A Benchmark for Cross-Domain Argumentative Stance Classification on Social Media

Jiaqing Yuan, Ruijie Xi, Munindar P. Singh
North Carolina State University
Raleigh, North Carolina 27606 USA
jyuan23@ncsu.edu, rxi@ncsu.edu, mpsingh@ncsu.edu

Abstract

Argumentative stance classification is key role in identifying authors' viewpoints on specific topics. However, generating diverse pairs of argumentative sentences across various domains is challenging. Existing benchmarks often come from a single domain or focus on a small set of topics. Additionally, manual annotation for accurate labeling is time-consuming and labor-intensive. To address these challenges, we propose leveraging website rules, readily available expert-curated content, and large language models to bypass the need for human annotation. Our approach produces a multidomain benchmark comprising 4,498 topical claims and 30,961 arguments from three sources, and spanning 21 domains. We benchmark the dataset in fully supervised, zero-shot, and few-shot settings, shedding light on the strengths and limitations of the different methodologies.

1 Introduction

Argumentation is a pervasive human activity present in various aspects of everyday life, which involves expressing viewpoints backed by reasons or attempting to persuade others towards a specific perspective (Guo, Zhang, and Singh 2020; Guo and Singh 2023; Sobhani, Inkpen, and Matwin 2015). A crucial challenge in argument mining is stance classification (Küçük and Can 2020), where the goal is to classify an argument's stance as either *favor*, *against*, or *neutral* regarding a given claim. For example, argument *The possession of nuclear weapons provides countries with a strong defense mechanism, deterring potential adversaries from launching attacks* can be classified as *against* the claim *All countries should give up their nuclear weapons.*

Social media sites are prominent venues for the exchange of ideas and arguments (AlDayel and Magdy 2021a). Claims are often simplified into noun phrase topics, such as "nuclear weapon" for the above example. Previous research has spent a lot of effort in constructing datasets concerning various topics. For example, Mohammad et al. (2016) constructed a dataset with tweets commenting on *Atheism*, *Climate change*, *Feminist*, *Hillary Clinton*, *Abortion*, and *Donald Trump.* Conforti et al. (2020b) studied public opinion toward four financial merger events, Glandt et al. (2021) investigated stance classification on three policies during Covid-19. Recently, Cruickshank, Soofi, and Ng (2024) collected YouTube video comments and annotated their stance towards the U.S. military.

One challenge to stance classification comes from the variety of stance topics (Allaway and McKeown 2020). Many prior benchmarks are not topically diverse. As mentioned above, they typically feature a handful of topics, each with a corpus of comments to facilitate the training of supervised models *dedicated* to that topic (Stab et al. 2018). The acquisition of stance labels relies on human annotators for ground truth (Küçük and Can 2020), which is time-consuming and difficult to scale up. Besides, most previous benchmarks focus on a single genre or source.

Accordingly, our objective is to construct a diverse and multisource stance classification benchmark without human annotation. Allaway and McKeown (2020) categorize stance classification into two categories based on the topic: *topic-phrase* and *topic-position*. For the former, the topic is typically a noun phrase (including proper noun), such as *nuclear weapon.* For the latter, the topic is a complete position claim such as *All countries should give up their nuclear weapons.* Notably, the argument we introduce at the beginning of this section would be classified as *favor* for the former and *against* for the latter. This reveals a major difference between topic-phrase and topic-position stance classification: the latter is context-dependent. Our benchmark focuses on topic-position, as we argue that topic-phrase can be easily converted into topic-position by constructing a positional claim, which has a more general form. Preferably, a truly intelligent stance classification system should be able to grasp the meaning of the topical claim and reverse its prediction when the claim reverses itself.

We construct our benchmark from three types of sources: a social media website, two debate websites, and arguments generated by large language models (LLMs). For the social media website, we leverage conversations from a subreddit called ChangeMyView[1] from Reddit, where a poster challenges other users to change the poster's opinion expressed by a positional title. The comments labeled by the poster as successful can be seen as counterarguments to the title (Guo, Zhang, and Singh 2020; Yuan and Singh 2023). On the debate websites, opposing arguments are curated by

[1]https://www.reddit.com/r/changemyview/

Topical Claim	Comment	Stance
Space exploration is a waste of money.	Instead of decreasing resources by space travel and such, we must deal with problems on Earth first. Why bother spending all this money on exploring space when we could be helping our own planet that us humans live on...	Favor
Animals have rights.	It makes no sense to give animals rights because they cannot makes decisions about what is right and wrong and will not try to treat us in an ethical manner in return...	Against
All student loan debt should be eliminated.	Schools already have a heavy workload and limited resources. Adding moral education to their curriculum may place an additional burden on teachers and administrators. It could divert valuable time and resources away from core academic subjects, potentially compromising the quality of education provided to students...	None

Table 1: Examples from our benchmark.

the users, who provide clear stance labels (Guo and Singh 2023). LLMs have been shown effective for data augmentation (Sahu et al. 2022; Yoo et al. 2021; Edwards et al. 2022). To further enrich the diversity of the benchmark, we use LLMs to generate arguments on both sides for a given topical claim. The foregoing steps yield a combined dataset of 4,498 topics and 30,961 arguments spanning 21 domains. Table 2 compares our benchmark with previous benchmarks. Using this benchmark, we tackle the following research questions.

RQ1 How does LLM-generated data benefit stance classification in real-world applications?

We conduct experiments on individual real-world datasets and on combined LLM-generated and real-world datasets to show the benefit of integrating LLM-generated data during training. We use traditional machine learning methods (AlDayel and Magdy 2021b) commonly applied in stance classification alongside pretrained LLMs like BERT (Devlin et al. 2019).

RQ2 To what extent do stance classification models generalize across topics and domains within a topic-position framework?

Topic-position stance classification offers a notable advantage due to its flexibility. It analyzes pairs of argumentative sentences instead of being limited to the topical noun phrase. This research question addresses how effectively stance classification can generalize across both topics (i.e., different topics from the same source) and sources (i.e., different sources covering similar or different topics).

RQ3 How does supervised finetuning compare to zero-shot and few-shot learning with LLMs for cross-domain stance classification?

Supervised finetuning and in-context learning are two common methods for adapting models to specific tasks. Research shows that LLMs can adapt well to new tasks in zero-shot and few-shot scenarios (Brown et al. 2020), which do not require any training data. In contrast, supervised finetuning substantially improves performance on data within the trained domain, but often fails to generalize effectively to new domains (Ng and Carley 2022a). This question addresses the relative effectiveness of these two approaches for cross-domain stance classification.

Findings We observe that incorporating LLMs generated data into the training process enhances in-domain performance of traditional machine learning techniques, including Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Bidirectional Long Short-Term Memory (BiLSTM) networks. However, this strategy yields inconsistent outcomes when applied to finetuning contemporary LLMs. We also observe that generative models consistently outperform classification models with supervised finetuning. LLMs yield commendable performance in zero-shot settings for cross-domain evaluation, though a substantial performance gap remains in comparison to in-domain supervised finetuning. Furthermore, in few-shot experiments, instruction-tuned LLMs consistently outperform their non-instruction-tuned counterparts, highlighting the effectiveness of instruction-tuning as a robust approach for adapting LLMs to downstream tasks.

Contributions Our contributions are twofold:

- We propose a scalable and extensible framework to construct a diverse and multisource benchmark for argumentative stance classification without human annotation.
- We implement and evaluate fully-supervised, zero-shot, and few-shot learning using LLMs. This thorough assessment facilitates a comparative analysis of various methodologies, emphasizing the efficacy of instruction-tuning for optimizing the performance of LLMs.

The rest of this paper is structured as follows: Section 2 reviews related work and positions our research in relation to existing studies. Section 3 introduces the proposed framework for constructing a multisource stance classification benchmark. Section 4 presents experiments using both traditional machine learning models and contemporary LLMs with supervised finetuning. Section 5 examines zero-shot and few-shot learning across various LLM families. Section 6 summarizes our work and discusses its broader impact.

2 Related Work

Stance is a speaker's evaluation of a proposition or topic. The proposition may be implicit as a *topic-phrase* (a noun phrase) or explicit as a *topic-position* (a positional claim) (Allaway and McKeown 2020). Datasets from early research

Authors	Diversity	Source	Topic Type	Size
Mohammad et al. (2016)	6 topics	Twitter	Phrase	4870
Stab et al. (2018)	8 topics	Google query	Phrase	25 492
Allaway and McKeown (2020)	4641 noun phrases	Debate website	Phrase	18 545
Ferreira and Vlachos (2016)	300 rumor claims	News article	Position	2595
Gorrell et al. (2019)	various claims	Twitter, Reddit	Position	8574
Bar-Haim et al. (2017)	55 claims	Debate website	Position	2394
Hanselowski et al. (2019)	6422 claims	Fact-check website	Position	19 439
This paper	*4498 claims*	*Reddit, Debate website, LLM*	*Position*	30 961

Table 2: Comparison with previous benchmarks.

originate from arguments in online debate forums (Somasundaran and Wiebe 2010; Murakami and Raymond 2010; Walker et al. 2012; Hasan and Ng 2014) and mostly fall under the topic-phrase category (AlDayel and Magdy 2021a). More recent datasets cover various topics (Sobhani, Inkpen, and Zhu 2017; Qazvinian et al. 2011; Mohammad et al. 2016; Conforti et al. 2020b; Li et al. 2021; Glandt et al. 2021). For topic-position stance classification, datasets primarily come from news articles, where headlines are used as the topic-phrase (Ferreira and Vlachos 2016; Habernal et al. 2018; Conforti et al. 2020a; Chen et al. 2019; Qazvinian et al. 2011). Many existing datasets are generated from one source in one domain and focus on comments for a small set of topics, followed by human annotation.

We emphasize the topic-position variant of stance classification because phrases can be transformed into positions by formulating an affirmative claim (e.g., *Abortion* maps to *Abortion should be legalized*). Unlike previous works, which rely on human annotators for labeling, we leverage website rules, readily available expert-curated content, and large language models to acquire faithful stance labels.

3 Benchmark Construction

We now describe the details for building our benchmark.

3.1 Dataset Collection

To enhance textual diversity, our benchmark is curated with content from three types of sources: a social media website, debate websites, and LLM generation.

Social media website ChangeMyView (CMV) is a subreddit (forum on Reddit) in which participants contribute their opinions and engage in discussions with the explicit aim of defending their perspectives. A typical CMV post adheres to a particular structure: it begins with the abbreviation "CMV: " signifying Change My View, followed by a concise representation of the author's viewpoint. Subsequently, the body of the post provides the rationale for their stance. If any of the comments made by the participants successfully manage to influence a shift in the author's viewpoint, the author acknowledges this change by awarding the commenter with a *delta*. Therefore, we extract the title, body, and delta-awarded comments, where the title corresponds to the designated topical claim, the body constitutes a supportive argument, and the comments bearing a delta reward act as counterarguments.

We leverage two existing CMV datasets (Tan et al. 2016; Al-Khatib et al. 2020) because Reddit now limits its API to return at most 1000 recent posts. To keep the text meaningful and concise, we select bodies and comments with a length between 20 to 200 words.

Debate websites We selected two sites dedicated to online debates, idebate.net and debatewise.org, due to their well-structured presentation of arguments for and against a position. These sites provide clear and comprehensive arguments, thus obviating the need for annotation. We captured the subjects of a debate along with the associated arguments.

Some of the arguments were excessively verbose. However, we observed that the stance of an argument can typically be discerned within the initial few sentences. Therefore, in the interest of improving the manageability of the data, we retained only the initial five sentences of each argument. This choice aligns with our restricting the CMV arguments to 20 to 200 words.

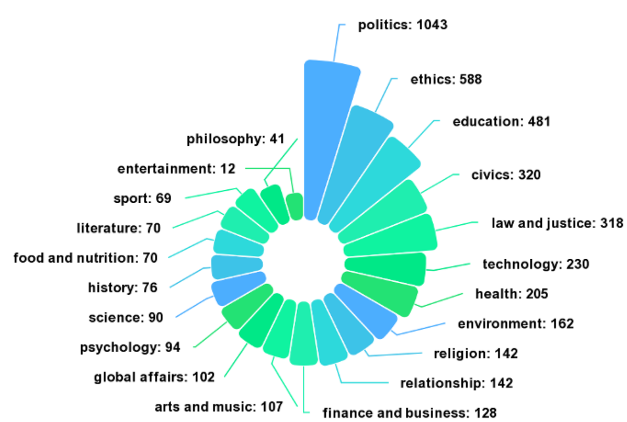


Figure 1: Domain distribution of the topics in our dataset.

3.2 LLM-Generated Arguments

We use GPT-3 to generate text because it produces coherent, contextually relevant, and fluent language. GPT models are pretrained on vast amounts of diverse data, enabling them to mimic human-like language patterns across various

Prompt 1: Generate a claim for each topic.
Given a question or topic, generate a controversial claim. **Input:** Should Halloween costumes be allowed in schools? **Output:**

Prompt 2: Generate three opposing arguments
Given a topic, write three distinct supporting arguments and three opposing arguments. You should write in 1st person view rather than 3rd person view. Don't explicitly say I support or oppose. Don't summarize the points at the beginning. **Topic**: Halloween costumes should be allowed in schools.

Prompt 3: Classify arguments into predefined categories.
Given a topic, classify which domain the topic falls into. Output the domain directly without other words. Some example domains are sport, environment, civics, history, education, politics, technology, literature, arts and music, science, ethics and animal, finance and business, global affairs, health, psychology, law and justice, relationship, nursing, religion, food and nutrition. You should pick the category that most closely matches the topic. If none of the categories matches, you can use a new category of your own. **Topic**: Halloween costumes should be allowed in schools. **Output**:

topics. Previous research has demonstrated GPT's ability to encode beliefs into argumentative texts (Alshomary et al. 2021) and to determine rumors and antagonistic relationships between Twitter users by detecting stances in replies and quotes (Villa-Cox et al. 2020). We employ GPT-3[2] to produce arguments from multiple perspectives on diverse subjects. This process involves three steps.

Topic collection We conducted a search with Google to identify contentious subjects and compiled such topics from ten web sources (listed in Appendix A).

Claim generation The topics collected from the web have various forms. We use Prompt 1 shown below to generate a claim for each topic.

Argument generation For each claim, we use Prompt 2, shown below, to generate three arguments from both sides.

	CMV	Debate	LLMG	Total
Topical claims	1873	982	1643	4498
Arguments	6407	14 696	9858	30 961
Favor	1863	7319	4929	14 111
Against	4544	7377	4929	16 850
None	5609	2904	4929	13 442
Args/Claim	6.4	17.9	9.0	6.9
Length/Arg	105.2	102.3	56.2	86.7

Table 3: Statistics for our benchmark. Args/Claim means the average number of arguments per claim, and Length/Arg means the average number of words for each argument.

Constructing neutral arguments The above methods assign stance labels of favor or against, as in some prior datasets. Generating a neutral stance, however, is difficult since a judgment of neutrality often depends on the annotator's interpretation. We seek to compel the LLM to focus on how the argument and the topical claim relate, moving beyond reliance on surface-level linguistic cues.

For each claim, we create neutral arguments by randomly selecting arguments for other claims. Consequently, we define a neutral stance as one that includes either irrelevant arguments or instances where no discernible stance can be inferred. One way is to randomly sample arguments. However, it may yield semantically distinct instances that are easily captured by the model. That is, classification is easier when the sampled arguments address completely different topics from the claim. To improve this, we use BERT to embed all claims and arguments. For each claim, we randomly sample three arguments falling within the similarity score range of $[0.3, 0.5]$. This criterion is motivated by the fact that highly similar arguments may include content that may convey an implied stance. Conversely, moderately similar arguments may seem to discuss related subjects but be subtly different, thereby forming more challenging examples. For example, the third claim in Table 1 concerns student loan debt, but the comment is about moral education: thus, it doesn't indicate a stance about the claim, though they both relate to education. We use "LLMG" to refer to the dataset generated by GPT-3.

3.3 Dataset Characteristics

Table 1 and Table 3 show some examples and the statistics for our dataset, respectively. Our dataset exhibits greater diversity than prior datasets. We apply GPT-3 to classify these arguments into predefined categories or generate novel domains, as shown by Prompt 3. This process yields over 100 domains across all topics. We consolidate these into 21 principal domains. Figure 1 illustrates the distribution of these domains across the topics.

3.4 Validate LLM-Generated Dataset

To evaluate LLMG's quality, we (1) manually verify that GPT-3 adheres to the guidelines and produces accurate responses and (2) compare the lexical diversity of real-world datasets and GPT-3 generated content.

Human verification First, we applied regular expressions to search for phrases like *as an AI, I cannot* and its variations, such as *as an AI, I can't*, in LLMG and found no such occurrences. Second, three independent raters labeled 200 randomly selected arguments from LLMG. We designed a survey to assess two key aspects of the GPT-3 response:

Acceptance: Does the sentence indicate that the AI refused to provide a response? For instance, does it contain variations of *as an AI, I cannot*?

[2] https://platform.openai.com/docs/guides/text-generation/chat-completions-api

Model	CMV			Debate			LLMG			CMV + LLMG			Debate + LLMG		
	Against	Favor	None	Against	Favor	None	Against	Favor	None	Against	Favor	None	Against	Favor	None
SVM	0.423	0.142	0.332	0.311	0.340	0.168	0.520	0.568	0.558	$0.463_{4.0}$	$0.332_{19.0}$	$0.426_{16.9}$	$0.480_{9.4}$	$0.492_{15.2}$	$0.265_{9.7}$
CNN	0.481	0.328	0.473	0.421	0.460	0.237	0.620	0.646	0.598	$0.546_{6.5}$	$0.434_{15.0}$	$0.554_{8.1}$	$0.552_{13.1}$	$0.506_{5.6}$	$0.286_{4.9}$
BiLSTM	0.524	0.372	0.531	0.471	0.474	0.356	0.662	0.645	0.651	$0.564_{4.0}$	$0.450_{16.9}$	$0.585_{7.8}$	$0.526_{5.4}$	$0.510_{3.6}$	$0.386_{3.0}$

Table 4: Macro-F1 scores with a single dataset versus using LLMG for weak supervision. The subscript numbers indicate performance improvement compared to a single dataset. The training and test splits are consistent with those in Table 6. The test set for CMV+LLMG and Debate+LLMG are the same as merely training using CMV and Debate, where the training set merges CMV and Debate with LLMG.

Model	Dataset	Against	Favor	None
SVM	CMV	0.40**	1.90***	0.94**
	Debate	0.40**	1.20**	0.67**
CNN	CMV	0.65**	1.06**	0.81***
	Debate	0.75***	1.31**	0.49**
BiLSTM	CMV	0.40**	0.78***	0.54**
	Debate	0.68**	0.91***	0.42**

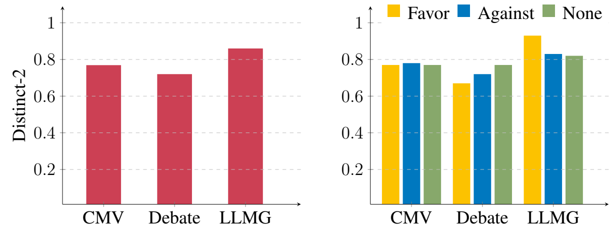
Table 5: Cohen's d Effect Sizes for SVM, CNN, and BiLSTM on the three datasets with and without LLMG. Here, ** and *** denote levels of statistical significance, where ** means a p-value of < 0.05 and ** means a p-value of < 0.001. The p-values are computed using a t-test of the differences in the means of Macro-F1 scores.

Accuracy: Does the response adhere to Prompt 2, specifically including three supporting and three opposing arguments in the correct order?

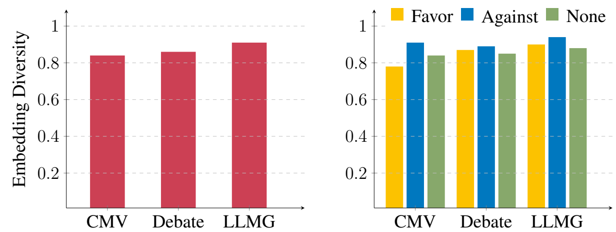
Each question could be answered as *Yes* or *No*. We observed that the raters, based on majority voting, found no instances of refusal to answer, and 99% of the responses adhered to the instructions in Prompt 2. The few answers that did not follow the prompt were instances where the first-person perspective was not used. Both tasks were rated as *strongly related*, with Cohen's Kappa scores of 1.0 for AI acceptance and 0.95 for AI accuracy. These results indicate that LLMG contains high-quality text.

Lexical and semantic diversity Figure 2a illustrates that sentences in LLMG exhibit greater lexical diversity than those in the CMV and Debate datasets. Lexical diversity is quantified using the metric called *distinct-2*—the number of unique bigrams and normalizing by the total number of words generated—which is a popular metric for lexical diversity (Li et al. 2016; Park, Yang, and Park 2019).

The overall distinct-2 scores—i.e., averaged across all sentences on Favor, Against, and None, respectively—are as follows: 0.774 for CMV, 0.719 for Debate, and 0.863 for LLMG. Specifically, the distinct-2 scores for Favor are 0.771 for CMV, 0.782 for Debate, and 0.767 for LLMG; for Against; they are 0.674 for CMV, 0.719 for Debate, and 0.770 for LLMG; and for None, they are 0.792 for CMV, 0.825 for Debate, and 0.823 for LLMG. This finding aligns with previous research indicating that machine-generated text often exhibits greater lexical diversity than



(a) Lexical diversity is calculated by bigram percentage.



(b) Semantic diversity is calculated as $(1-\sigma)$, where σ is the relevant cosine similarity between EBR embeddings.

Figure 2: Lexical and semantic diversity scores across datasets and labels.

human-authored text (Lee, Liang, and Yang 2022; Ravi et al. 2024).

Moreover, we compare semantic diversity across the three datasets. *Diversity* is determined by averaging the cosine similarity between the BERT embeddings of each instance. Semantic diversity is measured by averaging the cosine similarities between BERT embeddings for each instance. A lower similarity indicates higher diversity. Figure 2b illustrates that the overall semantic diversity values are 0.84 for CMV, 0.86 for Debate, and 0.91 for LLMG. This indicates that LLMG has the highest semantic diversity of the three datasets. Breaking down by label for CMV, Debate, and LLMG, respectively, we see that: for Favor, the diversity values are 0.78, 0.87, and 0.90; for Against, 0.91, 0.89, and 0.94; and for None, 0.84, 0.85, and 0.88.

In other words, LLMG shows the highest diversity across all labels, underscoring its superiority in capturing greater lexical and semantic variations than both CMV and Debate. Thus, LLMG is potentially better suited for more nuanced analyses than the other datasets, making it a valuable resource for investigating stance.

	train				dev				test			
	topics	args	favor	against	topics	args	favor	against	topics	args	favor	against
CMV	748	2633	744	1889	188	626	188	438	937	3148	931	2217
Debate	392	5860	2934	2926	99	1538	757	781	491	7298	3636	3662
LLMG	656	3936	1968	1968	165	990	495	495	822	4932	2466	2466
Total	1796	12 429	5646	6783	452	3154	1440	1714	2250	15 378	7033	8345

Table 6: Data distributions between train, dev, and test splits.

	BERT				T5				LLaMa			
	CMV	Debate	LLMG	Avg	CMV	Debate	LLMG	Avg	CMV	Debate	LLMG	Avg
CMV	0.789	0.477	0.557	0.608	0.771	0.489	0.752	0.679	0.820	0.513	0.652	0.662
Debate	0.765	0.594	0.738	0.699	0.529	0.616	0.791	0.645	0.665	0.672	0.822	0.720
LLMG	0.554	0.473	0.770	0.599	0.593	0.483	0.832	0.636	0.515	0.453	0.759	0.576
All	0.760	0.594	0.727	0.694	0.782	0.632	0.933	0.782	0.851	0.733	0.834	0.806

Table 7: Cross-dataset finetuning performance of macro-F1 for three models for the specified dataset component.

4 Fully Supervised Finetuning

We now address **RQ1** and **RQ2** by conducting experiments with both traditional machine learning models and LLMs.

4.1 Traditional Machine Learning Models

We conduct experiments using popular stance classification methods. AlDayel and Magdy (2021b) identify SVM, CNN, and BiLSTM as leading machine learning methods for stance classification. Therefore, we evaluate the effectiveness of LLMG as a weakly supervised approach for real-world datasets. We adopt Word2Vec embeddings—a well-known embedding approach. For BiLSTM and CNN, we fine-tune the models with a learning rate of $2\mathrm{e}{-4}$, AdamW optimizer, 0.5 dropout, and CrossEntropy loss.

Table 4 presents the performance of classifying the three stance labels with and without LLMG as weak supervision. The results show that LLMG greatly improves the performance, particularly for CMV's Favor label and Debate's Against label, with average performance gains of 10% and 9%, respectively.

These improvements across diverse datasets and stance labels demonstrate that incorporating AI-generated data enhances stance classification generalizability (Ng and Carley 2022b), reaffirming the benefits of LLMs in real-world applications (Lee, Liang, and Yang 2022; Ravi et al. 2024).

Moreover, Table 5 presents Cohen's d effect sizes across the three models on two datasets with respect to three stance categories: Against, Favor, and None. The results suggest varying levels of statistical significance, where higher values correspond to stronger effects. The SVM and CNN models are averaged over ten-fold cross-validation, while BiLSTM was trained for ten epochs. Similarly, CNN and BiLSTM models show significant gain in performance, especially in the Favor category for both datasets, with effect sizes ranging from 0.78*** to 1.31** across models. Overall, these results indicate that incorporating LLMG (Language Model Guidance) is effective in improving stance classification, as evidenced by the strong effect sizes, particularly in distinguishing stances that favor a position.

4.2 Large Language Models

We use *SLM* (S for Small) for the previous generation of language models, such as BERT (Devlin et al. 2019), to contrast them with LLMs. A prevalent method for classification using SLMs involves *finetuning*, which entails exposing a pretrained SLM to domain-specific data. However, finetuning is not always the optimal method for customizing LLMs and some research have suggested it could be detrimental to performance. Moreover, whereas finetuning is tractable for SLMs, it demands substantial computational resources for LLMs. Therefore, we compare SLMs and LLMs for supervised finetuning for stance classification. For finetuning BERT, we concatenate the topic and argument with the special token [SEP] and prepend the sequence with the special token [CLS] to form the template [CLS] + Topic + [SEP] + Argument. A three way classification head is added on top of the token [CLS] to perform the classification task.

For finetuning the generative models T5 and LLaMa, and use the same template as in the Training Prompt (below). We show a few concrete examples in Appendix C. As for autoregressive pretraining, we apply the maximum likelihood estimation, which involves minimizing the cross-entropy loss between the predicted probability distribution of the next token and the actual token for the whole sequence. At inference time, we simply remove the gold label from the prompt so that the model can make a prediction. The output length is limited to two.

4.3 Experimental Setup

Our evaluation involves (1) BERT (Devlin et al. 2019), recognized for its effectiveness in classification, (2) T5 (Raffel et al. 2020), a generative counterpart to BERT, and (3)

Training Prompt for Generative Models.
Classify the stance of the argument towards the topic as either *favor*, *against*, or *neutral*. Return the label only without any other text. **Topic**: {topic} **Argument**: {argument} **Label**: {label}

LLaMa-7b (Touvron et al. 2023), a popular LLM. We conduct the following experiments.

Finetuning with a single dataset To evaluate generalizability in stance classification, we assess how a model trained on one dataset performs on another dataset.

Finetuning with multiple datasets We extend the above evaluation to include finetuning on combined datasets.

Finetuning with varied sizes of training data We evaluate the effect of data size (from combined data) on finetuning. For all datasets, we adopt the macro-F1 metric, namely, the average F1 score for each label category (Favor, Against, None). For BERT (110M) and T5 (250M), we perform finetuning with all parameters. For LLaMa-7b, we apply the QLoRA (Dettmers et al. 2023) quantization technique, updating only 20 million parameters. Table 6 shows how we split the data into train, dev, and test sets. The hyperparameter settings for all models are shown in Appendix B.

4.4 Results

We now present the results of our experiments. Table 7 shows the results for finetuning with a single dataset and with all of the three datasets. These results reveal a persistent challenge across all models: a difficulty in adapting to new datasets when subjected to finetuning with one dataset, indicating the subtle differences between domains.

Notably, the best average performance is achieved by LLaMa-7b fine-tuned on the Debate dataset. For finetuning with multiple datasets, LLaMa-7b is the model with the highest average F1 across all three datasets. Despite having fewer fine-tuned parameters (20M compared to 110M and 250M), LLaMa-7b outperforms its counterparts, reflecting the power of LLMs in complex tasks. Both T5 and LLaMa-7 beat BERT, highlighting the advantage of using generative models over classification-oriented models for stance classification.

We now describe our ablation studies. Figure 3 presents the results for different amounts of training data. The three models demonstrate comparable and high training sample efficiency. Notably, with approximately 25% of the training data, each model achieves nearly 95% of its optimal performance.

5 Zero-Shot and Few-Shot Benchmarking

The zero-shot and cross-topic variants of stance classification are well-aligned since both involve topics not encountered during training. To address **RQ3**, we evaluate strict zero-shot and few-shot learning. Research suggests that the knowledge that LLMs possess is predominantly acquired through pretraining (Cruickshank and Ng 2023). This implies that LLMs possess the inherent capacity to address various tasks, provided they are suitably instructed

5.1 Experimental Setup

We focus on open LLMs (Touvron et al. 2023), to enhance accessibility. LLMs exhibit a variety of architectures and sizes, and whether they underwent instruction tuning during their training process. We employ LLaMa as the cornerstone of our study, because of its demonstrated superiority across multiple tasks and performance that is competitive with ChatGPT. We consider the 7B, 13B, 33B, and 65B configurations of LLaMa, as well as the 7B, 13B, 33B configurations of its instruction-tuned counterpart, Vicuna (Chiang et al. 2023). We also include another model family, UL2 (Tay et al. 2023), and its instruction-tuned counterpart FLAN-UL2, which has an encoder-decoder architecture and the 7B and 40B configurations of the Falcon family, with and without instruction tuning.

We conduct experiments with zero-shot and few-shot in-context learning. For all the LLMs, we use QLoRA to quantize them to 4 bits to reduce the need for GPU memory. QLoRA suffers little loss on a variety of tasks (Dettmers et al. 2023). Our experiments are run on a mixture of NVIDIA-A100, NVIDIA-A30, NVIDIA-A10, and NVIDIA-A6000 GPUs.

5.2 Results

The overall results are shown in Table 8. The main findings are summarized as follows.

Significance test for model performance We performed McNemar's test to assess the significance of model prediction differences among the top-performing models. This test was conducted in two stages. First, we compared the top models from each family, namely LLaMA-65B, Vicuna-33B, Falcon-40B-instruct, and FLAN-UL2-20B. Second, we compared zero-shot and 9-shot performances within these four families. The corresponding p-values are presented in Table 9. As observed, the differences across both settings are statistically significant ($p < 0.001$) across three datasets, with the exception of LLaMA-65B vs. Vicuna-33B on the CMV and LLMG datasets, and LLaMA-65B vs. FLAN-UL2-20B on the CMV dataset. Notably, all models exhibited significant differences between the zero-shot and 9-shot conditions, highlighting the critical benefit of few-shot examples in performance improvement.

Gap with upper bound Overall, we observe positive effects for model scaling. For all model families, larger models yield better performances across most settings. However, the best performance of FLAN-UL2, which achieves 41.17, 51.51, and 50.21 under zero shot for CMV, Debate, and LLMG, respectively, falls far behind the supervised approach, which suggests difficulty for LLMs to comprehend downstream tasks.

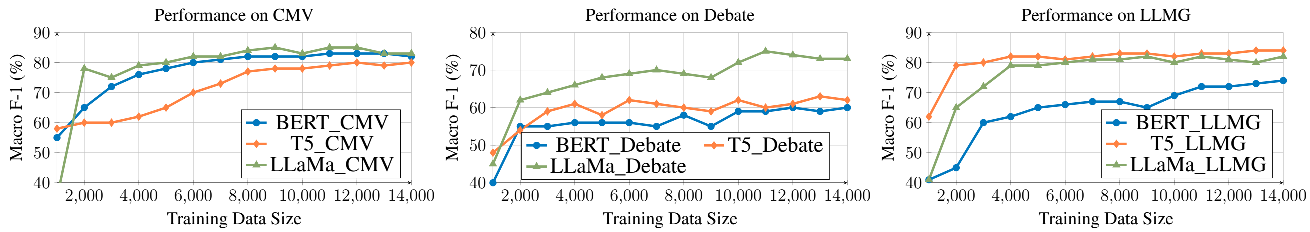


Figure 3: Performance of micro-F1 for BERT, T5, LLaMa on all partitions of the dataset with various training examples.

Models	**0-shot**			**3-shot**			**6-shot**			**9-shot**		
	CMV	**Debate**	**LLMG**	**CMV**	**Debate**	**LLMG**	**CMV**	**Debate**	**LLMG**	**CMV**	**Debate**	**LLMG**
LLaMA-7B	16.28	16.33	15.08	$17.26_{5.68}$	$29.07_{6.10}$	$24.45_{6.18}$	$23.46_{8.15}$	$33.80_{6.33}$	$29.32_{5.08}$	14.79	28.27	31.66
LLaMA-13B	9.68	20.48	17.19	$24.07_{7.56}$	$35.49_{5.08}$	$31.84_{8.08}$	$22.96_{6.99}$	$41.34_{1.34}$	$31.60_{3.79}$	12.55	40.84	30.50
LLaMA-30B	22.05	11.58	38.23	$32.63_{5.63}$	$42.68_{3.01}$	$42.00_{2.02}$	$28.68_{2.15}$	$44.06_{1.68}$	$37.03_{1.62}$	32.58	42.10	39.97
LLaMA-65B	33.32	36.96	**54.59**	$38.40_{4.71}$	$49.76_{2.38}$	$50.26_{3.31}$	$37.25_{5.69}$	$50.21_{2.24}$	$45.07_{2.55}$	41.76	50.73	53.65
Vicuna-7B	24.38	37.49	33.05	$19.83_{7.71}$	$28.92_{5.61}$	$29.94_{6.75}$	$15.95_{4.57}$	$30.63_{6.29}$	$30.70_{4.58}$	12.15	26.10	22.84
Vicuna-13B	25.59	30.15	47.11	$31.29_{6.27}$	$41.31_{6.47}$	$38.90_{7.47}$	$22.90_{2.23}$	$42.41_{1.22}$	$36.78_{4.19}$	14.95	42.56	40.59
Vicuna-33B	34.51	42.59	51.86	$31.70_{6.49}$	$47.54_{4.80}$	$41.40_{2.70}$	$29.92_{4.19}$	$48.94_{2.37}$	$38.50_{2.15}$	29.48	49.90	41.50
Falcon-7B	16.40	22.04	18.45	$27.58_{3.60}$	$30.33_{7.68}$	$30.80_{1.84}$	$26.70_{3.88}$	$34.40_{5.45}$	$32.55_{3.80}$	23.44	40.74	34.59
Falcon-7B-I	24.68	23.50	31.64	$24.42_{2.03}$	$23.70_{3.44}$	$20.81_{1.73}$	$10.67_{1.38}$	$20.29_{0.57}$	$18.08_{1.29}$	10.18	20.63	19.86
Falcon-40B	30.52	39.20	33.85	$24.73_{3.05}$	$34.54_{6.42}$	$28.87_{6.44}$	$22.01_{5.59}$	$32.67_{6.05}$	$29.22_{5.44}$	16.34	29.82	23.61
Falcon-40B-I	35.22	42.52	38.74	$29.87_{4.23}$	$35.07_{5.43}$	$32.74_{6.36}$	$27.60_{5.25}$	$33.07_{4.27}$	$31.44_{3.76}$	22.75	26.73	27.74
UL2-20B	31.93	37.32	36.38	$27.49_{3.22}$	$22.35_{3.26}$	$29.36_{8.94}$	$25.54_{1.87}$	$18.28_{4.57}$	$19.91_{4.23}$	22.59	11.67	17.01
FLAN-UL2-20B	**41.17**	**51.51**	50.21	$41.85_{0.84}$	$52.51_{0.29}$	$52.70_{0.55}$	$42.30_{0.63}$	$52.67_{0.31}$	$53.46_{0.62}$	42.93	52.60	54.28

Table 8: Zero-shot and few-shot in-context learning for various LLMs.

	CMV	Debate	LLMG
LLaMa-65B vs. Vicuna-33B	0.544	–	0.408
LLaMa-65B vs. Falcon-40B-I	–	–	–
LLaMa-65B vs. FLAN-UL2-20B	0.934	–	–
Vicuna-33B vs. Falcon-40B-I	–	–	–
Vicuna-33B vs. FLAN-UL2-20B	0.671	–	–
Falcon-40B-I vs. FLAN-UL2-20B	–	–	–
LLaMa-65B 0-shot vs. 9-shot	–	–	–
Vicuna-33B 0-shot vs. 9-shot	–	–	–
Falcon-40B-I 0-shot vs. 9-shot	–	–	–
FLAN-UL2-20B 0-shot vs. 9-shot	–	–	–

Table 9: McNemar's significance test. We show the p-values indicating nonsignificance and omitted all < 0.001.

Number of few-shot exemplars Exposing a model to more examples reliably improves performance across various tasks. However, our results are mixed. To understand variability, we randomly sampled 10 sets of examples for both 3-shot and 6-shot learning and calculated their mean and standard deviation. Some sets of examples show better performance than zero-shot. Nonetheless, the variability highlights the sensitivity of LLMs to specific examples. One exception is FLAN-UL2, the top performer, which maintains an average variance of 0.54, showcasing the consistency of its performance. Additionally, FLAN-UL2 demonstrates a robust improvement due to the increase in example input.

Impact of instruction tuning Instruction-tuning continually fine-tunes an LLM by exposing it to diverse instructions and their responses. Doing so enhances its ability to follow instructions. We observe that instruction-tuned models reliably outperform models of the same architecture and size that are not instruction-tuned. This is apparent by comparing Vicuna to LLaMa, Falcon-instruct to Falcon, and FLAN-UL2 to UL2. This observation highlights the effectiveness of instruction-tuning as a task-agnostic method for adapting LLMs to downstream tasks.

	CMV	Debate	LLMG
Against argument	39.54	60.24	35.01
Favor argument	54.93	69.75	47.31
Overall	44.17	65.01	40.87

Table 10: Percentage of using facts in argument for each dataset.

6 Discussion

We present a benchmark dataset compiled from three types of sources: a social media website, two debate websites, and arguments generated by large language models (LLMs). The resulting dataset covers a wide range of 4,498 topics, comprising 30,961 arguments distributed across 21 domains.

We demonstrate the usefulness of our dataset via three experimental approaches: fully supervised, zero-shot, and few-shot in-context learning with LLMs. Notably, our findings highlight the superior performance of generative models over classification models. LLMs, when used in a zero-shot scenario, demonstrate commendable performance, though

Facts Extraction Prompt.
Facts are objective statements that are verifiable. Arguments are subjective claims or positions. Given a topic and an argument, identify if the argument relies on any verifiable facts. Return the fact that the argument relies on. Return none if the argument does not rely on verifiable facts. Be concise in your response. **Topic**: {topic} **Argument**: {argument}

with a noticeable performance gap relative to the upper bound. Furthermore, instruction-tuned LLMs reliably outperform their non-instruction-tuned counterparts, emphasizing the effectiveness of instruction-tuning for adapting LLMs to downstream tasks.

Thus, our study establishes robust baselines for the created dataset and provides valuable insights that can guide the development of more generalized stance classification methods. This research not only advances our understanding of the performance dynamics among different learning approaches but also offers practical implications for optimizing the use of LLMs for stance classification. Next, we discuss the use of factual information in argument formulation and some limitations of our work.

6.1 Factual Information in Arguments

Using facts in arguments can improve their credibility and persuasiveness. Though stance classification is not concerned with factual accuracy, in real-world applications, ensuring the factuality of arguments generated by LLMs is essential for their responsible use.

We adopt an LLM not from the GPT family to analyze the prevalence of factual information in arguments. Specifically, we adopt Anthropic's *Claude 3.5 Sonnet*, one of the most advanced commercial models, with the facts extraction prompt.

Table 10 shows the percentages of arguments using facts in CMV, Debate, and LLMG. There is a clear difference between the use of facts in (the more casual) CMV discussions and formal debate arguments, where including facts is encouraged by the participants. Arguments produced by GPT-3 show a lower prevalence of facts, which may be due to limitations set by its built-in guardrails—for example, the model may avoid citing statistics or named entities unless explicitly prompted. Across all three datasets examined, the frequency of factual information in *favor* arguments is much higher than that in *against*.

6.2 Limitations and Future Work

This study faces several limitations. First, our proposed framework for the collection of diverse argumentative sentence pairs covering a variety of topics can be extended as needed to facilitate the collection of additional data. However, this framework is constrained by the types of sources from which stance labels can be extracted. While we investigate the utilization of LLMs to construct stance classification datasets, more sophisticated experiments would be beneficial for exploring the full potential of this approach.

Second, while our study examines stance classification from three types of sources, it is important to recognize that this task is applicable in a much broader array of contexts, such as news articles, tweets, and political discourse. Therefore, combining our dataset with other existing datasets from different domains could improve the generalizability of stance classification.

Third, while we focus on the adoption of GPT-3 for generating arguments, we do not directly compare LLMs for generating arguments. We defer comparative studies involving LLMs such as GPT-4, PaLM, or Claude to future research. Such an evaluation would enable more robust methods for benchmarking for stance classification and other social media problems.

Finally, future work could also consider the moral and social framing of arguments, especially in settings where responsibility attribution or ethical reasoning plays a central role. Prior work has shown that everyday arguments often reflect implicit moral norms (Xi and Singh 2023) and that blame assignment on social media is shaped by demographic and psychological factors (Xi and Singh 2024). Incorporating such dimensions could enhance the interpretability and real-world applicability of stance classification systems.

References

Al-Khatib, K.; Völske, M.; Syed, S.; Kolyada, N.; and Stein, B. 2020. Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness. In *58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 7067–7072. Association for Computational Linguistics.

AlDayel, A.; and Magdy, W. 2021a. Stance Detection on Social Media: State of the Art and Trends. *Information Processing & Management*, 58(4): 102597.

AlDayel, A.; and Magdy, W. 2021b. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4): 102597.

Allaway, E.; and McKeown, K. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8913–8931. Online: Association for Computational Linguistics.

Alshomary, M.; Chen, W.-F.; Gurcke, T.; and Wachsmuth, H. 2021. Belief-based generation of argumentative claims. *arXiv preprint arXiv:2101.09765*.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In Park, J. C.; Arase, Y.; Hu, B.; Lu, W.; Wijaya, D.; Purwarianti, A.; and Krisnadhi, A. A., eds., *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational*

Linguistics (Volume 1: Long Papers), 675–718. Nusa Dua, Bali: Association for Computational Linguistics.

Bar-Haim, R.; Bhattacharya, I.; Dinuzzo, F.; Saha, A.; and Slonim, N. 2017. Stance Classification of Context-Dependent Claims. In Lapata, M.; Blunsom, P.; and Koller, A., eds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 251–261. Valencia, Spain: Association for Computational Linguistics.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chen, S.; Khashabi, D.; Yin, W.; Callison-Burch, C.; and Roth, D. 2019. Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 542–557. Minneapolis, Minnesota: Association for Computational Linguistics.

Chern, I.-C.; Chern, S.; Chen, S.; Yuan, W.; Feng, K.; Zhou, C.; He, J.; Neubig, G.; and Liu, P. 2023. FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. arXiv:2307.13528.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Conforti, C.; Berndt, J.; Pilehvar, M. T.; Giannitsarou, C.; Toxvaerd, F.; and Collier, N. 2020a. STANDER: An Expert-Annotated Dataset for News Stance Detection and Evidence Retrieval. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4086–4101. Online: Association for Computational Linguistics.

Conforti, C.; Berndt, J.; Pilehvar, M. T.; Giannitsarou, C.; Toxvaerd, F.; and Collier, N. 2020b. Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1715–1724. Online: Association for Computational Linguistics.

Cruickshank, I. J.; and Ng, L. H. X. 2023. Use of large language models for stance classification. *arXiv preprint arXiv:2309.13734*.

Cruickshank, I. J.; Soofi, A.; and Ng, L. H. X. 2024. DIVERSE: A Dataset of YouTube Video Comment Stances with a Data Programming Model. arXiv:2403.03334.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; and Weston, J. 2024. Chain-of-Verification Reduces Hallucination in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 3563–3578. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Edwards, A.; Ushio, A.; Camacho-collados, J.; Ribaupierre, H.; and Preece, A. 2022. Guiding Generative Language Models for Data Augmentation in Few-Shot Text Classification. In Dragut, E.; Li, Y.; Popa, L.; Vucetic, S.; and Srivastava, S., eds., *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*, 51–63. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Ferreira, W.; and Vlachos, A. 2016. Emergent: A novel dataset for stance classification. In Knight, K.; Nenkova, A.; and Rambow, O., eds., *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1163–1168. San Diego, California: Association for Computational Linguistics.

Glandt, K.; Khanal, S.; Li, Y.; Caragea, D.; and Caragea, C. 2021. Stance Detection in COVID-19 Tweets. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1596–1611. Online: Association for Computational Linguistics.

Gorrell, G.; Kochkina, E.; Liakata, M.; Aker, A.; Zubiaga, A.; Bontcheva, K.; and Derczynski, L. 2019. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In May, J.; Shutova, E.; Herbelot, A.; Zhu, X.; Apidianaki, M.; and Mohammad, S. M., eds., *Proceedings of the 13th International Workshop on Semantic Evaluation*, 845–854. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Guo, Z.; and Singh, M. P. 2023. Representing and Determining Argumentative Relevance in Online Discussions: A General Approach. In *Proceedings of the 17th International AAAI Conference on Web and Social Media (ICWSM)*, 292–302. Limassol, Cyprus: AAAI Press.

Guo, Z.; Zhang, Z.; and Singh, M. P. 2020. In Opinion Holders' Shoes: Modeling Cumulative Influence for View

Change in Online Argumentation. In *Proceedings of the 29th Web Conference (WWW)*, 2388–2399. Taipei: ACM.

Habernal, I.; Wachsmuth, H.; Gurevych, I.; and Stein, B. 2018. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1930–1940. New Orleans, Louisiana: Association for Computational Linguistics.

Hanselowski, A.; Stab, C.; Schulz, C.; Li, Z.; and Gurevych, I. 2019. A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking. In Bansal, M.; and Villavicencio, A., eds., *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 493–503. Hong Kong, China: Association for Computational Linguistics.

Hasan, K. S.; and Ng, V. 2014. Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 751–762. Doha, Qatar: Association for Computational Linguistics.

Küçük, D.; and Can, F. 2020. Stance Detection: A Survey. *ACM Computing Surveys*, 53(1).

Lee, M.; Liang, P.; and Yang, Q. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, USA: Association for Computing Machinery. ISBN 9781450391573.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In Knight, K.; Nenkova, A.; and Rambow, O., eds., *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. San Diego, California: Association for Computational Linguistics.

Li, Y.; Sosea, T.; Sawant, A.; Nair, A. J.; Inkpen, D.; and Caragea, C. 2021. P-Stance: A Large Dataset for Stance Detection in Political Domain. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2355–2365. Online: Association for Computational Linguistics.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In Bethard, S.; Carpuat, M.; Cer, D.; Jurgens, D.; Nakov, P.; and Zesch, T., eds., *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41. San Diego, California: Association for Computational Linguistics.

Murakami, A.; and Raymond, R. 2010. Support or Oppose? Classifying Positions in Online Debates from Reply Activities and Opinion Expressions. In Huang, C.-R.; and Jurafsky, D., eds., *Coling 2010: Posters*, 869–875. Beijing, China: Coling 2010 Organizing Committee.

Ng, L. H. X.; and Carley, K. M. 2022a. Is my stance the same as your stance? A cross validation study of stance detection datasets. *Information Processing & Management*, 59(6): 103070.

Ng, L. H. X.; and Carley, K. M. 2022b. Is my stance the same as your stance? A cross validation study of stance detection datasets. *Information Processing & Management*, 59(6): 103070.

Park, C.; Yang, W.; and Park, J. 2019. Generating Sentential Arguments from Diverse Perspectives on Controversial Topic. In Feldman, A.; Da San Martino, G.; Barrón-Cedeño, A.; Brew, C.; Leberknight, C.; and Nakov, P., eds., *Proceedings of the Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 56–65. Hong Kong, China: Association for Computational Linguistics.

Qazvinian, V.; Rosengren, E.; Radev, D. R.; and Mei, Q. 2011. Rumor has it: Identifying Misinformation in Microblogs. In Barzilay, R.; and Johnson, M., eds., *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1589–1599. Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21(1).

Ravi, S.; Huber, P.; Shrivastava, A.; Shwartz, V.; and Einolghozati, A. 2024. Small But Funny: A Feedback-Driven Approach to Humor Distillation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13078–13090. Bangkok, Thailand: Association for Computational Linguistics.

Sahu, G.; Rodriguez, P.; Laradji, I.; Atighehchian, P.; Vazquez, D.; and Bahdanau, D. 2022. Data Augmentation for Intent Classification with Off-the-shelf Large Language Models. In Liu, B.; Papangelis, A.; Ultes, S.; Rastogi, A.; Chen, Y.-N.; Spithourakis, G.; Nouri, E.; and Shi, W., eds., *Proceedings of the 4th Workshop on NLP for Conversational AI*, 47–57. Dublin, Ireland: Association for Computational Linguistics.

Shafayat, S.; Kim, E.; Oh, J.; and Oh, A. 2024. Multi-FAct: Assessing Multilingual LLMs' Multi-Regional Knowledge using FActScore. *arXiv preprint arXiv:2402.18045*.

Sobhani, P.; Inkpen, D.; and Matwin, S. 2015. From Argumentation Mining to Stance Classification. In Cardie, C., ed., *Proceedings of the 2nd Workshop on Argumentation Mining*, 67–77. Denver, CO: Association for Computational Linguistics.

Sobhani, P.; Inkpen, D.; and Zhu, X. 2017. A Dataset for Multi-Target Stance Detection. In Lapata, M.; Blunsom, P.; and Koller, A., eds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 551–557. Valencia, Spain: Association for Computational Linguistics.

Somasundaran, S.; and Wiebe, J. 2010. Recognizing Stances in Ideological On-Line Debates. In Inkpen, D.; and Strapparava, C., eds., *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 116–124. Los Angeles, CA: Association for Computational Linguistics.

Stab, C.; Miller, T.; Schiller, B.; Rai, P.; and Gurevych, I. 2018. Cross-topic Argument Mining from Heterogeneous Sources. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3664–3674. Brussels, Belgium: Association for Computational Linguistics.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-Faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, 613–624. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450341431.

Tay, Y.; Dehghani, M.; Tran, V. Q.; Garcia, X.; Wei, J.; Wang, X.; Chung, H. W.; Shakeri, S.; Bahri, D.; Schuster, T.; Zheng, H. S.; Zhou, D.; Houlsby, N.; and Metzler, D. 2023. UL2: Unifying Language Learning Paradigms. arXiv:2205.05131.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, abs/2302.13971.

Villa-Cox, R.; Kumar, S.; Babcock, M.; and Carley, K. M. 2020. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. *arXiv preprint arXiv:2006.00691*.

Walker, M.; Tree, J. F.; Anand, P.; Abbott, R.; and King, J. 2012. A Corpus for Research on Deliberation and Debate. In Calzolari, N.; Choukri, K.; Declerck, T.; Doğan, M. U.; Maegaard, B.; Mariani, J.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 812–817. Istanbul, Turkey: European Language Resources Association (ELRA).

Xi, R.; and Singh, M. P. 2023. The Blame Game: Understanding Blame Assignment in Social Media. *IEEE Transactions on Computational Social Systems*, 10: 1–10.

Xi, R.; and Singh, M. P. 2024. Morality in the Mundane: Categorizing Moral Reasoning in Real-life Social Situations. In *Proceedings of the 18th International AAAI Conference on Web and Social Media (ICWSM)*, 1–12. Buffalo, New York: AAAI Press. To appear.

Yoo, K. M.; Park, D.; Kang, J.; Lee, S.-W.; and Park, W. 2021. GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2225–2239. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Yuan, J.; and Singh, M. P. 2023. Conversation Modeling to Predict Derailment. In *Proceedings of the 17th International AAAI Conference on Web and Social Media (ICWSM)*, 926–935. Limassol, Cyprus: AAAI Press.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? No, the dataset doesn't involve population-specific distributions
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? No
   (g) Did you discuss any potential misuse of your work? No
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? No
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, all material will be published later
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? No, no previous dataset is used.
   (b) Did you mention the license of the assets? No, no license is needed. The dataset will be open
   (c) Did you include any new assets in the supplemental material or as a URL? No
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, no personally identity is involved
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? No
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? No
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

Broader Impact

Stance classification focuses on identifying the position of an argument on a specific topic, not its factual correctness. Text generated by LLMs or online users sometimes includes inaccurate information. This concern is relevant when our dataset is used for downstream applications—such as generating arguments for persuasive writing or educational content—where factual correctness is critical (e.g., health or policy discussions). In such cases, factuality should be verified before deployment.

Arguments containing unverified information in our dataset may lead to the creation of models that generate unreliable outputs, which can contribute to the spread of misinformation (Bang et al. 2023). Existing research on verifying factual information falls into two primary approaches. One approach focuses on extracting atomic facts from content generated by LLMs and validating these facts against external references, such as Wikipedia or the web (Chern et al. 2023) results. Another approach relies on the model's internal mechanisms, employing strategies such as chain-of-thought reasoning to deliberate on the generated responses and implement self-correction (Dhuliawala et al. 2024).

The risk of misinformation highlights the necessity for the establishment of rigorous evaluation metrics to assess the factual accuracy of AI-generated text (Shafayat et al. 2024).

A Websites for collecting topics

1. https://blog.kialo-edu.com/lesson-ideas/classroom-debate-ideas/
2. https://www.weareteachers.com/controversial-debate-topics/
3. https://research.com/education/debate-topics-for-college-students
4. https://www.theedadvocate.org/political-debate-topics/
5. https://owlcation.com/academia/100-Debate-Topics
6. https://noisyclassroom.com/debate-topics/
7. https://www.myspeechclass.com/controversial-speech-topics.html
8. https://custom-writing.org/blog/debate-topics
9. https://www.5staressays.com/blog/speech-and-debate/debate-topics

B Experimental Hyperparameters

We provide the main hyperparameter settings for data generation with GPT-3, and the training of BERT, T5, and LLaMa below.

B.1 Argument Generation with GPT-3

- temperature: 0.7
- max_generated_tokens: 500

B.2 BERT Finetuning

- Number of training epochs: 10
- Learning rate: 5e-5
- Batch size: 16
- Optimizer: Adamw

B.3 T5 Finetuning

- Number of training epochs: 10
- Learning rate: 2e-5
- Batch size: 16
- Optimizer: Adamw

B.4 LLaMa-7B Finetuning

- Number of training epochs: 3
- LoRA rank: 8
- LoRA alpha: 16
- LoRA dropout: 0.05
- LoRA bias: none
- Target modules: q_proj, v_proj, o_proj, k_proj, up_proj, down_proj, gate_proj, embed_tokens, lm_head
- Learning rate: 2e-4
- Batch size: 64
- Warmup steps: 10
- Optimizer: paged_adamw_8bit
- max_input_length: 256

B.5 LLaMa-7B Inference

We perform inference with the fine-tuned LLaMa-7B in 4-bit with BitsAndBytes:

- load_in_4bit: True
- bnb_4bit_quant_type: nf4
- bnb_4bit_use_double_quantL: True
- bnb_4bit_compute_type: bfloat16
- max_generation_length: 2

C Training Examples

We shown a few more randomly chosen examples from each datasets of the benchmark for training in Table 11.

D Complementary Results

In this section, we provide complementary results for Table 7. Specifically, per-class F1 score for each finetuning setting is shown in Table 12, Table 13, and Table 14 for BERT, T5, and LLaMa, respectively.

Topical Claim	Comment	Stance
Transgender (specifically mtf) athletes have an unfair advantage in strength-based sports and should only be able to participate in their biological gender group	This is in response to the post that hit the front page recently which can be seen here: Biologically speaking, the male gender is already predisposed to stronger traits and as a response the Olympic Committee split competitions into two groups so that both genders have an ëqual playing field. But with more acceptance and social tolerance, along with more transgender people coming out, problems have arose and will continue to rise until this gets dealt with and, to me, it's unfair and should be given strict guidelines and have no gender fluidity in the eyes of competitors.	Favor
I feel that wealthy should pay a flat tax rather than the current progressive us tax system	I won't try to convince you that you're wrong about tax rates because I don't know enough about it. Just consider: doesn't it ever bother you that people just pick up and leave for the sake of a little more money? These people have no ties their location, or they have ties that they're willing to throw away for a cheaper house. It's more of an admirable goal to build those deep, binding ties. Approaching the world with love, making friends, and investing in a community will lead to a more satisfying life than approaching it with cold logic centered around money. Just picture the end of your life for a moment. Will you be happier if you kept more of your money for yourself? Or will you be happier if you're a respected member of a community? Find a place that's worth the higher taxes. Find a place you could fall in love with	Against
It is not okay to keep animals in zoos.	Education and conservation efforts can be achieved through alternative means, such as wildlife sanctuaries and educational programs. These alternatives allow for a more ethical approach to animal welfare, as they focus on providing a natural environment for the animals while still educating the public about conservation and wildlife protection.	Favor
The drinking age should be 18.	The human brain continues to develop until the mid-20s, particularly the prefrontal cortex responsible for decision-making and impulse control. Allowing 18-year-olds to consume alcohol may expose them to potential harm and hinder their brain development. Raising the drinking age to 21 provides additional years for the brain to mature and reduces the risk of long-term negative consequences associated with alcohol consumption.	Against
Government surveillance is essential for national security.	In an increasingly interconnected world, a country's security can be threatened by various factors beyond direct self-defense. Issues such as terrorism, cyber attacks, and the spread of weapons of mass destruction pose significant risks that may require proactive military action. Waiting until an attack occurs could result in catastrophic consequences that could have been prevented.	None

Table 11: More examples from our benchmark.

	CMV		**Debate**		**LLMG**	
	Against	Favor	Against	Favor	Against	Favor
CMV	0.738	0.840	0.559	0.219	0.622	0.236
Debate	0.658	0.845	0.733	0.421	0.660	0.547
LLMG	0.272	0.573	0.357	0.475	0.631	0.769
ALL	0.719	0.747	0.734	0.443	0.728	0.701

Table 12: Per class F1 score for BERT finetuning.

	CMV		**Debate**		**LLMG**	
	Against	Favor	Against	Favor	Against	Favor
CMV	0.664	0.724	0.558	0.249	0.687	0.582
Debate	0.440	0.563	0.623	0.477	0.840	0.786
LLMG	0.451	0.581	0.427	0.512	0.820	0.902
ALL	0.668	0.722	0.634	0.522	0.952	0.900

Table 13: Per class F1 score for T5 finetuning.

	CMV		**Debate**		**LLMG**	
	Against	Favor	Against	Favor	Against	Favor
CMV	0.788	0.880	0.451	0.429	0.622	0.664
Debate	0.625	0.561	0.683	0.551	0.833	0.801
LLMG	0.461	0.553	0.441	0.452	0.828	0.792
ALL	0.882	0.852	0.724	0.643	0.853	0.821

Table 14: Per class F1 score for LLaMa finetuning.