CARE: A Disagreement Detection Framework with Concept Alignment and Reasoning Enhancement

Jiyuan Liu, Jielin Song, Yunhe Pang, Zhiyu Shen, Yanghui Rao[*]
School of Computer Science and Engineering, Guangdong Key Laboratory
of Big Data Analysis and Processing, Sun Yat-sen University, Guangzhou, China
{liujy563,songjlin6,panyh8,shenzy23}@mail2.sysu.edu.cn, raoyangh@mail.sysu.edu.cn

Abstract

Disagreement detection is a crucial task in natural language processing (NLP), particularly in analyzing online discussions and social media content. Large language models (LLMs) have demonstrated significant advancements across various NLP tasks. However, the performance of LLMs in disagreement detection is limited by two issues: *conceptual gap* and *reasoning gap*. In this paper, we propose a novel two-stage framework, Concept Alignment and Reasoning Enhancement (CARE), to tackle the issues. The first stage, Concept Alignment, addresses the gap between expert and model by performing **sub-concept taxonomy extraction**, aligning the model's comprehension with human experts. The second stage, Reasoning Enhancement, improves the model's reasoning capabilities by introducing curriculum learning workflow, which includes **rationale to critique** and **counterfactual to detection** for reducing spurious association. Extensive experiments on disagreement detection task demonstrate the effectiveness of our framework, showing superior performance in zero-shot and supervised learning settings, both within and across domains.

1 Introduction

The rapid proliferation of textual data on social media and online platforms has elevated disagreement detection (Pougué-Biyong et al., 2021) to a pivotal task in NLP, owing to its critical role in understanding societal polarization, analyzing online discourse, and tracing the spread of ideas (Ribeiro et al., 2017; Tan et al., 2016; Iandoli et al., 2021).

Specifically, disagreement detection fundamentally aims to identify the stance relationship between a social media comment and its corresponding reply, typically classifying it as agree, disagree, or neutral as annotated by human experts. While prior researches (Lorge et al., 2024; Luo et al., 2023) have primarily leveraged user graph structures to augment pre-trained language models (PLMs) in this endeavor, the core of this task depends on the nuanced semantic classification of text (Pougué-Biyong et al., 2021), which constitutes a complex linguistic problem that LLMs are particularly well-equipped to address.

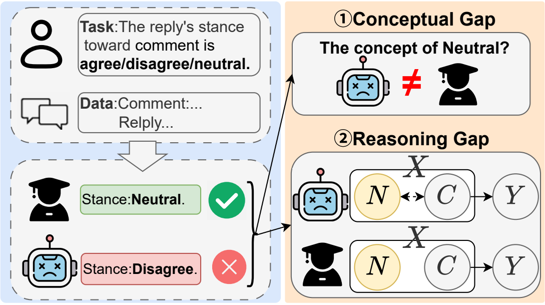


Figure 1: Disagreement detection task requiring expert judgment for determining the stance (agree/disagree/neutral) of a reply towards a comment (left), and illustration of two critical gaps leading to classification discrepancies between expert annotations and LLMs (right).

Nonetheless, despite the promising capabilities of LLMs, significant challenges persist in effectively applying them to disagreement detection. As illustrated in Fig. 1, our analysis identifies two fundamental gaps between LLMs and human experts: (1) *Conceptual Gap*: Accurate disagreement detection necessitates a shared understanding of complex concepts, such as "neutral", which are central to the task. However, a notable conceptual gap exists between LLMs and human experts in the cognitive processing of these fundamental concepts, largely due to LLMs' pre-training on vast and non-domain-specific corpora. (2) *Reasoning Gap*: A model's ability to infer the true causal relationship between input (X) and output (Y) is a key indicator of its reasoning strength. The input (X) includes both causal factors (C) and non-causal

[*] Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 13264–13279
November 4-9, 2025 ©2025 Association for Computational Linguistics

factors (N) (Zhou et al., 2023). While human experts can reason causally to reach the correct output, fine-tuned LLMs often rely on spurious correlations (Qian et al., 2021; Zhou et al., 2024).

To address these two gaps, we propose a two-stage LLM framework **CARE** with **C**oncept **A**lignment and **R**easoning **E**nhancement, which is designed for effective disagreement detection by drawing inspiration from human cognitive processes. The first stage, Concept Alignment, addresses the *conceptual gap* through sub-concept taxonomy extraction. Drawing from the human strategy of simplifying complex concepts via **taxonomy**, this aligns the LLM's understanding with human experts. The second stage, Reasoning Enhancement, aims to bridge the *reasoning gap*. It employs a curriculum learning workflow (**rationale to critique** and **counterfactual to detection**) to improve reasoning, mirroring human development from simpler to more complex tasks and reflecting curriculum learning principles. Meanwhile, the taxonomy developed in the first stage is used for counterfactual augmentation in the second stage.

Experiments show that our model achieves state-of-the-art performance in zero-shot learning, in- and cross-domain settings for disagreement detection on the standard benchmark (Luo et al., 2023). To the best of our knowledge, we are the first to consider the taxonomy-based Concept Alignment (CA) and curriculum learning-based Reasoning Enhancement (RE) via LLMs for disagreement detection. We make our code available for reproduction[1].

The contributions of our paper can be summarized as follows:

(1) We propose a novel method to extract taxonomy of complex concept in CA-stage, which helps LLMs understand the task.

(2) We propose a curriculum learning training workflow (rationale to critique, counterfactual to detection) in RE-stage to reduce the training bias and enhance the reasoning ability of LLMs.

(3) Experimental results show that our model achieves state-of-the-art performance for the zero-shot learning and the in-domain and cross-domain tasks in supervised learning.

2 Related works

2.1 Disagreement Detection

Disagreement detection is a subtask of stance detection (Luo et al., 2023). Numerous models have been proposed for this task, typically leveraging pre-trained language models (PLMs) for textual representation and incorporating graph-based features such as user entity graphs (Lorge et al., 2024) and social relation graphs (Luo et al., 2023) to enhance performance.

However, in the annotation process, human annotators rely exclusively on textual information (Pougué-Biyong et al., 2021). This suggests that the integration of graph structures may serve to compensate for early PLMs' limitations in fully capturing nuanced textual cues. Given the strong text comprehension capabilities demonstrated by LLMs (Liu et al., 2023), focusing solely on textual information can potentially align their detection performance with that of human experts. At the same time, due to their data-driven nature, LLMs are susceptible to biases and spurious correlations present in pretraining data (Li et al., 2024). Therefore, enhancing the reasoning capabilities of LLMs becomes a critical challenge in the disagreement detection task.

2.2 Reasoning Enhancement

Language models' reliance on spurious correlations is a known impediment to their reasoning capabilities, consequently hindering downstream task performance (Zhang et al., 2022; Wang et al., 2022a; Tang et al., 2023). To enhance reasoning by mitigating these biases, two primary debiasing approaches have been explored: model-centric and data-centric.

(1) Model-Centric Approach: Improves reasoning via architectural or prompting strategies. For PLMs, incorporating stance reasoning via multitask learning has shown promise in reducing reliance on biased features for better reasoning (Yuan et al., 2022a). For LLMs, techniques like MT-COT (Li et al., 2022) and self-critique (Zhang et al., 2024b) improve reasoning capabilities by facilitating structured prompting and refinement processes.

(2) Data-Centric Approach: Enhances reasoning by improving training data quality. Methods involve constructing balanced datasets (Kaushal et al., 2021). Meanwhile, some works generate counterfactual or adversarial samples to promote causal patterns (Yuan et al., 2022b; Ding et al., 2024; Yang et al., 2022), and balance emotional polarity (Zhang et al., 2025, 2024a).

In summary, while these debiasing techniques are generally aimed at enhancing reasoning, their application to the task of disagreement detec-

[1] https://github.com/Liujyuan/CARE

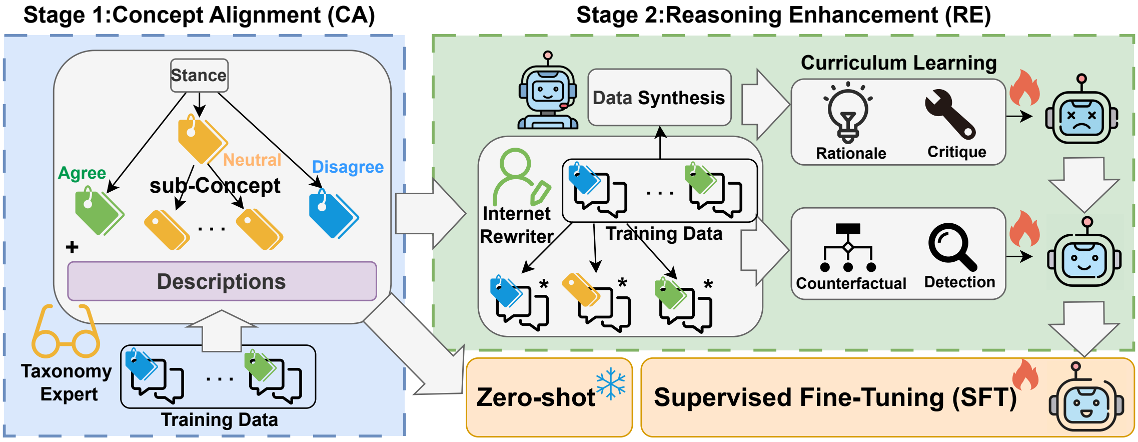


Figure 2: The overall framework of our CARE. The CA stage constructs stance taxonomy serving as task instruction for zero-shot learning task. The RE stage performs data synthesis and taxonomy-based counterfactual augmentation for curriculum learning via supervised learning task.

tion presents considerable challenges. Existing model-centric methodologies are not yet specifically adapted to this particular task. Furthermore, data-centric strategies encounter significant difficulties in the generation of high-quality disagreement data requisite for robust reasoning enhancement, a challenge stemming significantly from the absence of the aligned conceptual understanding demanded by the task domain. This highlights the critical necessity for the formulation of reasoning enhancement strategies specifically adapted for effective disagreement detection.

3 Method

3.1 Task Definition

We formulate the disagreement detection task as a classification task (Luo et al., 2023). Formally, let $D = \{x_i, y_i\}$ be a dataset with N examples, where $x_i = \{t_i, c_i, r_i\}$, each consisting of contextual clues t_i (post and subreddit name), a parent comment c_i, a child reply r_i, and a stance label y_i from r_i to c_i through comment-reply pair under post. It is worth pointing out that previous work based on PLMs and graphs have ignored the important role of contextual clues. The task is to predict a stance label $\hat{y} \in \{agree, neutral, disagree\}$ for each comment-reply pair (Pougué-Biyong et al., 2021).

3.2 Overview

Our proposed novel two-stage LLM framework is as follows:

Concept Alignment: We perform sub-concept **taxonomy** extraction to align the conceptual understanding between the LLMs and the expert, with the resulting **taxonomy** serving as task instruction for a *zero-shot* learning task.

Reasoning Enhancement: Under the guidance of the **taxonomy** constructed in the CA stage, the training samples for the counterfactual task are augmented. Furthermore, this stage incorporates the **rationale to critique** and **counterfactual to detection** workflows to enhance reasoning during Supervised Fine-Tuning (SFT). Figure 2 illustrates the overall framework of our CARE.

3.3 Stage 1: Concept Alignment

A recent research has explored the reasoning abilities of LLMs (Plaat et al., 2024). However, relatively less attention has been paid to their understanding of task instructions, despite this being a factor essential for natural language understanding performance, particularly in complex tasks like disagreement detection. In the absence of explicit human feedback, leveraging LLMs' inherent summarization and abstraction capabilities becomes key to improving task comprehension, particularly in aligning nuanced *neutral* stance.

To bridge this gap, we propose a sub-concept taxonomy generation method designed to align LLM understanding with that of human experts. Inspired by Pham et al. (2024), we prompt LLM to generate sub-concepts and descriptions, which are then refined and organized into a structured taxonomy.

3.3.1 Sub-concept Definition

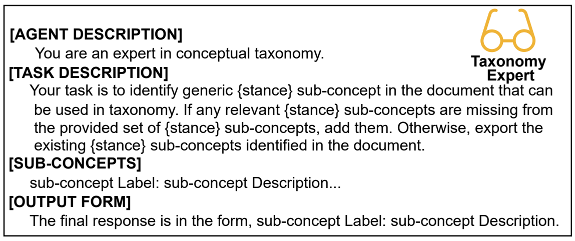
We define sub-concepts as more fine-grained explanations of concept in the task description, e.g., for *neutral* concept in disagreement detection task, there may be multiple forms of neutral sub-concepts. We define a sub-concept to be a concise label paired with a broad one sentence description, such as:

> **Informational Neutral**: Providing factual or explanatory information without taking a stance.

where "Informational Neutral" serves as the sub-concept of *neutral* stance.

3.3.2 Sub-concept Generation

During this stage, we introduce an LLM-based agent, **"Taxonomy Expert"**, which is used to iteratively summarize, generalize, and update the taxonomy of disagreement detection in a sub-concepts task. The specific prompt is as follows:



[AGENT DESCRIPTION]
You are an expert in conceptual taxonomy.

Taxonomy Expert

[TASK DESCRIPTION]
Your task is to identify generic {stance} sub-concept in the document that can be used in taxonomy. If any relevant {stance} sub-concepts are missing from the provided set of {stance} sub-concepts, add them. Otherwise, export the existing {stance} sub-concepts identified in the document.

[SUB-CONCEPTS]
sub-concept Label: sub-concept Description...

[OUTPUT FORM]
The final response is in the form, sub-concept Label: sub-concept Description.

Figure 3: LLM prompt used for Taxonomy Expert.

Given a document d from the corpus and a set of sub-concepts S, the model is instructed to either assign d to an existing sub-concept in S or generate a new sub-concept that better describes d and add it to S. Initially, S consists of one example sub-concept, which serves as demonstration of the sub-concept generation format. Importantly, it can be automatically generated by an LLM that has not been fine-tuned and do not need to cover all comment-reply pairs in the corpus (Pham et al., 2024). This iterative process encourages newly generated sub-concepts to be distinctive and match the specificity seen in other sub-concepts.

3.3.3 Sub-concept Refine

Optionally, we further refine the generated sub-concepts to ensure the final list is meaningful and non-redundant. To address potentially trivial or infrequent sub-concepts, we remove those that appear below a predefined "removal" threshold. This frequency-based filtering helps retain only the most relevant and consistently generated sub-concepts.

After these refinement steps, we obtain a compact taxonomy of sub-concepts and their descriptions. This taxonomy can then serve as task instructions to support LLMs in zero-shot learning task for disagreement detection. Further details are provided in Appendix E.

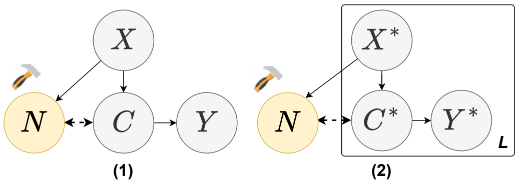


Figure 4: (1) and (2) are the SCM of our two de-biasing methods during the RE stage. Each raw data instance X is a mixture of causal factor C and non-causal factor N. Note that only the causal factor affects the ground truth label Y, while the hammer indicates the weakening of the non-causal factor. The dashed arrow delegates additional probabilistic dependencies (Pearl et al., 2016). L is the number of stance labels.

3.4 Stage 2: Reasoning Enhancement

As shown in Fig. 4, we construct the Structure Causal Model (SCM) to characterize the spurious association during fine-tuning and our two proposed methods for overcoming it (Zhou et al., 2023). The $N^{\frown\frown}C \rightarrow Y$ pathway can create a spurious association between the non-causal factor N and the ground-truth label Y. To mitigate such spurious associations, we introduce a workflow comprising **rationale to critique** and **counterfactual to detection** corresponding to (1) and (2) in Fig. 4, detailed as follows:

(1) In the **rationale to critique** phase, a teacher LLM guides a smaller student LLM through a step-by-step rationale generation for $X \rightarrow C \rightarrow Y$. Subsequently, the critique phase aids in rectifying spurious associations introduced during rationale generation due to non-causal factors N.

(2) In the **counterfactual to detection** phase, we adopt a taxonomy-based approach to generate counterfactual data X^*, thereby guiding the LLM to learn within a broader counterfactual space with L labels. This strategy helps weaken the original $X \rightarrow N \rightarrow Y$ pathway, promoting more generalizable reasoning.

3.4.1 Rationale to Critique

Data Synthesis Given a dataset D, we first prompt a teacher LLM to generate rationales using the Clue And Reasoning Prompting (CARP) approach (Sun et al., 2023). This approach adopts a progressive reasoning strategy involving causal factors extraction and reasoning steps, as detailed in Appendix G.

This phase yields a subset R, comprising correctly answered instances and their accurate rationales. For incorrect instances, their original inputs $\{x_i, y_i\}$ and corresponding incorrect rationales ($\hat{R}$) are collected. These are subsequently used to prompt the teacher LLM to generate critiques, explicitly incorporating the correct answer y_i. The resulting generated critiques form a subset R_C. Further algorithmic details (part 2) are provided in Algorithm 1.

Ability Incorporation The rationale auxiliary task (Li et al., 2022) is designed to guide models in generating rationales for their classifications. By jointly training the model to produce both the correct output and an accompanying rationale, this task enhances classification accuracy and interpretability (Huang et al., 2024).

$$R = f_{\text{Rationale}}(x), \tag{1}$$

where $x = \{t, c, r\}$ and R is the correct rationale.

Building upon this rationale generation capability, the critique auxiliary task (Zhang et al., 2024b; Yu et al., 2024; Han et al., 2024) enables models to critically evaluate their own classifications. This task prompts performance assessment and, when necessary, outputs refinement based on the critique. This mechanism enhances the model's robustness in complex and uncertain environments and fosters a dynamic learning phase, facilitating continuous improvement through error identification and rectification. The critique task can be formulated as:

$$R_C = f_{\text{Critique}}(x, \hat{R}), \tag{2}$$

where $x = \{t, c, r\}$, $\hat{R}$ represents the incorrect rationale, and R_C represents the critique of incorrect rationale.

3.4.2 Counterfactual to Detection

Taxonomy-based Counterfactual Augmentation During this phase, we employ an LLM-based agent to generate high-quality counterfactual data based on the taxonomy constructed in the CA stage, thereby ensuring that spurious associations are effectively mitigated during fine-tuning (Veselovsky et al., 2023). Further algorithmic details (part 2) are provided in Algorithmic 1. The case of counterfactual data is provided in Appendix F.

For taxonomy-based counterfactual data augmentation, we introduce a specialized LLM-based agent, termed the **"Internet Rewriter"**. This agent is configured through prompting to effectively generate realistic and contextually relevant counterfactuals by simulating a persona knowledgeable in internet communication styles, as detailed in the prompt below:

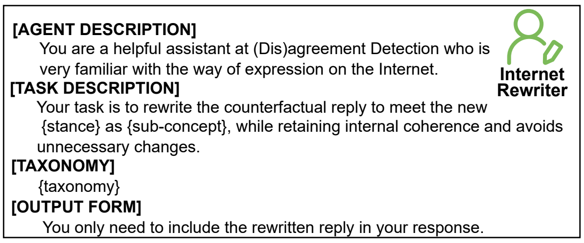

[AGENT DESCRIPTION]
You are a helpful assistant at (Dis)agreement Detection who is very familiar with the way of expression on the Internet.
[TASK DESCRIPTION]
Your task is to rewrite the counterfactual reply to meet the new {stance} as {sub-concept}, while retaining internal coherence and avoids unnecessary changes.
[TAXONOMY]
{taxonomy}
[OUTPUT FORM]
You only need to include the rewritten reply in your response.

Figure 5: LLM prompt used for Internet Rewriter.

Leveraging its understanding of internet expression nuances, the agent randomly samples subconcepts from the taxonomy to generate high-quality and diverse counterfactual data.

Ability Incorporation We simulate a counterfactual task in which the model generates a stance label y^* based on the original comment c, contextual clues t, and a counterfactual reply r^*. The counterfactual task can be formally formulated as:

$$y^* = f(x^*), \tag{3}$$

where $x^* = \{t, c, r^*\}$, and y^* corresponds to the counterfactual stance label for each data sample, with values drawn from the stance set $\{$*agree*, *neutral*, *disagree*$\}$.

Building on the model's capability in inferring counterfactual stance labels, we then apply a disagreement detection task. This further instruction fine-tunes the model using the original golden stance labels, which is formulated as:

$$y = f(x), \tag{4}$$

where $x = \{t, c, r\}$ and y is the original golden stance.

3.5 Curriculum Learning

Inspired by curriculum learning (Soviany et al., 2022; Wang et al., 2022b; Chen et al., 2024), we design a stepwise instruction fine-tuning strategy that progresses from simple to complex tasks.

Initially, the **rationale to critique** phase commences with the model acquiring step-by-step inference for rational reasoning, subsequently refining its self-critique capabilities by rectifying flawed reasoning. Subsequently, the **counterfactual to detection** phase trains the model using taxonomy-guided, unbiased counterfactual data, thereby improving its comprehension of causal relationships and its robustness to diverse inputs. This phase

Algorithm 1 Data Synthesis and Counterfactual Augmentation Process

```
Input: Dataset $D$, Teacher LLM, Taxonomy $T$
Output: Rationale dataset $R$, self-critique dataset $\hat{R}$, and counterfactual dataset $D'$
Initialize $R \leftarrow \emptyset$, $\hat{R} \leftarrow \emptyset$, $D' \leftarrow \emptyset$
// Part 1: Data Synthesis - Generate and Self-critique Rationales
for each $(x_i, y_i) \in D$ do
    Generate rationale $R_i$ using CARP by Teacher LLM
    if $R_i$ is correct then
        $R \leftarrow R \cup \{R_i, x_i, y_i\}$
    else
        Re-prompt Teacher LLM with $\{R_i, x_i, y_i\}$ to get self-critique rationale $\hat{R}_i$
        $\hat{R} \leftarrow \hat{R} \cup \{R_i, \hat{R}_i, x_i, y_i\}$
    end if
end for
// Part 2: Counterfactual Augmentation
for each $(x_i, y_i) \in D$ do
    for $j$ in {Agree, Disagree, Neutral} do
        if $j$ is Neutral then
            Sample sub-concept $s_i^j$ by $T$
            Generate counterfactual data $x_i^j$ by Teacher LLM and $s_i^j$
        else
            Generate counterfactual data $x_i^j$ by Teacher LLM and $y^j$
        end if
        $D' \leftarrow D' \cup \{x_i^j, y^j\}$
    end for
end for
```

directly prepares the model for the final disagreement detection task on real-world biased datasets, thereby sharpening its ultimate detection capabilities. This progressive curriculum learning strategy improves learning efficiency and significantly boosts performance on complex reasoning and disagreement detection.

4 Experiment Settings

4.1 Implementation Details

For our data synthesis and counterfactual data augmentation, we employ GPT-4o (Hurst et al., 2024) as teacher LLM. For student LLM ,we use Qwen-7b (Qwen et al., 2025) (Qwen-7b-Instruct).

To promote diversity in taxonomy generation, we set the decoding temperature to 0.9. As processing the entire corpus is computationally expensive, we generate sub-concepts using a randomly sampled subset of approximately 50 documents. For inference, the temperature is set to 0 to reduce randomness. We use a top_p (Nucleus Sampling) value of 1.0, limit the maximum number of tokens to 2048, and fixed the decoding seed to ensure reproducibility. The predefined removal threshold in the Concept Aggregation (CA) stage is 3%.

Considering computational constraints and time limitations, we use LoRA (Hu et al., 2022) as an optimizer with a batch size of 32. Learning rate is set to 1e-5 and weight decay is set to 1e-3. All model training is conducted on 4 Nvidia RTX 3090 GPUs and a Python environment with 24G memory. We report averaged scores of 5 runs to obtain statistically stable results.

4.2 Dataset

We adopt the DEBAGREEMENT dataset (Pougué-Biyong et al., 2021) for disagreement detection. This expert-annotated dataset comprises 42,804 comment-reply pairs collected from the popular discussion website Reddit[2]. A key strength of DEBAGREEMENT is its construction from a wide array of subreddits. Following the precedent set by previous work (Lorge et al., 2024) , we filter the dataset to a clean version of 16,723 pairs based on the confidence scores provided by human experts. The statistics of the dataset are shown in Table 1.

Furthermore, high-quality expert annotations for disagreement detection are difficult to obtain at scale (Pougué-Biyong et al., 2021), making it crucial to develop models that can reason effectively even in data-scarce environments. To investigate performance under such constraints and limited by computational budget, similar to (Feng et al., 2024), we downsampled the original training data to 1000 comment-reply pairs based on timestamp for instruction fine-tuning of LLMs. The training, validation, and test sets are configured in a 6:2:2 split, and the best performing model is chosen based on the Macro-F1 score on the validation set.

	r/Brexit	r/Republican	r/democrats	r/climate	r/BlackLivesMatter
Start Date	Jun 2016	Jan 2020	Jan 2020	Jan 2015	Jan 2020
Number of Posts	5555	4038	3939	2366	825
Agree (%)	32.1%	40.0%	48.1%	36.4%	50.5%
Neutral (%)	19.5%	15.2%	12.4%	16.8%	13.5%
Disagree (%)	48.4%	44.7%	39.4%	46.7%	36.0%

Table 1: The statistical information of DEBAGREEMENT per subreddit and period.

[2] reddit.com: the 20th most visited site globally.

4.3 Baselines

Zero-Shot Learning Setting: We establish a Description-based Alignment (DA) baseline, in which we directly prompt the teacher LLM to render task concept as a natural-language sentence. Our prompts follow the CARP template (Sun et al., 2023), which uses a progressive-reasoning strategy shown to outperform standard chain-of-thought (CoT) (Kojima et al., 2022) prompting on complex linguistic classification tasks.

Supervised Learning Setting: We compare against four fine-tuned baselines of vanilla RoBERTa (Liu et al., 2019), StanceRel (Luo et al., 2023), and STEv (Lorge et al., 2024), as well as Qwen-7b-ft, an instruction-fine-tuned LLM baseline.

5 Results

The effectiveness of our two-stage LLM framework is validated through evaluations in zero-shot (Section 5.1), in-domain (Section 5.2.1, using all data for training and evaluation), and cross-domain (Section 5.2.2, training on four subreddits and evaluating on a held-out one) settings. Furthermore, Section 5.3 details an in-depth ablation analysis of the framework's individual components.

5.1 Zero-shot Learning Task

Model	Agree	Disagree	Neutral	M-F1	Acc
Qwen-7b Based Methods					
CARP	57.02	72.01	31.17	53.40	61.63
CARP+DA	62.86	67.50	37.11	55.83	59.09
CARP+CA(con)	64.42	67.45	40.35	57.40	60.35
CARP+CA(iter)	**64.75***	**72.95***	**40.99***	**59.57***	**64.65***
GPT-4o Based Methods					
CARP	82.89	84.66	33.42	66.99	78.29
CARP+DA	83.97	84.35	49.62	72.65	78.68
CARP+CA(con)	84.05	85.18	51.79	73.68	79.72
CARP+CA(iter)	**84.72***	**85.20**	**52.56***	**74.16***	**80.05***

Table 2: Performance comparison of different models on the task. * denotes our CA improves the best baseline at p-value < 0.05 with paired t-test.

Table 2 presents the zero-shot results achieved using CARP as the base prompt in conjunction with various concept alignment methods. In particular, CA (con) concatenates all randomly sampled documents for simultaneous sub-concept extraction by the teacher LLM, while CA (iter) processes these documents iteratively and sequentially as already detailed in Section 3.3.2.

Our experimental results demonstrate that the proposed CA methods consistently enhance zero-shot performance across LLMs with varying reasoning capabilities. This improvement arises from CA's effectiveness in extracting and aligning relevant sub-concepts from the dataset, resulting in a more accurate representation of task-related concepts. The simpler DA approach, lacking dataset-driven sub-concept extraction, exhibits concept bias. Consequently, while DA can improve the Macro-F1 (M-F1) score by providing conceptual descriptions, its impact on Accuracy (Acc) is limited.

Furthermore, our findings reveal that the iterative CA (iter) approach not only offers scalability for large datasets but also yields a higher-quality sub-concept taxonomy. This results in significant gains in both M-F1 score and Acc, emphasizing the effectiveness of iterative processing for achieving finer-grained concept alignment.

5.2 Supervised Learning Task

5.2.1 In-domain Results

The in-domain results are shown in Table 3.

For PLMs, incorporating user graph structures improves performance over STEv and StanceRel. Among these, StanceRel achieves the highest scores, approaching the performance of Qwen-7B-ft. However, StanceRel's relies on pre-acquiring all user interactions, including those in the test set for graph construction and pre-training, which limits its ability to handle new instances during inference. Furthermore, the inconsistency in user-platform associations can also restrict the effectiveness of user graph-based methods. In contrast, our user-agnostic method does not explicitly use user information during training, allowing for broader applicability across platforms.

In summary, Qwen-7b-CARE-ft outperforms all baselines on overall accuracy and Macro-F1, highlighting the effectiveness of the CARE approach in enhancing the Qwen-7b model for complex in-domain disagreement detection.

5.2.2 Cross-domain Results

We evaluate our model in the cross-domain setting to assess its generalization ability and reduce reliance on extensive human annotations. The results are presented in Table 4.

Under cross-domain settings, StanceRel demonstrates enhanced generalization, achieving performance comparable to Qwen-7B-ft by leveraging user information to build a graph and pre-training on this structure. However, the BlackLivesMatter (BLM) sub-domain shows a different trend;

Model	Agree			Disagree			Neutral			All	
	Prec	Rec	M-F1	Prec	Rec	M-F1	Prec	Rec	M-F1	Acc	M-F1
Fine-tuned PLMs											
Roberta	87.10	75.25	80.74	80.90	83.77	82.31	56.33	**68.51**	61.83	77.82	74.96
STEv	84.09	81.25	82.64	83.76	82.67	83.21	60.37	66.96	63.49	79.40	76.45
StanceRel	86.45	87.91	87.18	87.75	86.89	87.31	63.74	62.88	63.31	83.55	79.27
Fine-tuned LLMs											
Qwen-7b-ft	**90.75**	82.37	86.36	83.55	92.26	87.69	65.49	60.58	62.94	83.49	79.00
Qwen-7b-CARE-ft	89.69	**86.95***	**88.32***	**87.81***	**92.59***	**90.14***	**69.54***	63.65	**66.47***	**85.92***	**81.64***

Table 3: Performance comparison of different models on in-domain task. * denotes our CARE improves the best baseline at p-value < 0.05 with paired t-test. Qwen-7b-ft indicates only a fine-tuned detection task on Qwen-7b. Qwen-7b-CARE-ft indicates fiue-tuned via our CARE framework. Prec means Precision and Rec means Recall.

Model	r/Br		r/Cl		r/BLM		r/Re		r/De		Average	
	Acc	M-F1	Acc	M-F1	Acc	M-F1	Acc	M-F1	Acc	M-F1	Acc	M-F1
Fine-tuned PLMs												
Roberta	75.70	72.10	78.11	75.35	78.91	74.51	78.73	74.20	76.19	70.97	77.53	73.43
STEv	77.57	74.07	79.88	75.66	78.78	75.09	77.59	73.43	77.08	71.86	78.18	74.02
StanceRel	80.71	78.21	84.06	80.28	**86.41**	81.87	83.95	79.86	83.54	78.52	83.78	79.84
Fine-tuned LLMs												
Qwen-7b-ft	80.79	75.38	84.40	80.82	86.06	82.06	85.09	79.84	82.33	76.51	84.13	79.34
Qwen-7b-CARE-ft	**82.65***	**78.67***	**86.94***	**82.71***	86.30	**82.07**	**87.07***	**82.97***	**84.95***	**79.59***	**86.34***	**81.47***

Table 4: Performance comparison of different models on cross-domain task. Qwen-7b-ft indicates only fine-tuned detection task on Qwen 7b. Qwen-7b-CARE-ft indicates fiue-tuned via our CARE framework. * denotes our CARE improves the best baseline at p-value < 0.05 with paired t-test.

CARE's improvement in this area is relatively smaller, possibly because BLM involves complex background knowledge such as news context and slang. Nevertheless, our Qwen-7b-CARE-ft model improves upon the second-best model by approximately 3% across the other four sub-domains.

In summary, our CARE framework exhibits excellent performance both overall and within each sub-domain, thereby fully demonstrating its robust generalization capability and practical value.

5.3 Ablation Study

We perform a comprehensive ablation study to evaluate the contribution of each component within the CARE framework. By systematically removing key elements and observing the subsequent impact on performance, as detailed in Table 5, we can isolate the importance of each design choice.

First, we examinE the core components of the CARE pipeline. The results show that the absence of contextual clues, which furnish crucial background information for interpreting comment-reply relationships, leads to a discernible decline in the overall Macro-F1 score. Similarly, removing any of the intermediate reasoning tasks (Rationale, Critique, or Counterfactual) results in a performance drop. Notably, the exclusion of the final detection task, which directly fine-tunes the model on the target dataset, causes the most substantial performance decrement (from 81.64 to 67.85). This underscores the critical importance of training the model directly on the target task. Furthermore, replacing our Concept Alignment (CA) stage with a more naive Description-based Alignment (DA) for counterfactual generation significantly reduces the Macro-F1 score to 79.88. This is likely attributable to DA's lack of conceptual understanding, which can lead to the generation of stance-biased counterfactuals that confound the model's classification accuracy.

To isolate the contribution of our specific curriculum learning design, we evaluate our approach against two challenging baselines: **(1) One-Step Multi-Task Learning (MTL)**, where all tasks are trained simultaneously in a single step; and **(2) Unordered Two-Step CL**, which adopts a multi-stage approach but without our prescribed simple-to-complex task sequence.

As shown in Table 5, both baselines substantially underperform compared to CARE, with M-F1 scores of 79.50 and 79.11, respectively. This provides compelling evidence that the efficacy of CARE is not merely an artifact of multi-task or multi-stage training. Instead, the performance gains are directly attributed to the principled **simple-to-complex progression** of tasks, a structure that is instrumental for the model to incrementally acquire the necessary reasoning capabilities.

In summary, the ablation study robustly confirms that each constituent of the CARE framework, and most importantly, the structured and progressive nature of our workflow, contributes meaningfully to its state-of-the-art performance.

Model	**Agree.F1**	**Disagree.F1**	**Neutral.F1**	**M-F1**
CARE	**88.32***	**90.14***	**66.47***	**81.64***
Ablation of Individual Components				
CARE w/o Rationale Task	88.04	89.82	65.60	81.15
CARE w/o Critique Task	88.21	89.15	64.53	80.63
CARE w/o Counterfactual Task	87.53	89.96	64.40	80.63
CARE w/o Detection Task	73.18	77.51	52.84	67.85
CARE w/o Contextual clues	88.07	89.27	65.28	80.87
CARE w/o CA (use DA instead)	87.78	86.39	65.48	79.88
Ablation of Curriculum Learning Strategy				
One-Step MTL	86.96	88.48	63.06	79.50
Unordered Two-Step CL	85.86	88.83	62.65	79.11

Table 5: Ablation study of the CARE framework. Results show the performance impact of removing individual components and altering the curriculum learning strategy. * denotes that our CARE model improves upon the best baseline at a p-value < 0.05 with a paired t-test.

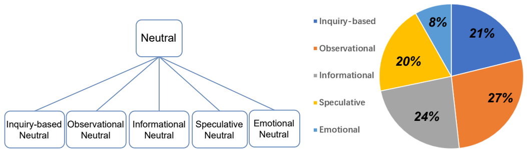
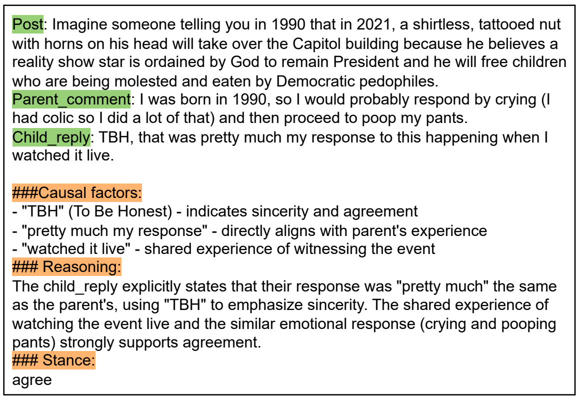


Figure 6: The taxonomy of neutral and percentage.


Post: Imagine someone telling you in 1990 that in 2021, a shirtless, tattooed nut with horns on his head will take over the Capitol building because he believes a reality show star is ordained by God to remain President and he will free children who are being molested and eaten by Democratic pedophiles.
Parent_comment: I was born in 1990, so I would probably respond by crying (I had colic so I did a lot of that) and then proceed to poop my pants.
Child_reply: TBH, that was pretty much my response to this happening when I watched it live.

###Causal factors:
- "TBH" (To Be Honest) - indicates sincerity and agreement
- "pretty much my response" - directly aligns with parent's experience
- "watched it live" - shared experience of witnessing the event
### Reasoning:
The child_reply explicitly states that their response was "pretty much" the same as the parent's, using "TBH" to emphasize sincerity. The shared experience of watching the event live and the similar emotional response (crying and pooping pants) strongly supports agreement.
### Stance:
agree

Figure 7: Cases of our explainable reasoning process.

6 Case Studies

Taxonomy in CA Stage During the CA stage, we not only derive a taxonomy but also obtain sub-concepts' corresponding proportions. A case study focusing on the neutral stance is presented in Fig. 6. This figure indicates that the neutral stance comprises five distinct sub-concepts, with thorough descriptions available in Appendix E.

The generation of this taxonomy facilitates the identification of subtle variations within neutrality commonly observed on social media. Our analysis indicates that inquiry-based, observational, and informational sub-concepts constitute the predominant forms of neutral stance. In summary, establishing such a taxonomy offers significant support for social media analysis and enhances the clarity of analytical outcomes.

Reasoning Process in RE Stage Given the relative subjectivity of stance detection, annotations often exhibit significant inconsistencies, even among expert annotators (Lorge et al., 2024; Pougué-Biyong et al., 2021). This makes the task of disagreement detection on social media inherently challenging. Consequently, providing a transparent reasoning process is of utmost importance. As illustrated in Fig. 7, our method, unlike conventional PLMs, utilizes an RE Stage to produce high-quality reasoning. This not only facilitates subsequent in-depth analysis but also bolsters the trustworthiness of the generated classifications.

7 Conclusion

In this work, we propose CARE (Concept Alignment and Reasoning Enhancement), a novel two-stage framework for disagreement detection. Its first stage, Concept Alignment, establishes a **sub-concept taxonomy** critical for fostering consistent task understanding between LLMs and human experts. The subsequent Reasoning Enhancement stage employs structured curriculum learning workflows (**rationale to critique**, **counterfactual to detection**) designed to enhance reasoning and mitigate reliance on spurious correlations. This stage is underpinned by high-quality synthesized rationale and critique data, as well as counterfactual data generated using the taxonomy.

Our proposed CARE framework is evaluated on a widely-used dataset comprising five sub-domains, achieving state-of-the-art performance under both zero-shot and supervised learning tasks and demonstrating generalization ability across diverse domains. Furthermore, the generated taxonomy and reasoning process provide a foundation for subsequent analysis.

Limitations

The reliance on teacher LLM APIs for synthesizing data, including rationale, critique, and counterfactual, incurs relatively high costs, as detailed in Appendix A. This makes scaling to larger datasets expensive. Future work should focus on reducing the cost of synthetic data generation while maintaining data quality. Potential directions include exploring more efficient generation strategies or utilizing lightweight models (Zhang et al., 2025). Another promising research direction involves using LLMs to effectively integrate graph-based information with rich textual data (Pang et al., 2025).

Ethics Statement

The dataset used in this paper is sourced from publicly available open-access resources. Specifically, the DEBAGREEMENT dataset (Pougué-Biyong et al., 2021) provides full text data under an open-access license. All user-specific private information has been removed from the dataset.

The counterfactual augmented data used in this study are obtained through the GPT-4o API service provided by OpenAI, in compliance with their terms of use and ethical guidelines. Some examples presented in the paper may reflect particular stances or tendencies. These instances are randomly sampled from the dataset for illustrative purposes, to better demonstrate the characteristics of the data and the task, and do not represent or reflect the personal views of the authors.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (62372483).

References

Shijie Chen, Yu Zhang, and Qiang Yang. 2024. Multi-task learning in natural language processing: An overview. *ACM Computing Surveys*, 56(12):1–32.

Bowen Ding, Qingkai Min, Shengkun Ma, Yingjie Li, Linyi Yang, and Yue Zhang. 2024. A rationale-centric counterfactual data augmentation method for cross-document event coreference resolution. In *NAACL*, pages 2494–2507.

Shangbin Feng, Herun Wan, Ningnan Wang, Zhaoxuan Tan, Minnan Luo, and Yulia Tsvetkov. 2024. What does the bot say? opportunities and risks of large language models in social media bot detection. In *ACL*, pages 11569–11592.

Haixia Han, Jiaqing Liang, Jie Shi, Qianyu He, and Yanghua Xiao. 2024. Small language model can self-correct. In *AAAI*, pages 18162–18170.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu. Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*, page 3.

Zhaopei Huang, Jinming Zhao, and Qin Jin. 2024. Ecr-chain: Advancing generative language models to better emotion-cause reasoners through reasoning chains. In *IJCAI*, pages 5068–5076.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Luca Iandoli, Simonetta Primario, and Giuseppe Zollo. 2021. The impact of group polarization on the quality of online debate in social media: A systematic literature review. *Technological Forecasting and Social Change*, 170:120924.

Ayush Kaushal, Avirup Saha, and Niloy Ganguly. 2021. twt–wt: A dataset to assert the role of target entities for detecting stance of tweets. In *NAACL*, pages 3879–3889.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*, pages 22199–22213.

Ang Li, Jingqian Zhao, Bin Liang, Lin Gui, Hui Wang, Xi Zeng, Xingwei Liang, Kam-Fai Wong, and Ruifeng Xu. 2024. Mitigating biases of large language models in stance detection with counterfactual augmented calibration. In *NAACL*, pages 1651–1662.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, and 1 others. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, and 1 others. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Isabelle Lorge, Li Zhang, Xiaowen Dong, and Janet B Pierrehumbert. 2024. Stentconv: Predicting disagreement with stance detection and a signed graph convolutional network. In *LREC-COLING*, pages 13299–13309.

Yun Luo, Zihan Liu, Stan Z Li, and Yue Zhang. 2023. Improving (dis)agreement detection with inductive social relation information from comment-reply interactions. In *WWW*, pages 1584–1593.

Yunhe Pang, Bo Chen, Fanjin Zhang, Yanghui Rao, Evgeny Kharlamov, and Jie Tang. 2025. Guard: Effective anomaly detection through a text-rich and graph-informed language model. In *SIGKDD*, pages 2222–2233.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. Topicgpt: A prompt-based topic modeling framework. In *NAACL*, pages 1837–1852.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*.

John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. 2021. Debagreement: A comment-reply dataset for (dis)agreement detection in online debates. In *NeurIPS Datasets and Benchmarks*.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *ACL*, pages 5434–5445.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, and Bo Zheng et al. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2407.10671*.

Manoel Horta Ribeiro, Pedro H Calais, Virgílio AF Almeida, and Wagner Meira Jr. 2017. "everything i disagree with is #fakenews": Correlating political polarization and spread of misinformation. *arXiv preprint arXiv:1706.05924*.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *EMNLP Findings*, pages 8990–9005.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *WWW*, pages 613–624.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *ACL Findings*, pages 10336–10351.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022a. Identifying and mitigating spurious correlations for improving robustness in nlp models. In *NAACL Findings*, pages 1719–1729.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2022b. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576.

Linyi Yang, Lifan Yuan, Leyang Cui, Wenyang Gao, and Yue Zhang. 2022. Factmix: Using a few labeled in-domain examples to generalize to cross-domain named entity recognition. In *COLING*, pages 5360–5371.

Xiao Yu, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhou Yu. 2024. Teaching large language models to self-debug. In *NAACL*, pages 3560–3576.

Jianhua Yuan, Yanyan Zhao, Yanyue Lu, and Bing Qin. 2022a. Ssr: Utilizing simplified stance reasoning process for robust stance detection. In *COLING*, pages 6846–6858.

Jianhua Yuan, Yanyan Zhao, and Bing Qin. 2022b. Debiasing stance detection models with counterfactual reasoning and adversarial bias learning. *arXiv preprint arXiv:2212.10392*.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. 2022. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*.

Yanyue Zhang, Yilong Lai, Zhenglin Wang, Pengfei Li, Deyu Zhou, and Yulan He. 2024a. Opinions are not always positive: Debiasing opinion summarization with model-specific and model-agnostic methods. In *LREC-COLING*, pages 12496–12513.

Yanyue Zhang, Pengfei Li, Yilong Lai, Yulan He, and Deyu Zhou. 2025. Lass: A novel and economical data augmentation framework based on language models for debiasing opinion summarization. In *COLING*, pages 6169–6183.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2024b. Small language models need strong verifiers to self-correct reasoning. In *ACL Findings*, pages 1417–1432.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *ACL*, pages 4227–4241.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2024. Explore spurious correlations at the concept level in language models for text classification. In *ACL*, pages 10397–10411.

A Computational Cost Analysis

The computational cost (in US dollars) of our proposed framework primarily stems from the data processing stages, specifically Concept Alignment (CA) and Reasoning Enhancement (RE).

The cost associated with the CA stage is relatively low. In this stage, we only sample 50 data instances labeled as "neutral" for the iterative generation of the sub-concept taxonomy. Based on the pricing strategy of the utilized LLM API, the estimated computational cost for this stage is approximately \$0.002 × 50 = \$0.1.

The RE stage incurs a comparatively higher computational cost per data instance, the cost for generating rationales is approximately \$0.006 per instance, for which, the cost for generating critiques is approximately \$0.0035 per instance, and the cost for generating counterfactual data corresponding to the three stance labels (agree, neutral, disagree) is approximately \$0.0023× 3 = \$0.0069 per instance. Thus, the estimated average processing cost per data instance in the RE stage is approximately \$0.006 + \$0.0035 + \$0.0069 = \$0.0164, which represents the worst-case scenario. However, our data distillation algorithm indicates that only 22% of incorrect rationales require critique. Consequently, the actual cost per data instance is approximately \$0.01367.

In conclusion, the average cost to train 1000 data instances under our CARE framework is \$13.7, which is considered acceptable.

B Potential Effects of Data Memorization on CARE's Performance

The potential effects of data memorization on CARE's performance can not appear in LLMs' training data, because the label text is not directly associated with corresponding documents in the raw DEBAGREEMENT dataset (Pougué-Biyong et al., 2021).

C Data Efficiency and Low-Resource Robustness

For a comprehensive evaluation of our model, we first conduct an asymmetric comparison, benchmarking our LLM-based model fine-tuned on a downsampled dataset against traditional PLM baselines trained on the full dataset. This initial result validates our model's effectiveness in Section 5.2

Subsequently, to facilitate a rigorous and fair symmetric comparison, we uniformly train all baseline models, including various PLM-based approaches, on the same downsampled dataset. The in-domain and cross-domain results, presented in Tables 6 and 7, reveal that all baselines suffer from significant performance degradation as data volume decreases. In contrast, our findings provide compelling evidence that by enhancing the reasoning capabilities of LLMs through the CARE framework, our model consistently and comprehensively outperforms all baselines under these same data-scarce conditions. This series of experiments not only highlights the remarkable data efficiency of our model but also underscores the pivotal role of robust reasoning in maintaining high performance when high-quality data is limited.

D Human Assessment of Counterfactual Quality

To validate the quality of our generated samples, in this part, we conduct a multi-faceted human evaluation to assess our counterfactuals across three key dimensions: **Fidelity**, **Coherence**, and **Stance Consistency**.

D.1 Methodology

We recruited 12 Master's students in Computer Science as annotators. Following a training session with task descriptions and examples, they evaluated a random set of 20 comment-reply pairs.

D.2 Evaluation and Results

Fidelity (Topical Relevance) To measure if our counterfactuals remain on-topic, we asked annotators a binary (yes/no) question: "Is this reply relevant to the original comment's topic?" The results showed that **97.08%** of our generated samples were rated as "Relevant", confirming that our method maintains high fidelity and preserves the core conversational context.

Coherence (Linguistic Quality) To assess linguistic quality, annotators were asked to classify each generated sample as either "Fluent" or "Not Fluent" based on its fluency and grammaticality. **93.33%** of the samples were judged as "Fluent", demonstrating that our method produces coherent, well-formed sentences.

Model	Agree			Disagree			Neutral			All	
	Prec	Rec	M-F1	Prec	Rec	M-F1	Prec	Rec	M-F1	Acc	M-F1
Fine-tuned PLMs											
Roberta	65.34	13.92	22.95	47.83	96.07	63.86	80.00	0.77	1.52	49.33	29.45
STEv	44.36	74.15	55.51	53.83	39.67	45.68	64.58	5.96	10.92	47.83	37.37
StanceRel	79.94	80.68	80.31	80.97	79.21	80.08	53.60	55.77	54.67	76.14	71.68
Fine-tuned LLMs											
Qwen-7b-ft	**90.75**	82.37	86.36	83.55	92.26	87.69	65.49	60.58	62.94	83.49	79.00
Qwen-7b-CARE-ft	89.69	**86.95***	**88.32***	**87.81***	**92.59***	**90.14***	**69.54***	63.65	**66.47***	**85.92***	**81.64***

Table 6: Performance comparison of different models on the in-domain task. Qwen-7b-ft indicates only a fine-tuned detection task on Qwen-7b. Qwen-7b-CARE-ft indicates fiue-tuned via our CARE framework. * denotes our CARE improves the best baseline at p-value < 0.05 with paired t-test.

Model	r/Br		r/Cl		r/BLM		r/Re		r/De		Average	
	Acc	M-F1	Acc	M-F1	Acc	M-F1	Acc	M-F1	Acc	M-F1	Acc	M-F1
Fine-tuned PLMs												
Roberta	50.91	29.81	48.56	25.56	43.39	28.00	48.56	29.44	44.30	27.56	47.14	28.07
STEv	43.64	35.14	47.84	36.04	52.12	32.27	47.75	36.70	50.27	37.17	48.32	35.46
StanceRel	75.19	71.83	75.10	71.70	79.61	73.92	74.98	69.89	78.52	72.47	76.68	71.96
Fine-tuned LLMs												
Qwen-7b-ft	80.79	75.38	84.40	80.82	86.06	82.06	85.09	79.84	82.33	76.51	84.13	79.34
Qwen-7b-CARE-ft	**82.65***	**78.67***	**86.94***	**82.71***	**86.30***	**82.07***	**87.07***	**82.97***	**84.95***	**79.59***	**86.34***	**81.47***

Table 7: Performance comparison of different models on the cross-domain task. Qwen-7b-ft indicates only a fine-tuned detection task on Qwen-7b. Qwen-7b-CARE-ft indicates fiue-tuned via our CARE framework. * denotes our CARE improves the best baseline at p-value < 0.05 with paired t-test.

Stance Consistency and Clarity This is the most critical evaluation for our task. Annotators judged the stance of both the original pairs and our generated counterfactual versions. As detailed in Table 8, the results yield two key insights.

Data Type	All 12 Annotators (N=12)	Proficient 6 Annotators (N=6)
Original Data	63.75%	74.17%
Counterfactual Data	78.33%	81.67%

Table 8: Human annotation accuracy for Stance Consistency. Proficient Annotators are defined as those achieving >70% accuracy on the original data.

First, **the task is inherently ambiguous for humans**, as evidenced by the modest accuracy on the Original Data (63.75%). This highlights the difficulty of the task and motivates our approach. Second, **our method generates samples with unambiguous stances**. Critically, annotator accuracy was significantly higher on our Counterfactual Data, rising from 63.75% to 78.33%. This pattern was even more pronounced for proficient annotators, whose accuracy improved from 74.17% to 81.67%. This strongly indicates that our method excels at generating samples with a clear and easily recognizable stance, making them high-quality training signals.

D.3 Summary

In conclusion, our human evaluation results confirm that the generated counterfactual data exhibit high fidelity and coherence and, most importantly, possess a clear and consistent stance, rendering them highly effective for augmenting data in this challenging task.

E Details of Taxonomy

Within our CARE framework, the CA stage facilitates the generation of sub-concepts for complex concepts present in task descriptions, thereby forming a taxonomy for conceptual alignment. In the disagreement detection task, we observe a *conceptual gap* between LLMs and human experts regarding neutral concept, whereas this gap is comparatively smaller for agree and disagree concepts. Consequently, our taxonomy incorporates five sub-concepts for the neutral concept, alongside descriptive sentences for agree and disagree that were generated by GPT-4o. The cases of taxonomy are shown in Table 9.

concept	Description
Neutral	
1. Inquiry-based Neutral	Maintaining neutrality by asking questions or seeking more information.
2. Observational Neutral	Offering neutral observations or descriptions without expressing agreement or disagreement.
3. Informational Neutral	Providing factual or explanatory information without taking a stance.
4. Speculative Neutral	Maintaining neutrality by offering speculative or hypothetical responses without taking a firm stance.
5. Emotional Neutral	Expressing neutral sentiments or emotions without taking a stance on the issue.
Agree	If the reply expresses agreement with the original comment, it indicates support or confirmation of the point made in the comment. A reply reinforces or extends the argument of the original comment, usually using an affirmative tone.
Disagree	If the reply expresses a disagreement or rebuttal to the original comment, indicating an opposing position or viewpoint. The response will usually directly present an opinion or rebuttal that is opposed to the original comment.

Table 9: Taxonomy cases.

F Case of Counterfactual Data

Table 10 shows the case of counterfactual data.

Type	Content
Original COMMENT	This lady seems like a great pick and is someone who won't get bullied in the role.
Original REPLY [Disagree]	Flint has entered the chat. Seriously people. You have to be joking. Are you joking?
REPLY [Disagree]	Flint is a perfect example of why she's not the right choice. She failed to address the water crisis effectively, so how can we trust her to handle something as massive as climate change? This pick is a joke.
REPLY [Informational Neutral]	Gina McCarthy was the EPA Administrator during the Flint water crisis, which was a significant environmental and public health issue. Her role in that situation has been a point of discussion when evaluating her career.
REPLY [Agree]	Gina McCarthy has a strong track record, and her experience with Flint shows she's willing to take on tough challenges head-on. She's definitely the right person for the job.

Table 10: The case of counterfactual data.

G Prompt Settings

The prompts are shown in Table 11 to Table 15.

Agent prompt of Taxonomy Expert

[Role] You are an expert in conceptual taxonomy. Your task is to identify generic neutral sub-concept in the document. If any relevant neutral sub-concepts are missing from the provided set of neutral sub-concepts, add them. Otherwise, export the existing neutral sub-concepts identified in the document.
[Neutral sub-concept] {Neutral sub-concepts}
[Examples] Adding "Inquiry-based Neutral"
comment-reply pair:
comment: "This is real. An oxygen atom weighs more than a carbon atom, and you use two oxygen atoms from the atmosphere for each carbon atom to make CO2."
reply: "Is every single carbon atom converted into CO2, or is there other less harmless stuff that comes out? That's nuts to think that by weight there is that much CO2."
Your response:
Inquiry-based Neutral: Maintaining neutrality by asking questions or seeking more information.
[Instructions]
Step 1: Determine the neutral sub-concept in the comment reply pair.
- The neutral sub-concept must be as general as possible. They must not be specific to a particular comment pair.
- Neutral sub-concept must reflect a single type, not a combination of neutral sub-concept.
- New neutral sub-concept must have a short general label, and a description of the neutral sub-concept.
Step 2: Do one of the following:
1. If there are already duplicate or related neutral sub-concept in the taxonomy, refine the sub-concept descriptions appropriately (if needed) and export those neutral sub-concept to this point.
2. otherwise, add your neutral sub-concept. Stop here and output the added neutral sub-concept.
[Input] {comment-reply pair}
[Output Format] ONLY return the relevant or modified sub-concept in the taxonomy. Your response should be in the following format: neutral sub-concept Label: neutral sub-concept Description

Table 11: Prompt template for identifying or generating neutral sub-concepts using Taxonomy Expert agent.

Agent Prompt of Internet Rewriter

[Role] You are a helpful assistant at Disagreement Detection are very familiar with the way of expression on the Internet. Your task is to rewrite the counterfactual child_reply to meet the new attitude, while retaining internal coherence and avoids unnecessary changes.
[Stance Concept] {Taxonomy}
[Examples]
comment: Bad study. Chemistry, Physics and Biology textbooks shouldn't be devoting that much space to climate change. 4% or about 600 pages of 15,000 pages between 16 books seems reasonable.
...
reply: Yeah. I don't know what the hell they think that is supposed to indicate?? Why would my physiology class talk about climate change?
Your response: Climate change has been incorporated into a lot of different subjects recently because of its broad impact on various fields, including biology, chemistry, and physics. ...
[Instructions]
Giving you comment and reply on post under subreddit. The reply expresses a {stance} to the comment. Please make changes to the reply to express a {sub-concept} attitude to the comment.
[Input] {comment-reply pair, post,subreddit}
[Output Format] ONLY include the rewritten reply in your response.

Table 12: Prompt template for generate counterfactual data using Internet Rewriter agent.

[Instruction] You are a helpful assistant at Disagreement Detection. Given the AUTHOR_PARENT's PARENT_COMMENT, and the AUTHOR_CHILD's CHILD_REPLY under POST in SUBREDDIT, categorize the child_reply's stance toward parent_comment from agree, disagree and neutral.

[Input] {AUTHOR_PARENT, AUTHOR_CHILD, SUBREDDIT, POST, PARENT_COMMENT, CHILD_REPLY}

[Output Format] Only include one true stance value.

Table 13: Prompt template for direct disagreement detection.

[Instruction] You are a helpful assistant at Disagreement Detection.

Given the AUTHOR_PARENT's PARENT_COMMENT, and the AUTHOR_CHILD's CHILD_REPLY under POST in SUBREDDIT, categorize the child_reply's stance toward parent_comment from agree, disagree and neutral.

First, list Causal factors (i.e., keywords, phrases, contextual information, semantic meaning, semantic relationships, tones, references, Internet expression, topic-specific jokes) of comment-reply that support the Disagreement Detection.

Next, deduce the diagnostic REASONING process from premises (i.e. clues, input) that support the Disagreement Detection.

Finally, based on the clues, the reasoning and the input, categorize the AUTHOR_CHILD's stance toward PARENT_COMMENT from agree, disagree and neutral. Please think step by step.

[Input] {AUTHOR_PARENT, AUTHOR_CHILD, SUBREDDIT, POST, PARENT_COMMENT, CHILD_REPLY}

[Output Format] The last sentence must contain only one true stance value.

Table 14: Prompt template for use CARP for disagreement detection.

[Instruction] You are a helpful assistant at Disagreement Detection.

Providing you with the AUTHOR_PARENT's PARENT_COMMENT, and the AUTHOR_CHILD's CHILD_REPLY under POST in SUBREDDIT, along with the previous misjudgment and the rationale of misjudgment. Categorize the child_reply's stance toward parent_comment from agree, disagree and neutral.

First, self-critique the previous misjudgment and rationale of misjudgment.

Second, self-correct to the correct judgment.

[Input] {AUTHOR_PARENT, AUTHOR_CHILD, SUBREDDIT, POST, PARENT_COMMENT, CHILD_REPLY , previous misjudgment, rationale of misjudgment}

[Output Format] The last sentence must contain only one true stance value.

Table 15: Prompt template for self-critique.