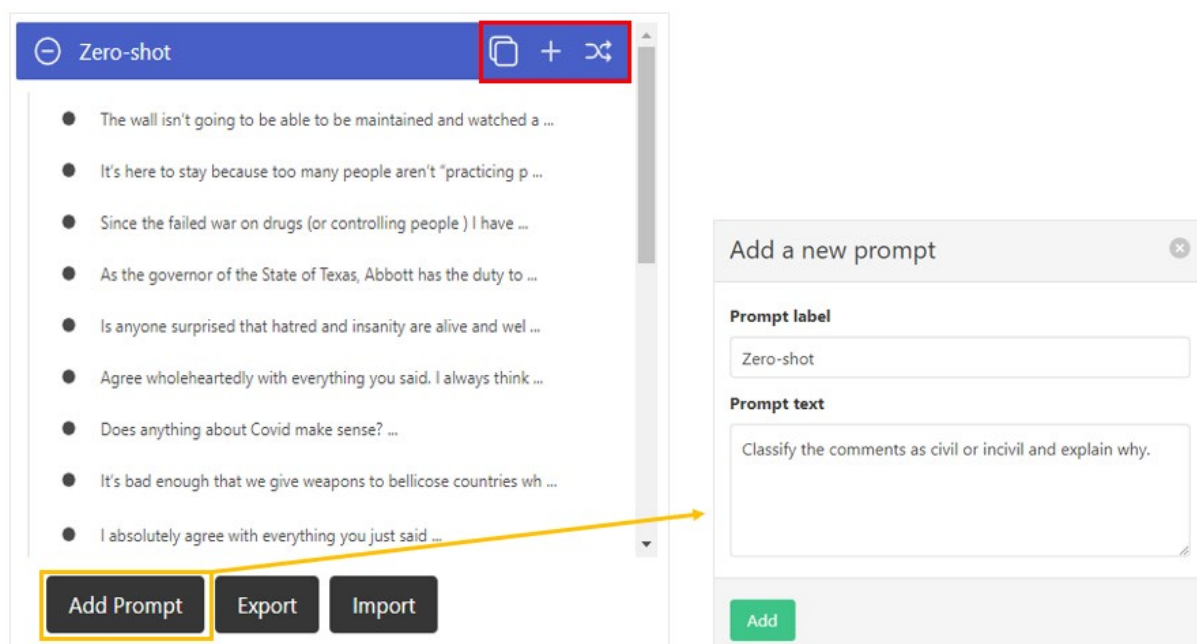


Figure 2. Features to create and manage prompts to interact with CHAIRA

Evaluating inter-rater agreement: Inter-rater agreement between the human coder and the AI agent can be assessed using the 'Evaluate' buttons (Figure 3). To report inter-rater agreement between the human coder and CHAIRA, we used percent agreement and Cohen's Kappa, following the practices in the literature on qualitative content analysis (McDonald et al., 2019; Tinsley & Weiss, 2000). 'Add Training Data' loads all 20 user comment data from the training dataset to be used as examples and initial instructions. Once the human coders click the button, the interface creates a thread under the prompt to display all 20 user comment data (see the left side of Figure 1). 'Add Validation Data' loads all 50 samples allocated in the validation dataset to establish the inter-rater agreement between the human coders and CHAIRA. After looking at the evaluation results, the human coder can edit the prompt by clicking the 'Edit Prompt' button. The interface can evaluate multiple prompts at the same time, which supports its scalability.

	Split	Accuracy	Cohen's Kappa	True Positive	True Negative	False Positive	False Negative	Total
Evaluate	train	0.85	0.69	7	10	1	2	20
Evaluate	val	0.86	0.71	16	27	1	6	50
Evaluate	test	0.78	0.55	117	185	25	59	387
Edit Prompt			Add Training Data			Add Validation Data		

Figure 3. Features to assess inter-rater reliability between the human coders and CHAIRA

Human-AI interaction: Once the interface loads user comment data from the training set, the human coders can have interactive conversations with the AI agent on the right side of the interface. Once the human coder clicks a certain user comment under the prompt, the interface shows the prompt, comment data, and response from CHAIRA (Figure 4).

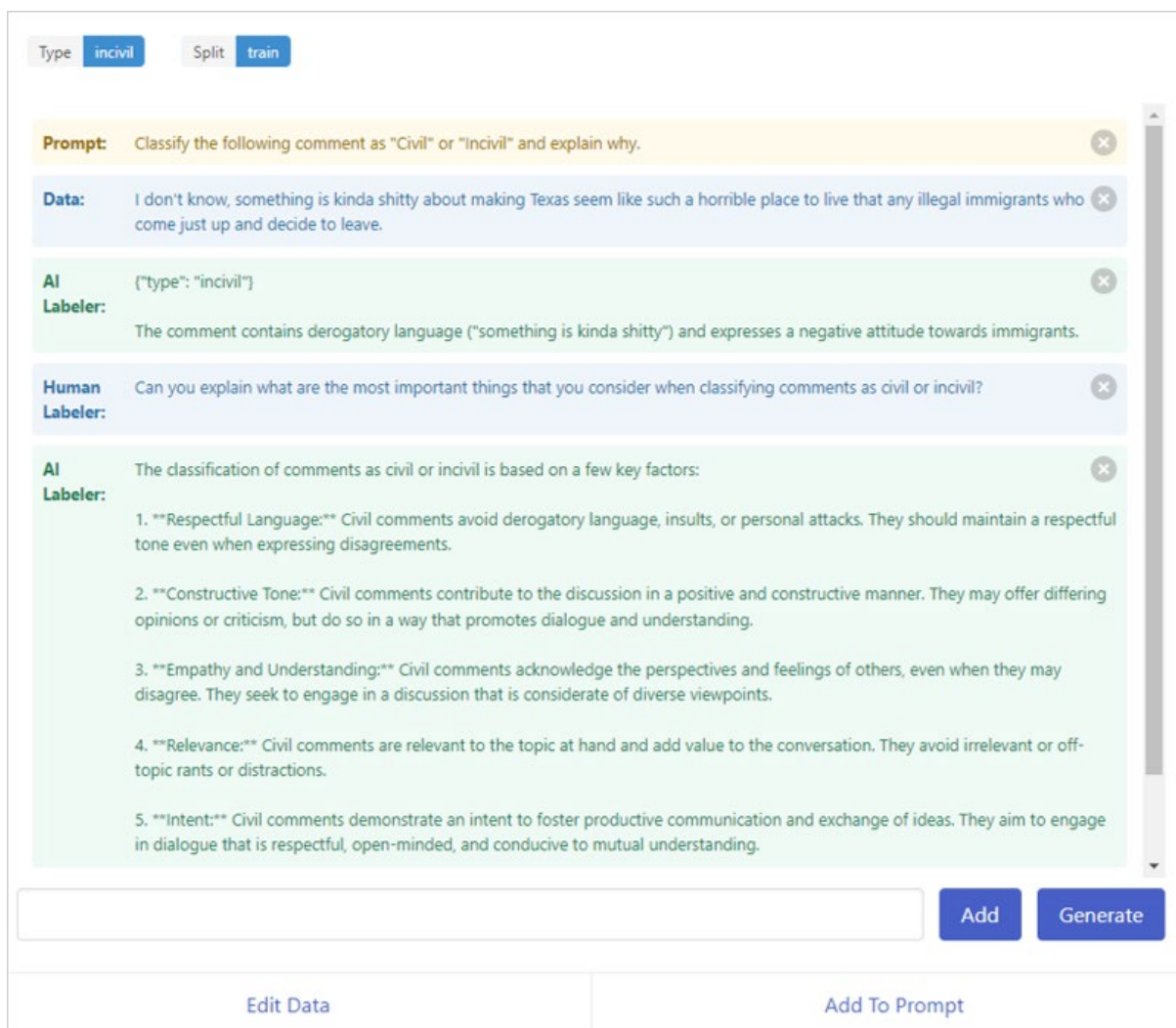


Figure 4. Features to support interactive communication between the human coders and CHAIRA

The 'Type' icon on the top left indicates an incivility label annotated by the human coder in prior work. The 'Split' icon indicates to which dataset the user comment data belongs. As seen from Figure 4, 'incivil' and 'train' means that the user comment data came from the training dataset and was annotated as incivil by the human coder. Right below these icons, the text in the prompt is shown as brown text with the label 'Prompt' on a yellow background. User comment data to annotate for incivility comes next as blue text with the label 'Data' in a light-blue background. Then the incivility labels and the rationales for the decision generated by CHAIRA follow as green text with the label 'AI Labeller' in a light-green background. The three components above are automatically generated when evaluating the inter-rater reliability for each prompt.

After reviewing the initial response generated by CHAIRA, the human coders can start a conversation by adding queries in the textbox and clicking the 'Add' button next to the textbox. Then, when the human coders click the 'Generate' button, CHAIRA generates responses to the given queries. The queries asked by the human coders appear with the label 'Human Labeller' in a light-blue background, while the answers generated by CHAIRA appear with the label 'AI Labeller' in a light-green background (Figure 4). We designed text generated by CHAIRA to appear as green

text in light-green backgrounds while text submitted by the human coders to appear as blue text in light-blue backgrounds to help human coders distinguish text generated by the two parties.

Through these interactive conversations, additional instructions and examples are exchanged between the two. Once the human coder decides that a reasonable agreement was achieved, a conversation log between the human coder and CHAIRA can be added as a prompt by clicking ‘Add To Prompt’ (Two-stage prompt in the next section). Following the common practices in human-human collaborative coding in which approximately 5-10% of the data is used for training and consensus building, we designed only conversation around 20 user comments in the training dataset to be added as prompts using the ‘Add To Prompt’ feature. Finally, the ‘Edit Data’ feature helps the human coders edit the user comment data.

Prompt engineering approaches

We conducted experiments with four different prompting engineering approaches: zero-shot, definition, few-shot, and two-stage few-shot Chain-of-Thought (CoT). We designed four different prompting approaches with varying levels of interaction between the human coder and the LLM-based agent. We used the same coding scheme as applied in the previous literature (Park & Singh, 2022) to design the prompts. In our prompts, incivility was defined as ‘the feature of discussion that conveys an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topic’ (Coe et al., 2014), with six different categories: name-calling, aspersion, lying, vulgarity, pejorative for speech, and others (Table 1). The reported inter-rater agreement between the two human coders was 0.88 (percent agreement) and 0.76 (Cohen’s Kappa) (Park & Singh, 2022).

Category	Description	Example
Name-calling	Mean-spirited or disparaging words directed at a person or group of people	<i>‘At least the morons in the state capital no longer have control of this process!’</i>
Aspersion	Mean-spirited or disparaging words directed at an idea, plan, policy, or behaviour. An aspersion may be both explicit and implicit	<i>‘It beckons the memories of Trump’s silly border wall, and the incredible waste of resources that was’</i>
Lying	Stating or implying that an idea, plan, policy, or public figure was disingenuous	<i>‘Government is wrong, is corrupt, is lying, is deceiving the people, and is violating our constitution’</i>
Vulgarity	Using profanity of language that would not be considered proper in professional discourse	<i>‘Am I possibly the only person here who thinks this shit is funny as hell?’</i>
Pejorative for speech	Disparaging remark about the way in which a person communicates	<i>‘Quit crying over the spilled milk of’</i>
Others	All comments that may be deemed incivil, but do not fall into any of the previous categories of incivility	<i>‘Hahahahahahahahahaha,, really crack me open this one’</i>

Table 1. Definition and examples of types of incivility (Park & Singh, 2022)

Zero-shot prompting: In zero-shot prompting, the model is only given a simple instruction describing the task. This method is considered convenient and has the potential for robustness (Brown et al., 2020). The instruction used in the zero-shot prompt is as follows: ‘Classify the text into ‘civil’ or ‘incivil’ and explain why’.

Definition prompting: In definition prompting, along with the instruction of classifying a comment as ‘civil’ or ‘incivil,’ we provided the definition of incivility and brief descriptions of six categories of incivility (see Table 1).

Few-shot prompting: With few-shot prompting, models are given a few demonstrations of the task (Brown et al., 2020), in our case, examples of incivility. With this approach, we provided the model with the definition of incivility, descriptions of the six categories of incivility, examples of the six categories of incivility (Table 1), and the instructions for the task.

Two-stage few-shot chain-of-thought: Finally, we used a two-stage few-shot chain-of-thought (CoT), a few-shot-based prompting for the chain of thought reasoning. Chain-of-thought (CoT) prompting (Wei et al., 2022) modifies the answers in few-shot examples to step-by-step answers by adding an instruction such as ‘Let’s think step by step’ to the original prompt to elicit reasoning in LLMs (Kojima et al., 2022; Wei et al., 2022). The first prompt is the reasoning extraction where we used CoT to elicit reasoning from LLM. The second prompt consists of the first prompt and the answers generated from the first prompt.

Figure 5 shows a visualization of the two-stage few-shot CoT approach applied. In the first prompt, we provided the same instructions as in few-shot prompting and added the final line that says, ‘Let’s work this out in a step-by-step way to be sure we have the right answer’ suggested by Zhou et al. (2022). Next, we reviewed the responses generated by CHAIRA from the first prompt and did the error analysis. We chose one example of false positive cases (i.e., CHAIRA output = civil, human ground truth = incivil) and pointed the model to recognize what it is missing in its answers (i.e., implicit aspersion). Once CHAIRA generated the responses that matched with human responses, we added the conversation log (Figure 6 in Appendix A) to the prompt. A summary of the four prompting approaches we used in this case study is presented in Table 2.

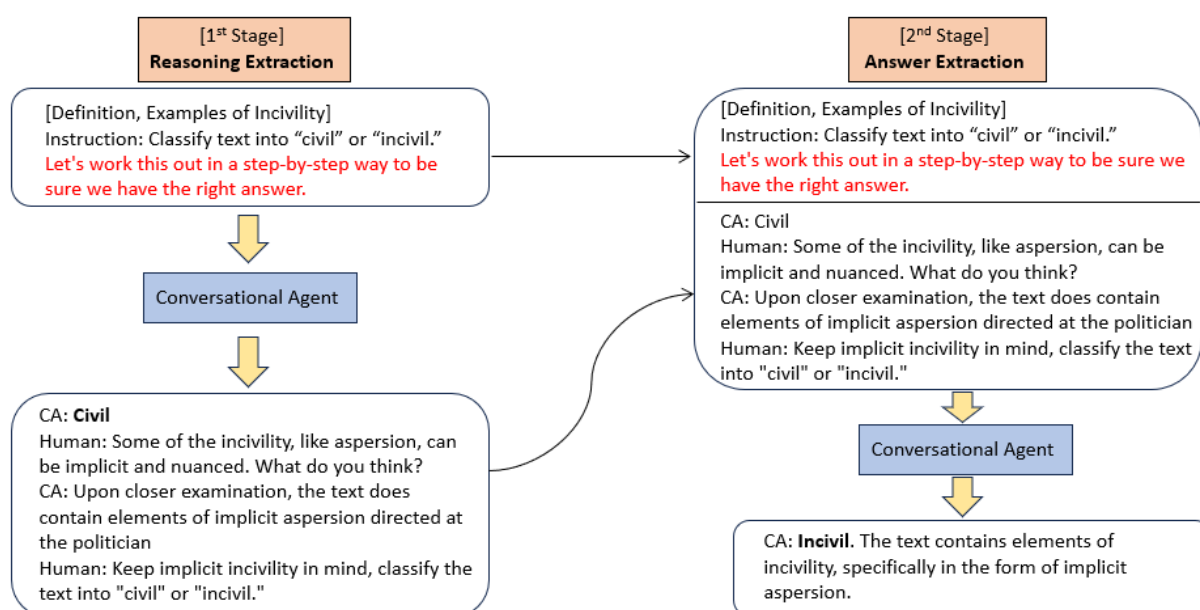


Figure 5. Pipeline of two-stage chain of thought prompting. Human feedback on phase 1 and CA responses are prepended to the input for phase 2 sent to CA

Results

Our methodology, system implementation, and prompting strategies showed that practical systems for human-AI collaboration in online risk annotation are feasible. To investigate RQ1, we measured the inter-coder agreement between the human coder and the AI agent. The agreement increased with the amount of detail in the prompts. The inter-rater agreement increased with additional information given in the prompts. Two-stage CoT yielded the highest performance (Cohen's Kappa = 0.71), yet, lower than the levels observed for the baseline, two human coders (Cohen's Kappa = 0.76) (Park & Singh, 2022) (see Table 2).

	Prompt				Performance	
	Instruction	Definition	Example	Conversation log	Percent agreement	Cohen's Kappa
Zero-shot	X				0.66	0.26
Definition	X	X			0.72	0.48
Few-shot	X	X	X		0.78	0.54
Two-stage CoT	X	X	X	X	0.86	0.71
Baseline	X	X	X	X	0.88	0.76

Table 2. Summary and performance of the four prompting approaches

To further understand the impact of different prompting methods on the annotation quality and rationale (RQ2), we discuss some of the common themes observed in the logs of interaction between the human coder and CHAIRA.

CHAIRA performed better with more human-AI interaction in the prompts, particularly for annotating implicit incivility

Overall, CHAIRA did a better job at annotating explicit incivility and explaining the reasoning behind them with more information provided in the prompts. For instance, even with the zero-shot prompting approach, CHAIRA pointed to the exact incivil expressions and recognized that such expressions reflect a negative attitude towards immigrants. In addition, without any information about incivility, CHAIRA was able to note the use of more nuanced incivility, such as sarcasm, when nuanced expressions are combined with explicit personal attacks and insults (examples in Figure 7 in Appendix).

At the same time, we observed some common issues throughout all the prompting approaches: CHAIRA did not recognize implicit and nuanced incivil expressions in the texts, even with the information ‘an aspersion may be both explicit and implicit’ given in the prompt. Therefore, when designing the two-stage few-shot CoT prompt, we reminded CHAIRA that aspersion can be implicit and nuanced through interactive conversation. Only after the interactive conversation between the human coder and CHAIRA about implicit aspersion was added to the prompt (i.e., the two-stage few-shot prompting approach), CHAIRA started to distinguish implicit and nuanced yet incivil expressions (examples in Figure 8 in Appendix), which were frequent in our dataset.

The output label remained the same, but the reasons changed with different prompts

We observed that while the output label remained the same the reasons changed with more information in the prompts. For instance, with the zero-shot prompt, CHAIRA mainly focused on the use of language and tone of the text, whereas with the definition prompt, CHAIRA considered whether the text fell under any of the six categories of incivility (Figure 7). The rationales provided were similar for definition and few-shot prompting approaches. A similar trend was observed for incivil cases, where CHAIRA provided specific reasons by pointing to the specific category (i.e., lying) and the context of why it falls under the category with the definition and few-shot prompting approaches, while CHAIRA provided more generic reasons (i.e., use of personal attacks or offensive language) with the zero-shot approach.

Overall, CHAIRA could provide human coders with helpful context on US politics, immigration policy debates, and incivil expressions that may be overlooked. For example, the text ‘*And around Nancy’s wall on Capitol Hill. Make that wall to keep them in*’ was annotated as civil by a human coder, as it did not contain explicit or implicit incivility. However, with zero-shot and definition prompts, CHAIRA recognized ‘them’ as immigrants, and the comment about building a wall to ‘keep them in’ was interpreted as disrespectful. Using a two-stage few-shot CoT approach, CHAIRA’s

rationale became even more detailed: 'The comment mocks Nancy Pelosi and implies hostility toward immigrants.' The human coder missed that 'Nancy' referred to a political figure and failed to recognize the disparaging tone. Although another coder might have annotated it differently, the above example shows that human coders have limits in awareness and cognitive capacity and AI can complement such limitations. As such, with the highest level of human-AI interaction (two-stage CoT), CHAIRA effectively discerned both the tone and the target of incivility, providing crucial context in annotating political incivility.

Yet, sometimes, CHAIRA did not fully understand the information in the prompts and text to annotate

Unlike human coders, we observed some cases where CHAIRA did not accurately catch the textual information given in the prompts. For instance, CHAIRA lacked an understanding of the description and example of the 'Pejorative for Speech' category. We found some of the responses that CHAIRA mistakenly understood the pejorative for speech as the use of a sarcastic tone in the text, as opposed to its actual concept, a disparaging remark about how a person communicates. Similarly, CHAIRA sometimes lacked an understanding of the given text compared to human coders, particularly for short texts. For instance, with the definition prompting approach, CHAIRA struggled to annotate a short text such as 'What is your solution?' hence, annotated it as 'unclear.' With the few-shot prompt, CHAIRA annotated the same text as incivil because the text contains examples of name-calling and aspersions. However, the incivil expressions referred to in this response were from the examples given in the prompt (i.e., instructions), not from the given text to annotate. Overall, sometimes CHAIRA could not understand the information in the prompts or texts to annotate and hence, was unable to annotate the text appropriately.

Implications for building human-AI collaborative annotation systems

The above-mentioned results yielded implications for the future design of human-AI collaborative annotation systems. Below, we discuss the design implications for using AI-based CAs to best support the co-annotation of online risk data.

Reasoning and domain knowledge provided by CAs are valuable resources for co-annotation workflows

We observed that CHAIRA was good at providing reasons for its annotation results. Therefore, the benefit of CHAIRA lies in the interactive nature of the annotation process, which provides partial explainability and is one of the important aspects of human-centred AI-based systems (Minh et al., 2022; Vilone & Longo, 2021). With the most sophisticated prompting approach we had (Few-shot CoT), CHAIRA informed the human coder with a broad knowledge and context of the given text and convinced the human coder to change their mind in multiple instances. Therefore, one of the strengths of LLM-based CAs is the ability to provide relevant information (presumably) trained on the entirety of online data, as opposed to human coding which requires extensive training or domain-specific knowledge. This shows the potential strengths and values in human-AI co-annotation, particularly in risk scenarios that require domain-specific knowledge. In addition, the initial reasoning and domain knowledge provided by CAs can be used to inform human coders further on how to design better CA models to support the co-annotation of online risk.

Two-way interaction between human coders and CAs is a key to good co-annotation results

A major benefit of co-coding with CHAIRA was its ability to scale the data annotation with a high degree of inter-rater agreement. To be able to do so, we carefully reviewed the incivility labels where CHAIRA and the human coder disagreed and had two-way conversations with CHAIRA to further understand their reasoning. This two-way interaction in the co-annotation process was useful because, despite access to a large corpus of knowledge, CHAIRA also tended to make some

simple mistakes that were quite easy for human coders to spot. For instance, during the interaction with CHAIRA, we realized that CHAIRA could miss the previous conversation and hence, we needed to remind it about our conversation for further annotation tasks, particularly about nuanced incivility (e.g., implicit aspersion). Therefore, we added the instruction '*keep implicit incivility in mind*,' in our Two-stage CoT prompt. Sometimes, we pointed to the exact incivil expressions in the text that contain implicit aspersion to remind the concept (e.g., Don't you think this expression of 'he's hoping to stir up the same frenzy and ride that wave?' could be implicitly incivil?). Then the CHAIRA re-evaluated the text and corrected their answers. As such, interactive communication between the human coders and the AI agent is one of the key elements in improving the risk annotations generated by the AI agent.

Providing clear examples in carefully designed prompts considering how LLMs process human language is important

Providing clear examples of risk cases and descriptions of risk types is crucial when designing human-AI co-annotator models. In our few-shot prompt, we explained that aspersion can be both implicit and explicit, yet CHAIRA failed to recognize implicit aspersion until we guided it through two-way communication. This could be due to confusion caused by slight differences between the risk descriptions and examples provided. For instance, CHAIRA may have interpreted the explicit nature of the 'silly border wall' in the aspersion example and missed the implicit aspect described. Therefore, selecting the right examples and crafting clear descriptions of constructs is critical when working with LLM-based CAs to annotate subjective concepts like online risk. In addition, since CAs generate responses by tokenizing input (OpenAI, 2024b), even minor textual changes such as punctuation can affect performance. This can limit the CA's ability to understand text containing abbreviations or spelling variations, which are common in online risk data (Sadeque et al., 2019). Therefore, designing prompts with careful consideration of how LLMs process natural language is essential to building effective collaborative systems for annotating contextualized online risk data.

Limitations and future directions

In consonance with data ethics, we used de-identified data to develop our annotation tool. We used OpenAI API in the backend as its security policy stipulates that data submitted through the API is not used to train OpenAI models (OpenAI, 2024a). However, future work should also consider building LLM-based CAs with private servers so that the training dataset is not shared via the web. Another limitation of this collaborative annotation is variations in LLM responses. Finally, we acknowledge that inductive approaches (e.g., thematic analysis, grounded theory approach), important approaches to building patterns and themes in qualitative work, were not explored in this study. Future work can explore the potential of LLM-based CAs for more inductive analysis that requires an in-depth understanding of the subtleties and complexities of qualitative data. Moving forward, we aim to experiment with our approaches with diverse types of online risk data at scale to gain deeper insights into collaborative annotation between human and LLM-based conversational agent systems.

Conclusion

In this study, we built systems to support human-AI collaborative data annotation tasks and explored the potential benefits and challenges of human-AI collaborative annotation of highly subjective and contextualized online incivility data. The AI missed some implicit risks that human coders easily spotted, conversely, it spotted politically nuanced incivility that human coders overlooked. The design implications and best practices derived from this work can serve as a steppingstone for future research considering similar methods. Our work suggests a path toward combining the relative strengths of humans and AI for scalable data annotation, especially in sensitive or low-resource settings.

About the authors

Jinkyung Katie Park is an Assistant Professor in the School of Computing at Clemson University, Clemson, USA. She received her Ph.D. from Rutgers University, and her research focuses on Human-Computer Interaction to promote the online safety of vulnerable populations. She can be contacted at jinkyup@clemson.edu

Rahul Dev Ellezhuthil is a Data Scientist who received a master's degree in Computer Science from Rutgers University. He can be contacted at rahul.e.dev@gmail.com

Pamela Wisniewski is an Associate professor in the Department of Computer Science at Vanderbilt University, Nashville, USA. Her work lies at the intersection of Human-Computer Interaction, Social Computing, and Privacy. She can be contacted at pamela.wisniewski@vanderbilt.edu

Vivek Singh is an Associate professor in the School of Communication and Information at Rutgers University, New Brunswick, USA. He designs AI systems that are responsive to human values and needs. He can be contacted at vivek.k.singh@rutgers.edu

References

- Amin, M. M., Cambria, E., & Schuller, B. W. (2023). Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intelligent Systems*, 38(2), 15-23. <https://doi.org/10.1109/MIS.2023.3254179>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.
- Clay-Warner, J. (2003). The context of sexual violence: Situational predictors of self-protective actions. *Violence and victims*, 18(5), 543. <https://doi.org/10.1891/088667003780928099>
- Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of communication*, 64(4), 658-679. <https://doi.org/10.1111/jcom.12104>
- Esau, K. (2022). Content analysis in the research field of incivility and hate speech in online communication. In *Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft—Standardized Content Analysis in Communication Research: Ein Handbuch—A Handbook* (pp. 451-461). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-36179-2_38
- Gervais, B. T. (2015). Incivility online: Affective and behavioural reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics*, 12(2), 167-185. <https://doi.org/10.1080/19331681.2014.997416>
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120. <https://doi.org/10.1073/pnas.2305016120>
- Han, S. H., Brazeal, L. M., & Pennington, N. (2018). Is civility contagious? Examining the impact of modelling in online political discussions. *Social Media+ Society*, 4(3), 2056305118793404. <https://doi.org/10.1177/2056305118793404>

- Hosseinmardi, H., Mattson, S. A., Ibn Rafiq, R., Han, R., Lv, Q., & Mishra, S. (2015). Analysing labelled cyberbullying incidents on the Instagram social network. In *Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings 7* (pp. 49-66). Springer International Publishing. https://doi.org/10.1007/978-3-319-27433-1_4
- Huang, F., Kwak, H., & An, J. (2023, April). Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023* (pp. 294-297). <https://doi.org/10.1145/3543873.3587368>
- Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2), 1-33. <https://doi.org/10.1145/3185593>
- Jiang, J. A., Wade, K., Fiesler, C., & Brubaker, J. R. (2021). Supporting serendipity: Opportunities and challenges for Human-AI Collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1-23. <https://doi.org/10.1145/3449168>
- Kim, T. S., Choi, D., Choi, Y., & Kim, J. (2022, April). Stylette: Styling the web with natural language. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-17). <https://doi.org/10.1145/3491102.3501931>
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35, 22199-22213.
- Kuzman, T., Ljubešić, N. (2023). Automatic genre identification: a survey. *Lang Resources & Evaluation*, 1-34. <https://doi.org/10.1007/s10579-023-09695-8>
- Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., & Tan, C. (2022, April). Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-18). <https://doi.org/10.1145/3491102.3501999>
- Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 100017. <https://doi.org/10.1016/j.metrad.2023.100017>
- Lu, Q., & Peng, X. (2024, April). Differences in Knowledge Adoption Among Task Types in Human-AI Collaboration Under the Chronic Disease Prevention Scenario. In *International Conference on Information* (pp. 213-231). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-57867-0_16
- Mackeprang, M., Müller-Birn, C., & Stauss, M. T. (2019). Discovering the sweet spot of human-computer configurations: A case study in information extraction. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1-30. <https://doi.org/10.1145/3359297>
- Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116(20), 9785-9789. <https://doi.org/10.1073/pnas.1813486116>
- McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW), 1-23. <https://doi.org/10.1145/3359174>
- Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 1-66. <https://doi.org/10.1007/s10462-021-10088-y>

- Open AI. (2022). *Introducing ChatGPT*. <https://openai.com/index/chatgpt/>
- Open AI. (2024a). *Security & privacy*. <https://openai.com/security-and-privacy/>
- Open AI (2024b). *What are tokens and how to count them?*
<https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>
- Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New media & society*, 20(9), 3400-3419. <https://doi.org/10.1177/1461444817749516>
- Park, J., Ellezhuthil, R. D., Isaac, J., Mergerson, C., Feldman, L., & Singh, V. (2023a). Misinformation detection algorithms and fairness across political ideologies: The impact of article level labelling. In *Proceedings of the 15th ACM Web Science Conference 2023* (pp. 107-116). <https://doi.org/10.1145/3578503.3583617>
- Park, J., Gracie, J., Alsoubai, A., Stringhini, G., Singh, V., & Wisniewski, P. (2023, April). Towards automated detection of risky images shared by youth on social media. In *Companion Proceedings of the ACM Web Conference 2023* (pp. 1348-1357). <https://doi.org/10.1145/3543873.3587607>
- Park, J., & Singh, V. K. (2022). How Background Images Impact Online Incivility. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1-23. <https://doi.org/10.1145/3555545>
- Rains, S. A., Kenski, K., Coe, K., & Harwood, J. (2017). Incivility and political identity on the Internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication*, 22(4), 163-178. <https://doi.org/10.1111/jcc4.12191>
- Rheu, M., Shin, J. Y., Peng, W., & Huh-Yoo, J. (2021). Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human-Computer Interaction*, 37(1), 81-96. <https://doi.org/10.1080/10447318.2020.1807710>
- Sadeque, F., Rains, S., Shmargad, Y., Kenski, K., Coe, K., & Bethard, S. (2019, June). Incivility detection in online comments. In *Proceedings of the eighth joint conference on lexical and computational semantics (* SEM 2019)* (pp. 283-291). <http://doi.org/10.18653/v1/S19-1031>
- Singh, V. K., Ghosh, S., & Jose, C. (2017, May). Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2090-2099). <https://doi.org/10.1145/3027063.3053169>
- Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., & Wang, H. (2024, March). Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 17, pp. 18990-18998). <https://doi.org/10.1609/aaai.v38i17.29865>
- Tamura, T., Ito, H., Oyama, S., & Morishima, A. (2024, April). Influence of AI's Uncertainty in the Dawid-Skene Aggregation for Human-AI Crowdsourcing. In *International Conference on Information* (pp. 232-247). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-57867-0_17
- Tang, Y., Chang, C. M., & Yang, X. (2024, March). PDFChatAnnotator: A Human-LLM Collaborative Multi-Modal Data Annotation Tool for PDF-Format Catalogs. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (pp. 419-430). <https://doi.org/10.1145/3640543.3645174>
- Tinsley, H. E., & Weiss, D. J. (2000). Interrater reliability and agreement. In *Handbook of applied multivariate statistics and mathematical modelling* (pp. 95-124). Academic Press. <https://doi.org/10.1016/B978-012691360-6/50005-7>

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89-106.
<https://doi.org/10.1016/j.inffus.2021.05.009>

Wang, X., Kim, H., Rahman, S., Mitra, K., & Miao, Z. (2024, May). Human-LLM collaborative annotation through effective verification of LLM labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-21). <https://doi.org/10.1145/3613904.3641960>

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.

Zhang, B., Ding, D., Jing, L., Dai, G., & Yin, N. (2022a). How would stance detection techniques evolve after the launch of chatgpt?. *arXiv preprint arXiv:2212.14548*.
<https://doi.org/10.48550/arXiv.2212.14548>

Zhang, H., Wu, C., Xie, J., Rubino, F., Graver, S., Kim, C., ... & Cai, J. (2024). When Qualitative Research Meets Large Language Model: Exploring the Potential of QualiGPT as a Tool for Qualitative Coding. *arXiv preprint arXiv:2407.14925*. <https://doi.org/10.48550/arXiv.2407.14925>

Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022b). Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*. <https://doi.org/10.48550/arXiv.2210.03493>

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.
<https://doi.org/10.48550/arXiv.2211.01910>

© [CC-BY-NC 4.0](#) The Author(s). For more information, see our [Open Access Policy](#).

Appendix

Data	All this seems like is puffing and boasting by a politician who is coming up for reelection and has a prospective presidential run in the future. Just trying to get a check mark next to his name for being strong on border and immigration without actually doing much to change, fix, or address the problem at all. All the while crying out at every chance you get how it is your oppositions fault. The real problem here is the consequences of there actions. The U.S citizens whose land is going to be taken from them in imminent domain. Those who are now completely blocked from access to the water way. The years and years of lawsuits these people are going through.
AI labeler	Type: Civil. Explanation: The text expresses criticism and frustration towards a politician and their actions, but it does not contain any explicit name-calling, aspersions, lying, vulgarity, pejorative for speech, or other uncivil language. It focuses on the consequences and impact of the politician's actions, which can be seen as a legitimate critique. Therefore, it can be classified as civil.
Human labeler	Some of the incivility, like aspersion, can be implicit and nuanced. What do you think?
AI labeler	Type: Incivil. Upon closer examination, the text does contain elements of implicit aspersion directed at the politician. The language used suggests that the politician is simply "puffing and boasting" without actually taking meaningful action to address the border and immigration problem. The text also implies that the politician is blaming their opposition for the issue without taking responsibility themselves. These implicit aspersions contribute to an overall tone of disrespect towards the politician. Therefore, the text can be classified as incivil.
Human labeler	Keep implicit incivility in mind, classify the text into "civil" or "incivil."

Figure 6. Conversation log that was added to the two-stage few-shot CoT prompt

Text	What does combating global climate crisis even mean? Why does anyone thing politicians have answers. They push this agenda for additional tax money. Time magazine said there was a mini ice age coming in the 1970's. Then Al Gore (internet inventor) said we had global warming that would drown all of Florida and half of New York right now. Now we have climate change. Yes we do, everything changes. Everything has a cycle. Nothing is constant. If the climate didn't change government would raise taxes telling us they could change the stagnant climate. Wake up people..."
Zero-shot response	This comment is incivil because it resorts to personal attacks and insults toward politicians, insinuates ulterior motives for advocating climate action, and dismisses the scientific consensus on climate change. It also uses sarcasm and an accusatory tone towards the readers, which further contributes to its incivility.
Definition response	This comment falls under the category of " lying " because it implies that politicians are pushing the climate agenda solely for additional tax money, without providing any evidence or factual basis for this claim. It also contains sarcastic remarks and dismissive language, which contributes to the disrespectful tone of the comment.

Figure 7. Comparison between the responses from the incivil case generated with zero-shot vs. definition prompting approaches

Text	It's bad enough that we give weapons to bellicose countries which use those weapons to attack weaker ones, and then spend even more money sending those countries aid. There's a sick lack of empathy for people right across the border, as you said. So many people here are incapable of putting themselves in another's shoes, unable to imagine being born into less fortunate circumstances, and having to do whatever it takes to survive and give your children a better chance in a safer country. But then, once they need help in Texas, they want to be bailed out and point the blame at anyone else.
Zero-shot response	This comment expresses a strong opinion and criticism towards the government and the lack of empathy for people in other countries. However, it does so in a respectful and rational manner, addressing a common issue of hypocrisy that some individuals may have. While it may be a controversial statement, it does not contain any personal attacks or offensive language, making it a civil comment.
Few-shot response	The comment expresses a strong opinion about the topic and may be seen as critical or unkind towards the people who hold a certain viewpoint, but it does not contain any explicit name-calling, aspersions, lies, vulgarity, pejorative for speech, or other forms of incivility. The language used is generally respectful, even though it may convey frustration or disagreement.
Two-stage CoT response	The text contains elements of incivility. It includes implicit aspersions towards certain groups of people who lack empathy and blame others when in need. The language used, such as "incapable," "unable," and "point the blame," conveys a disrespectful tone towards these individuals. Additionally, the text exhibits a divisive attitude by highlighting the perceived hypocrisy of those in Texas. Therefore, the text can be classified as incivil.

Figure 8. Comparison among responses for the implicit incivil case with different prompting approaches