ZeroStance: Leveraging ChatGPT for Open-Domain Stance Detection via Dataset Generation

Chenye Zhao[*♦] **Yingjie Li**[*♣] **Cornelia Caragea**[♦] **Yue Zhang**[♣]

[♦]Computer Science, University of Illinois Chicago
[♣]School of Engineering, Westlake University
[♦]{czhao43,cornelia}@uic.edu
[♣]{liyingjie,zhangyue}@westlake.edu.cn

Abstract

Zero-shot stance detection that aims to detect the stance (typically against, favor or neutral) towards unseen targets has attracted considerable attention. However, most previous studies only focus on targets from a single or limited text domains (e.g., financial domain), and thus zero-shot models cannot generalize well to unseen targets of diverse domains (e.g., political domain). In this paper, we consider a more realistic task, i.e., *open-domain stance detection*, which aims at training a model that is able to generalize well to unseen targets across multiple domains of interest. Particularly, we propose a novel dataset generation method *ZeroStance*, which leverages ChatGPT to construct a synthetic open-domain dataset *CHATStance* that covers a wide range of domains. We then train an open-domain model on our synthetic dataset after proper data filtering. Extensive results indicate that our model, when trained on this synthetic dataset, shows superior generalization to unseen targets of diverse domains over baselines on most benchmarks. Our method requires only a task description in the form of a prompt and is much more cost-effective and data-efficient than previous methods. Our code and data are available at https://github.com/chenyez/ZeroStance.

1 Introduction

The task of stance detection is to identify the attitude (e.g., favor, against or neutral, etc.) of a given text with respect to a specific target of interest (Küçük and Can, 2020; AlDayel and Magdy, 2020; Xu et al., 2022; Li et al., 2023a; Zhao et al., 2023; Liu et al., 2023; Arakelyan et al., 2023; Ko et al., 2023). Until recently, typical stance detection task was in-domain (Mohammad et al., 2016; Li and Caragea, 2019; Siddiqua et al., 2019; Li et al., 2021b; Upadhyaya et al., 2023; Li and Caragea, 2023) in which the training and test sets share the same set of targets. Most recent works began considering a cross-domain setup, i.e., cross-target stance detection (Augenstein et al., 2016; Xu et al., 2018; Wei and Mao, 2019; Zhang et al., 2020; Liang et al., 2021) where models are trained on labeled data of a training target and tested on a destination target that is unseen during training. However, cross-target task requires human knowledge about any destination target and how it is related to the training target (Allaway and McKeown, 2020), which limits models' ability to generalize to a wide variety of unseen targets. Hence, there has been an emerging trend to explore zero-shot stance detection (Allaway and McKeown, 2020; Liang et al., 2022b; Li et al., 2023b), which aims to determine the stance towards unseen targets at the inference stage without requiring human knowledge about unseen targets or their relation to training targets.

Recent studies (Allaway et al., 2021; Liu et al., 2021; Liang et al., 2022a,b) often conduct zero-shot evaluations with training and unseen destination targets originating from a single or limited domains, and thus zero-shot models perform poorly on out-of-domain data. For instance, the WT-WT dataset (Conforti et al., 2020) exclusively comprises targets within the financial domain for both training and evaluation sets, leading to poor performance of models trained on it when applied to political domain data. Xu et al. (2022) investigate *open-domain stance detection* which aims to train a model that can generalize well to unseen targets of multiple domains. However, their approach still relies on the texts and targets of existing datasets, and thus limits its applicability to a narrow spectrum of domains.

To address the limitations of prior works, we propose a novel open-domain dataset generation approach, *ZeroStance*, which aims to promote model's generalization to unseen targets of diverse domains by using external knowledge from pow-

[*]Both authors contributed equally to this research.

Findings of the Association for Computational Linguistics ACL 2024, pages 13390–13405

erful large language models (LLMs). Motivated by the remarkable success of LLMs such as ChatGPT in text generation (Brown et al., 2020; Ouyang et al., 2022; Min et al., 2023; Zhou et al., 2023), we present *CHATStance*, a synthetic dataset created through *ZeroStance* that leverages ChatGPT to generate high-quality, human-like targets (controversial claims) and texts spanning a variety of domains. We take into account the attribute diversity during the generation by incorporating desired attributes (including domains, geographical locations, and writing styles) as constraints in the prompts. By comparing the performance of *ZeroStance* with previous baselines, we observe a substantial underperformance of the latter, uncovering the effectiveness of our approach for open-domain stance detection. Table 1 shows a generated sample from our *CHATStance* dataset.

Our contributions are summarized as follows:

- We propose a novel dataset generation approach *ZeroStance* for open-domain stance detection that greatly improves the zero-shot performance by improving the data diversity and requires no training data (text or target) of existing datasets.
- Extensive results on six stance datasets show that the model, when trained on our synthetic open-domain dataset, demonstrates better generalization to unseen targets of diverse domains over models trained on human-annotated datasets (with or without data augmentation).
- We present an open-domain dataset *CHATStance*, which is much more data-efficient and cost-effective. Notably, the model trained on only 33% (around 7k instances) of the *CHATStance* outperforms the model that is trained with more than 70k instances from human-annotated datasets. Moreover, the total cost of creating *CHATStance* is only around $3, making it over two thousand times cheaper than human annotation.

Target:	The use of child soldiers in warfare is a serious human rights violation.
Text:	Child soldiers are robbed of their childhoods and forced to engage in violence. They are often subjected to physical and psychological abuse as well as exploitation. The use of child soldiers perpetuates cycles of violence and contributes to ongoing conflicts. Children who are recruited into armed groups are also denied the right to an education and access to basic healthcare needs.
Stance:	Favor

Table 1: An example of *CHATStance*.

2 Related Work

2.1 Zero-shot Stance Detection

Zero-shot stance detection that aims to detect the stance toward completely unseen targets has drawn considerable attention in recent years. Allaway and McKeown (2020) introduce a dataset for zero-shot stance detection and propose a target-grouped attention method to implicitly capture the relationships between targets. Later, Allaway et al. (2021) extract target-invariant features to generalize across topics using adversarial learning. Meanwhile, Liu et al. (2021) exploit the structural-level and semantic-level information to strengthen the generalization abilities of zero-shot models. Contrastive learning has also shown its effectiveness by generalizing the target-based stance features to unseen targets or improving the quality of augmented data (Liang et al., 2022a,b; Li and Yuan, 2022). Li et al. (2023b) propose a teacher-student framework that leverages generated keyphrases as augmented targets to improve the performance of zero-shot models. However, most previous works only perform training and evaluations on targets from a single domain (Mohammad et al., 2016; Conforti et al., 2020) or limited domains (Allaway and McKeown, 2020). Models trained on targets of these domains cannot generalize well to unseen targets of more diverse domains. In contrast, we tackle the task from the perspective of data quality by proposing a dataset generation method, which is *orthogonal* to most existing methods.

2.2 Data Augmentation

Data augmentation has been widely used to boost the model performance. A common data augmentation method is word replacement. Wang and Yang (2015) replace words with their top-n similar words based on pre-trained word embeddings (Mikolov et al., 2013). Similarly, Zhang et al. (2015) replace words with their synonyms in WordNet (Miller, 1995). Wei and Zou (2019) present random word replacement techniques that consist of four types of operations (e.g., synonym replacement, random deletion, etc). Wu et al. (2019) utilize segment embeddings of BERT (Devlin et al., 2019) as class indicators to generate augmented samples by predicting the masked word with the masked language

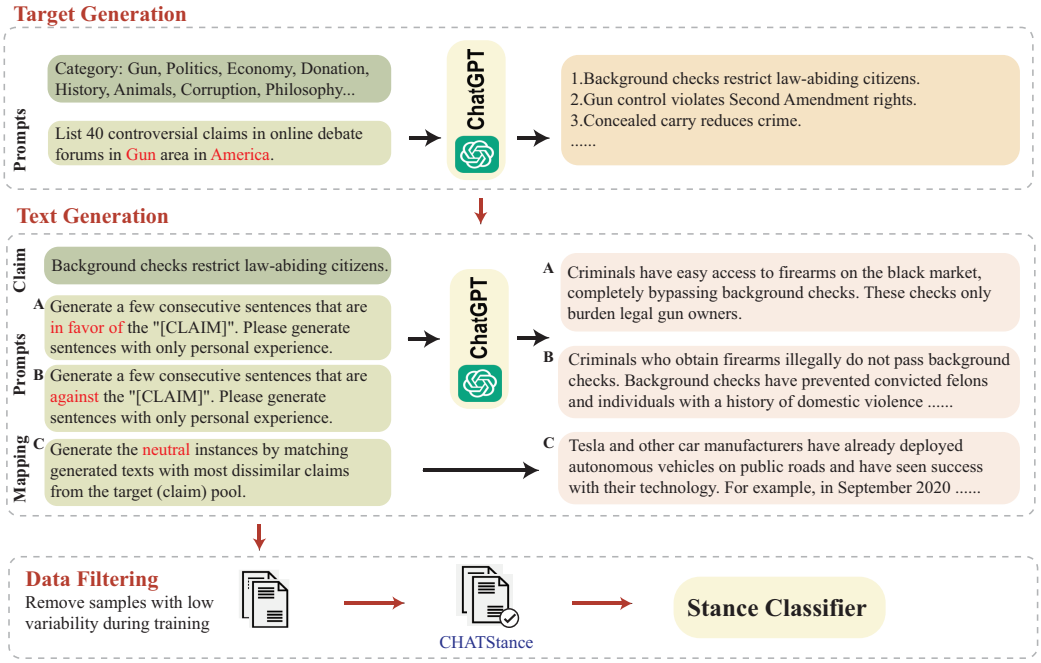

Figure 1: The overall framework of *ZeroStance*.

modeling (MLM) objective. Another commonly used strategy for augmentation is back-translation (Yu et al., 2018), which translates the training sample from one language (e.g., English) to another (e.g., French) and then translates it back to the original language. In stance detection, Li and Caragea (2021) generate samples by performing MLM with target and label information. Li and Yuan (2022) propose a data augmentation framework for that generates synthetic training data for unseen targets by adversarial learning and contrastive learning. Xu et al. (2022) design a masking mechanism to generate data with pre-trained language models. Li et al. (2023b) propose a target augmentation method that augments targets from the original training set and exploited a teacher-student framework to improve the model performance. However, the aforementioned data augmentation methods usually rely on existing datasets to generate augmented samples, which limits the generated samples to a restricted set of domains. Thus, zero-shot models trained with augmented data cannot generalize well to diverse domains. In this paper, inspired by the recent success of LLMs (Dai et al., 2023; Ubani et al., 2023; Zhu et al., 2023; Kuzman et al., 2023; Huang et al., 2023; Gilardi et al., 2023), we construct an open-domain dataset *CHATStance* that covers a wide range of domains using ChatGPT. Our proposed method requires no training data from existing datasets, but a task description of stance detection.

3 Method

In this section, we introduce the implementation of *ZeroStance* which includes three steps: *target generation*, *text generation* and *data filtering*. During target generation, ChatGPT is used to generate relevant targets given the category information extracted from the debate forum. For text generation, generated targets and stance labels are fed to ChatGPT in the form of prompts to generate corresponding texts. Then, we perform data filtering to remove potentially noisy and less informative data. The overall framework of *ZeroStance* is shown in Figure 1.

3.1 Problem Formulation

For zero-shot stance detection, we have two disjoint sets of targets: S for seen targets t^s and U for unseen targets t^u. Suppose a given training set $D^s = \{(x_i^s, t_i^s, y_i^s)\}_{i=1}^{N_s}$ and a test set $D^u = \{(x_i^u, t_i^u)\}_{i=1}^{N_u}$, where x_i^s is a sequence of words, t_i^s is the corresponding target and $y_i^s \in$ {Against, Favor, Neutral} is the stance label. The objective of zero-shot stance detection task is to predict the stance y_i^u given a document x_i^u and an unseen target t_i^u based on a model that is trained on documents x^s and seen targets t^s in D^s. Note that documents x^u and unseen targets t^u do not necessarily come from the same or similar domains of training set D^s.

Example 1	**Category:** Politics
America:	Police brutality towards people of color is not a widespread problem.
Europe:	The UK's decision to leave the European Union will lead to economic prosperity.
Asia:	The Singapore government is too heavy-handed in regulating free speech.
Example 2	**Category:** Health
America:	Marijuana should be legalized for medicinal purposes.
Europe:	Genetically modified foods are safe to consume.
Asia:	Traditional Chinese medicine is more effective than modern medicine.

Table 2: Examples of generated claims.

3.2 Target Generation

First, we create a high-quality seed list by extracting pre-defined generic categories from *kialo*[1], which is a structured online debate platform where users provide supporting and opposing claims for each claim related to a controversial topic. *Kialo* includes diverse set of controversial claims that are tagged under pre-defined generic categories such as *Politics*, *Racism*, *Music*, *Society* and *Economy*. The complete seed list is shown in Appendix A.

Second, motivated by the recent success of LLMs on data annotation (Zhu et al., 2023; Kuzman et al., 2023; Huang et al., 2023; Gilardi et al., 2023) and data augmentation (Dai et al., 2023; Ubani et al., 2023), we leverage ChatGPT to generate controversial claims for each category from the seed list with the prompt "*List 40 controversial claims in online debate forums in [CAT] area in [GEO]*", where *[CAT]* and *[GEO]* are special tokens that are replaced with a real category from the seed list and a geographic region (America, Europe or Asia), respectively. Since controversial topics of different geographic regions are usually different for the same category (e.g., for the category *Politics*, the controversial topics could be "police brutality towards people of color" in America and "regulating free speech in Singapore" in Asia), we generate claims for each region to cover more diverse topics. At the end, we generate a pool of around 24k controversial claims. Examples of generated claims are shown in Table 2.

3.3 Text Generation

After target generation, we further apply ChatGPT for text generation. We input a stance label y_i^s (favor or against) and a generated claim t_i^s from the previous step into ChatGPT and prompt ChatGPT to generate the corresponding text x_i^s. Previous works and datasets have shown that people usually express their stance towards specific topics by providing examples (Mohammad et al., 2016), sharing individual experience (Liu and Fahmy, 2011) or discussing relevant topics (Li et al., 2021a). Therefore, we consider three different prompts for text generation, which well represent how people express their stance towards controversial topics. For example, the prompt of sharing individual experience is "*Generate a few consecutive sentences that are [STANCE] the "[CLAIM]". Please generate sentences with only personal experience.*", where *[STANCE]* and *[CLAIM]* are stance label (in favor of or against) and generated claim, respectively. All three prompts used in text generation are shown in Appendix B. We randomly sample 3k claims from the target pool for each prompt and generate 18k instances in terms of three prompts and two stance labels (against and favor). Examples of generated texts are shown in Table 3. More examples are shown in Appendix B.

Note that we only generate *against* and *favor* instances, and do not generate *neutral* instances with ChatGPT since ChatGPT may not well understand what a neutral stance is and it is even quite challenging for humans to write a neutral opinion towards a topic in the task of stance detection (Li et al., 2021a). Therefore, similar to previous work (Allaway and McKeown, 2020), we automatically generate neutral instances by matching generated texts with most dissimilar claims from the target pool that are obtained based on the cosine-similarity of sentence embeddings (Reimers and Gurevych, 2019). We generate and sample 3k neutral instances for each prompt type and construct a dataset of 27k synthetic instances in total.

3.4 Data Filtering

Swayamdipta et al. (2020) has shown that training dynamics (i.e., variability and confidence) can be used to improve the out-of-distribution (OOD) performance. *Variability* is defined as follows:

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E} \left(p_{\theta^{(e)}}(y_i^* | x_i) - \hat{\mu}_i\right)^2}{E}} \quad (1)$$

where $p_{\theta^{(e)}}$ denotes the model's probability with parameter $\theta^{(e)}$ at the end of the e^{th} epoch, x_i and y_i^* denote the input instance and its label, respec-

[1] https://www.kialo.com/tags

Target:	Genetic modification causes cancer.	**Prompt:** Share personal experience.
Text (Against):	Growing up, my grandfather had a genetically modified heart valve and he lived a long and healthy life without any issues related to cancer. I remember eating genetically modified foods as a child and have never experienced any negative health consequences. My friend, who has a genetically modified immune system, has not developed cancer and is leading a fulfilling and healthy life.	
Text (Favor):	As a child, my family bought GMO produce from the grocery store without realizing it. As we transitioned to non-GMO, organic produce, we noticed significant improvements in our health and overall well-being. It wasn't until years later that we learned about the potential link between GMOs and cancer.	

Table 3: Examples of generated texts given claims and stance labels.

tively, E is the total number of training epochs and confidence $\hat{\mu_i}$ is defined as below.

$$\hat{\mu_i} = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}(y_i^*|x_i) \quad (2)$$

Training instances are classified into three types: 1) High variability instances exhibit frequent fluctuations in true class probabilities, introducing ambiguity and are informative for enhancing models' OOD performance. 2) High confidence, low variability instances indicate easy-to-learn scenarios with consistently reliable and correct predictions, offering limited value for models' further improvement. 3) Low confidence, low variability instances represent hard-to-learn scenarios with consistent prediction errors, often due to mislabeling during annotation (Swayamdipta et al., 2020).

In this paper, to enhance model generalization to unseen domains, we adopt the approach of Swayamdipta et al. (2020) by excluding instances with low *variability*, covering both easy-to-learn (less informative) and hard-to-learn (potentially noisy) cases. We calculate the *variability* for each instance after training and remove p percent of instances with the lowest *variability* from the training dataset. Consequently, we develop *CHATStance*, a stance detection dataset for open domains. Details on *CHATStance*'s data statistics after filtering are available in Appendix C.

4 Experimental Settings

In this section, we first introduce the baseline human-annotated datasets (§4.1). Then we discuss our evaluation setup (§4.2). Last, we describe baselines used in our experiments (§4.3).

4.1 Human-Annotated Datasets

covid COVID-19-Stance (Glandt et al., 2021) consists of tweets related to four targets in the COVID-19 domain: *Anthony Fauci*, *Stay-at-Home Orders*, *Wear a Face Mask*, and *Keeping School Closed*.

pstance The P-STANCE dataset (Li et al., 2021a) consists of tweets from the political domain. The dataset includes three targets: *Donald Trump*, *Joe Biden*, and *Bernie Sanders*.

sem16 The SemEval-2016 Task 6 dataset (Mohammad et al., 2016) is composed of tweet-target pairs centered around five targets, namely *Atheism*, *Feminist Movement*, *Legalization of Abortion*, *Hillary Clinton*, and *Climate Change is a Real Concern*.

wtwt The Will-They-Won't-They dataset (Conforti et al., 2020) consists of a large number of annotated tweet-target pairs from the financial domain, including five merger and acquisition operations (e.g., *Merger of CVS Health and Aetna*).

ibm30k The IBM-Rank-30k dataset (Gretz et al., 2020) includes 30k annotated text-target pairs. Targets encompass 71 selected controversial topics (e.g., *We should abolish capital punishment*).

vast The VAST dataset (Allaway and McKeown, 2020) includes news comments from the The New York Times *Room for Debate* section that contains a large number of targets from multiple domains.

The dataset split are presented in Appendix D. We also analyze the data contamination between the *CHATStance* and the human-annotated datasets, as presented in Appendix E.

4.2 Evaluation Setup

Following previous work (Hardalov et al., 2021), we consider two evaluation setups for baselines: (1) no training, i.e., unsupervised baselines that make stance predictions directly without requiring any training data; (2) out-of-domain, i.e., leave-one-dataset-out training on existing human-annotated datasets. Specifically, for six evaluation datasets, we leave one dataset out as the target dataset and take the rest five datasets as source datasets. We train and validate models using training and validation sets of source datasets and test them on the test set of the target dataset. To evaluate our pro-

posed *ZeroStance*, we train and validate the model on *CHATStance* dataset and test it on test sets of six human-annotated datasets. In adherence to the zero-shot setup, we exclude any targets (claims) from *CHATStance* that match targets present in the evaluation datasets. Like prior works (Allaway and McKeown, 2020; Glandt et al., 2021), we employ the macro-averaged F1 across all stance classes as our evaluation metric.

4.3 Baselines

Unsupervised Baselines. We consider five baselines that can work on the unsupervised scheme, which are applied directly to make stance predictions. **Random Guess** is a baseline that randomly predicts the stance label. **GPT2** (Radford et al., 2019) infers the stance label using the document's perplexity. **BART-MNLI** and **RoBERTa-MNLI** apply BART (Lewis et al., 2020) and RoBERTa (Liu et al., 2019) pre-trained on the MNLI dataset (Williams et al., 2018) for stance prediction. **ChatGPT** is a strong zero-shot baseline that directly predicts the stance based on a task description (Zhang et al., 2023), as detailed in Appendix F.

Data Augmentation Baselines. We also compare *ZeroStance* with previous data augmentation methods. **RoBERTa** represents the RoBERTa-large model without data augmentation. We then consider the following data augmentation methods applied to RoBERTa. **BT** (Yu et al., 2018) employs a back-translation technique where English sentences are initially translated to French and then re-translated back to English. **EDA** (Wei and Zou, 2019) augments the dataset using four operations: random deletion, random swap, synonym replacement, and random insertion. **OpenStance** (Xu et al., 2022) generates training data based on texts or targets of existing datasets. **TTS** (Li et al., 2023b) first performs target augmentation based on a keyphrase generation model. A teacher-student learning framework is employed to improve target diversity by assigning pseudo stance labels to the augmented targets.

In our experiments, data augmentation baselines are trained in the out-of-domain setup. We perform data augmentation based on the source datasets. Models are trained on the combination of original source data and augmented data and evaluated on the out-of-domain data to understand its adaptability to an unseen dataset.

ZeroStance. *ZeroStance* is our proposed approach that constructs a synthetic open-domain dataset *CHATStance* with ChatGPT. As discussed in §3.3, three different types of prompts (*prompt 1*: provide examples, *prompt 2*: share personal experience and *prompt 3*: discuss relevant topics) are used to improve the diversity of generated texts. We further filter out 1% of instances with the lowest *variability* scores to improve the data quality. More details on data filtering are provided in Appendix G. Finally, we train and validate the RoBERTa-large model on *CHATStance* and test it on six benchmark datasets. Hyperparameters adopted in our experiments are shown in Appendix H.

5 Results and Discussions

In this section, we first compare our approach with previous unsupervised baselines and data augmentation methods (§5.1). We then perform ablation studies to understand the effectiveness of data filtering and prompts that are used for text generation (§5.2). We also investigate the impact of the data size to *ZeroStance* by training models on different sizes of the *CHATStance* dataset (§5.3). Next, we explore the performance of augmenting existing datasets with our *CHATStance* dataset (§5.4) and utilizing each human-annotated dataset as the open-domain dataset (§5.5).

5.1 Comparison with Baselines

We compare *ZeroStance* with unsupervised and data augmentation baselines. Results are shown in Table 4. First, we observe that *ZeroStance* outperforms unsupervised baselines on most datasets. In particular, *ZeroStance* outperforms the best baseline ChatGPT on four out of six datasets. We observe that when used as a stance classifier, ChatGPT struggles to accurately discern stances toward certain targets, such as "Atheism" in the sem16 dataset. Specifically, a large number of *against* and *favor* instances related to "Atheism" are misclassified as *neutral* by ChatGPT. In contrast, our proposed *ZeroStance* improves the zero-shot model by harnessing ChatGPT to produce a plethora of texts and targets (such as "*Christianity is intolerant of other religions and beliefs.*") from diverse domains. This suggests that, rather than directly applying ChatGPT for stance prediction, leveraging knowledge from ChatGPT to generate the dataset for strong baselines (specifically, RoBERTa in our study) proves to be more effective for open-domain stance detection.

Model	pstance	ibm30k	sem16	vast	covid	wtwt	Avg
No training							
Random Guess	51.25	51.51	30.85	33.48	33.87	30.50	38.58
GPT2	50.33	50.12	29.44	36.90	35.79	31.10	38.95
BART-MNLI	65.85	72.56	36.51	39.55	26.65	26.41	44.59
RoBERTa-MNLI	71.27	79.04	41.45	53.63	38.08	36.21	53.28
ChatGPT	**81.22**	88.12	55.80	71.91	66.67	**44.92**	**68.11**
Out-of-domain							
RoBERTa	74.33	62.37	55.56	64.65	66.12	30.47	58.92
+BT	76.07	57.62	55.75	68.12	64.87	28.46	58.48
+EDA	74.97	53.95	55.25	71.22	63.72	29.16	58.05
+OpenStance	75.68	51.37	57.14	71.70	61.68	29.48	57.84
+TTS	76.51	48.23	53.58	72.21	64.78	31.39	57.78
+ZeroStance	77.63	**94.42**	**59.28**	**72.24**	**67.54**	31.60	67.12

Table 4: Comparison of ZeroStance (RoBERTa + CHATStance) with unsupervised and data augmentation baselines.

Model	pstance	ibm30k	sem16	vast	covid	wtwt	Avg
ZeroStance	**77.63**	**94.42**	59.28	**72.24**	**67.54**	31.60	**67.12**
w/o *prompt 1*	75.83	94.36	60.30	71.85	62.16	29.68	65.70
w/o *prompt 2*	74.11	93.06	50.40	71.06	64.80	**33.25**	64.45
w/o *prompt 3*	75.66	93.70	**60.54**	71.18	66.51	31.98	66.60
w/o *prompt 1,2*	75.91	92.66	52.29	71.52	64.48	32.38	64.87
w/o *prompt 1,3*	75.15	93.02	58.66	69.94	57.70	31.93	64.40
w/o *prompt 2,3*	77.27	92.47	56.71	70.87	65.81	31.00	65.69
w/o *data filtering*	75.00	94.33	55.67	70.07	65.69	32.48	65.54

Table 5: Ablation studies of our approach on human-annotated datasets.

Second, in the out-of-domain setup, *ZeroStance* consistently outperforms RoBERTa across all datasets, achieving an average improvement of 8.2% on six datasets. Notably, *ZeroStance* exhibits an improvement of 32.05% on the ibm30k dataset, which suggests that the model trained on our synthetic dataset generalizes well to domains of ibm30k, which are not well captured by other models trained on existing datasets.

Third, *ZeroStance* shows improvements over all data augmentation methods across all datasets, achieving an average improvement of 9.08% in average F1-macro. This result demonstrates that our proposed method, which aims to cover a wide range of domains, proves to be more beneficial in open-domain stance detection than previous augmentation methods that merely augment data from limited domains. Note that *ZeroStance* demonstrates superior performance with a much smaller training set (around 21k) in contrast to RoBERTa and data augmentation methods such as BT that rely on substantially larger training sets of approximately 70k and 140k, respectively. This suggests that our proposed *ZeroStance* is much more data-efficient in open-domain stance detection.

5.2 Ablation Study

We conduct an ablation study to investigate the effectiveness of different components in *ZeroStance*. First, we study the impact of different prompts used in text generation. For a fair comparison, we gather an equivalent amount of data as the *CHATStance* dataset but using only one or two types of prompts. Second, we investigate the effectiveness of our *data filtering* approach by training the model on the original collected data without filtering. Results are shown in Table 5.

First, the removal of one type of prompt (e.g., "w/o *prompt 1*") leads to decrease in the performance on most datasets. Similar results can be observed when we only use one type of prompt (e.g., "w/o *prompt 1, 2*"). This implies that our proposed approach with all three prompts can effectively guide ChatGPT to generate more diverse texts, which strengthen the generalization abilities of zero-shot models. Second, the removal of *data filtering* ("w/o *data filtering*") results in worse performance on most datasets, which indicates that removing the data that the model is most decisive about (low-variability) can effectively improve the OOD performance, which is consistent with the observations of previous work (Swayamdipta et al., 2020). In Appendix I, we perform a human evaluation based on the filtered instances and show that our data filtering approach can effectively identify examples that are less informative and noisy.

Dataset	pstance	ibm30k	sem16	vast	covid	wtwt	Avg
CHATStance (100%)	**77.63**	**94.42**	**59.28**	72.24	67.54	31.60	**67.12**
CHATStance (16%)	72.89	91.68	51.78	70.77	60.47	31.76	63.23
CHATStance (33%)	75.63	93.49	56.95	71.12	63.34	32.31	65.47
CHATStance (66%)	75.80	92.80	58.53	**72.72**	66.32	31.85	66.34
CHATStance (200%)	74.38	93.96	55.79	71.56	**67.80**	**32.43**	65.99

Table 6: Results of different sizes of training set of *CHATStance*.

Model	pstance	ibm30k	sem16	vast	covid	wtwt	Avg
RoBERTa	74.33	62.37	55.56	64.65	66.12	30.47	58.92
+ZeroStance	**77.63**	**94.42**	**59.28**	72.24	67.54	31.60	67.12
+ZeroStance-Aug	75.98	93.67	56.70	**77.25**	**67.74**	**32.02**	**67.23**

Table 7: Augmenting existing human-annotated stance detection datasets with our *CHATStance* dataset.

Dataset	pstance	ibm30k	sem16	vast	covid	wtwt
pstance	-	55.34	37.31	33.25	32.30	28.46
ibm30k	75.84	-	40.42	36.80	36.06	30.07
sem16	65.29	62.51	-	63.37	61.40	21.60
vast	77.03	45.91	49.54	-	51.31	22.73
covid	62.69	50.53	27.85	59.06	-	**39.25**
wtwt	62.94	45.68	10.36	17.48	17.93	-
CHAT	**77.63**	**94.42**	**59.28**	**72.24**	**67.54**	31.60

Table 8: Comparison of each human-annotated dataset and *CHATStance* (CHAT) as the open-domain dataset. RoBERTa-large is trained on a single dataset (row) and evaluated on test sets of evaluation datasets (column). '-' indicates that we ignore the in-domain performance for the zero-shot setup.

5.3 Impact of Data Size

To understand the impact of data size on performance, we change the training size of *CHATStance* by randomly sampling subsets that comprise 16%, 33%, and 66% of the original training set and doubling the size of the dataset (200%) by generating more texts and targets. Then we train the RoBERTa-large model on each selected set. Results are shown in Table 6. We observe that performance decreases as we use a smaller training set of the *CHATStance* dataset. This result indicates the necessity of developing a large dataset. Besides, we observe that further increasing the size of the dataset decreases the model performance on four out of six datasets. A plausible reason is, as data size increases, the diversity of targets generated by ChatGPT may saturate due to the repetitive generation and an accumulation of similar data could potentially degrade model generalization. Interestingly, with just 7k training instances (33%) of the *CHATStance* dataset, our model surpasses the RoBERTa baseline (as shown in Table 4), which is trained on approximately 70k instances, on five out of six datasets. This underscores our approach's superior data efficiency over existing human-annotated datasets.

5.4 Data Augmentation with *CHATStance*

In this section, we explore the effectiveness of utilizing *CHATStance* as an augmented set for human-annotated datasets. Specifically, models are trained on the combination of *CHATStance* and five existing datasets in the out-of-domain setup (*ZeroStance-Aug*) and evaluated using the left-out dataset. The results are shown in Table 7. We observe that *ZeroStance-Aug* shows improvements over *ZeroStance* on three datasets, but the overall improvement in average F1-macro is not significant (67.23% for *ZeroStance-Aug* vs. 67.12% for *ZeroStance*). This indicates that mixing data from specific domains with an open-domain dataset does not necessarily improve the model's generalization to unseen domains.

5.5 *CHATStance* vs. Human-Annotated Datasets

We explore the potential of using each human-annotated dataset as an open-domain dataset and compare its efficacy with our *CHATStance* dataset. Specifically, we train and validate the RoBERTa model on training and validation sets of each dataset and then evaluate its performance on test sets of all datasets. The comparison between each human-annotated dataset and *CHATStance* is presented in Table 8. We observe that models trained on other datasets consistently lag behind the model trained on our *CHATStance* dataset across all test sets. Notably, even models trained on multi-domain datasets such as *vast* show worse performance than the model trained on *CHATStance*. This underscores the limited domain diversity of previous human-annotated datasets when compared to our *CHATStance* dataset.

We also perform a quality analysis on *CHATStance* and compute the total cost of data generation, which are shown in J and K, respectively.

6 Conclusion

In this paper, we propose a novel dataset generation approach for open-domain stance detection, which aims to train a model that performs well on unseen targets from all domains of interest. We leverage the ChatGPT to construct a synthetic open-domain dataset from scratch by generating controversial topics of diverse domains and the corresponding texts without training samples of existing datasets. Experimental results indicate that the model trained on our synthetic dataset shows better generalization to unseen targets of diverse domains over baselines on most benchmark datasets. Moreover, our proposed method is more cost-effective and data-efficient than previous methods.

Limitations

One limitation of our dataset is that it focuses only on America, Europe, and Asia to maintain consistency with established methodologies in prior research. However, we are keen to expand our dataset to a more diverse geographical scope, including regions like Africa in future work, enhancing the robustness and diversity of our findings. The other limitation is that we use ChatGPT for data generation, which could unintentionally introduce biases such as generating texts of fixed patterns. However, we mitigate this by considering diverse domains and attributed prompts (providing examples, sharing individual experience and discussing relevant topics) during the generation.

Ethical Statement

We gather targets and texts solely based on category names from a public debate website and our proposed prompts, ensuring ethical integrity by not including user-identifiable information and offensive content in ChatGPT's inputs. Although OpenAI has made significant efforts to mitigate toxicity issues, we have further enhanced security by integrating Google's Perspective API for toxicity detection in our generated content. Only after clearing this toxicity assessment are samples retained.

Acknowledgements

We would like to thank anonymous reviewers for their insightful comments to help improve the paper. This work is supported by the National Natural Science Foundation of China (NSFC) Key Project under Grant Number 62336006, the Pioneer and "Leading Goose" R&D Program of Zhejiang under Grant Number 2022SDXHDX0003 and the Ministry of Science and Technology of China Key Project under Grant Number 2022YFE0204900. We also thank the National Science Foundation for support from grants IIS-2107487 and ITE-2137846 which supported the research and the computation in this study. Yue Zhang is the corresponding author.

References

Abeer AlDayel and Walid Magdy. 2020. Stance detection on social media: State of the art and trends. *arXiv preprint arXiv:2006.03644*.

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4756–4767.

Erik Arakelyan, Arnav Arora, and Isabelle Augenstein. 2023. Topic-guided sampling for data-efficient multi-domain stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13448–13464.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very

large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

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

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7805–7813.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Yunyong Ko, Seongeun Ryu, Soeun Han, Youngseung Jeon, Jaehoon Kim, Sohyun Park, Kyungsik Han, Hanghang Tong, and Sang-Wook Kim. 2023. Khan: Knowledge-aware hierarchical attention networks for accurate political stance prediction. In *Proceedings of the ACM Web Conference 2023*, page 1572–1583.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1):1–37.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification. *arXiv preprint arXiv:2303.03953*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yang Li and Jiawei Yuan. 2022. Generative data augmentation with contrastive learning for zero-shot stance detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6985–6995.

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6299–6305.

Yingjie Li and Cornelia Caragea. 2021. Target-aware data augmentation for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1850–1860.

Yingjie Li and Cornelia Caragea. 2023. Distilling calibrated knowledge for stance detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6316–6329.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023a. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021a. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2021b. Improving stance detection with multi-dataset learning and knowledge distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6332–6345.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2023b. Tts: A target-based teacher-student framework for zero-shot stance detection. In *Proceedings of the ACM Web Conference 2023*, page 1500–1509.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022a. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, page 2738–2747.

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive graph for cross-target stance detection. In *Proceedings of the Web Conference 2021*, page 3453–3464.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022b. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3152–3157.

Xudong Liu and Shahira Fahmy. 2011. Exploring the spiral of silence in the virtual world: Individuals' willingness to express personal opinions in online versus offline settings. *Journal of Media and Communication Studies*, 3(2):45–57.

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*.

Zhengyuan Liu, Yong Keong Yap, Hai Leong Chieu, and Nancy Chen. 2023. Guiding computational stance detection with expanded stance triangle framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3987–4001.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2):1–40.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1868–1873.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. 2023. Zeroshotdataaug: Generating and augmenting training data with chatgpt. *arXiv preprint arXiv:2304.14334*.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. A multi-task model for emotion and offensive aided stance detection of climate change tweets. In *Proceedings of the ACM Web Conference 2023*, page 3948–3958.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text

classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Penghui Wei and Wenji Mao. 2019. Modeling transferable topics for cross-target stance detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1173–1176.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783.

Hanzi Xu, Slobodan Vucetic, and Wenpeng Yin. 2022. OpenStance: Real-world zero-shot stance detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–324.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2023. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, page 649–657.

Chenye Zhao, Yingjie Li, and Cornelia Caragea. 2023. C-STANCE: A large dataset for Chinese zero-shot stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13369–13385.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guan Wang, Kaichao Zhang, Cheng Ji, Qi Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

Sci-Fi, Privacy, Computers, Intelligence, Marijuana, Legislation, Donations, Sustainability, Currency, Medicine, Parenting, DDoS, Health, Coronavirus, Humans, Clothing, International, Programming, Morality, Ethics, Drugs, Bitcoin, Jobs, Refugees, Terrorism, Immigration, Housing, Charity, Christianity, Future, Baby, Cryptocurrency, Conscription, Social Media, Basketball, Music, Psychology, Recht, Cryptography, Sexuality, Aliens, Microsoft, Islam, Decentralization, Crime, Diplomacy, Caste, Antinatalism, Development, Film, Epistemology, Space, Economics, Abortion, Socialism, Bible, Taxation, Mental Health, Travel, Vegetarian, Feminism, Mathematics, Finance, C++, Movies, Poverty, Computing, Elections, LGBTQ, Software, Education, Climate Change, Capitalism, Marvel, Reform, Environment, Bacteria, Religion, Electricity, Research, Flat Earth Theory, College, Climate, Racism, Philosophy, Fandom, Diversity, Marriage, Cars, Society, Artists, Star Wars, IT, Games, America, Crypto, Human Rights, Death, Learning, Video Games, Consciousness, Law, Domestic Violence, Nature, Policy, Government, Comics, Blockchain, Business, Languages, Energy, Copyright, Fantasy, Advertising, Aircraft, History, Parliament, Capital Gains, Democracy, Women, China, School, False Theories, Gender, Christmas, Crowd Manipulation, Economy, Justice, Apple, Pandemic, Acting, Military, Transgender, Cannabis, Money, Culture, Affirmative Action, Vaccines, Ecology, Diet, Food, Politics, Animals, DNA, Fashion, Sports, Democrats, Nintendo, Atheism, Art, God, UK, Open Source, COVID-19, Android, Biology, Security, Regulation, Safety, Daesh, Brexit, India, TV, Healthcare, Animal Rights, War, Communication, Evolution, GMO, Nuclear Weapons, Violence, Parents, Botnet, Student, Catalonia, Economic Inequality, Death Penalty, Science Fiction, Discrimination, Judaism, Corruption, Gun, Mexico, Migrants, Globalization, EU, Cancer, Sex, Airbnb, AI, Commerce, Taxes, Vegan, Science, Censorship, Entertainment, Police, Relationships, Equality, United States, Altruism, Ethereum, Disease

Table 9: The full seed list of categories used for target generation.

Provide examples:	Generate a few consecutive sentences that are [STANCE] the "[CLAIM]". Please provide examples related to the claim in these sentences.
Personal experience:	Generate a few consecutive sentences that are [STANCE] the "[CLAIM]". Please generate sentences with only personal experience.
Relevant topics:	Generate a few consecutive sentences that are [STANCE] the "[CLAIM]". Please generate sentences by discussing topics or events that are related to the claim.

Table 10: Three Prompts for Text Generation.

A Seed Categories for Target Generation

The full seed list of categories for target generation is shown in Table 9. We generate the list by extracting pre-defined generic categories from *kialo*, which is an online debate platform that covers controversial topics from a wide range of domains.

B Prompts and More Examples of Text Generation

Motivated by observations of previous works (Mohammad et al., 2016; Liu and Fahmy, 2011; Li et al., 2021a), we consider three types of prompts for text generation, which are shown in Table 10. We input a generated claim and a stance label (against or in favor of) to ChatGPT and prompt ChatGPT to generate the corresponding text. We observe that ChatGPT can precisely capture the difference among our prompts and is able to generate high-quality texts as required. More examples of text generation using different prompts are shown in Table 11.

C Data Statistics of CHATStance

CHATStance includes near 27k synthetic instances that cover a wide range of domains. We split the dataset into training and validation sets in an 80/20 fashion. Data statistics of *CHATStance* can be seen in Table 12.

D Dataset Split

Train, validation and test sets are used as provided for covid, pstance, ibm30k and vast datasets. For sem16, training and test sets are provided and we split the training set into training and validation sets using an 85/15 split. For wtwt, we split the whole dataset into training, validation and test sets using an 80/10/10 split.

E Data Contamination

We evaluate our proposed method on six human-annotated stance datasets, which are *pstance* (Li et al., 2021a), *ibm30k* (Gretz et al., 2020), *sem16* (Mohammad et al., 2016), *vast* (Allaway and McKeown, 2020), *covid* (Glandt et al., 2021) and *wtwt* (Conforti et al., 2020).

Given the possibility that ChatGPT could have been trained using existing stance detection datasets, we investigate the potential data contamination by measuring the similarity between our *CHATStance* dataset (D_1) and each existing stance detection dataset (D_2). Following the previous work (Ubani et al., 2023), we adopt the Percent Word Overlap metric to measure the similarity between D_1 and D_2. Specifically, first, for any two text samples originating from the datasets, we remove stop words and punctuation. Second, we obtain the word overlap score by counting the number of unique words in both samples and dividing

Example 1	**Prompt:** Provide examples.
Target:	Legal age for drinking should be lowered in some Asian countries.
Text (Against):	Many Asian countries already struggle with alcohol-related issues such as drunk driving, violence, and health problems. Lowering the legal age for drinking will only exacerbate these issues. For example, in Vietnam, where the legal age for drinking is 18, alcohol is a factor in over 30% of road accidents. In addition, lowering the legal age will encourage underage drinking and make it easier for minors to obtain alcohol.
Text (Favor):	Studies have shown that the earlier people start drinking, the less likely they are to develop drinking problems. For example, in countries like Japan and China, where alcoholic beverages are a significant part of the culture, many young people start drinking in their teens.
Example 2	**Prompt:** Discuss relevant topics.
Target:	The use of fur in fashion is unethical and should be banned.
Text (Against):	Fashion is about choice, and wearing fur is a choice that some people make. Why should we tell them they cannot wear it if they find it stylish and comfortable? Furs are a natural material that has been used since ancient times for clothing, blankets, and other purposes. Many fur trimmings and accessories are made from by-products of the meat industry, which would otherwise go to waste.
Text (Favor):	For years, animal welfare organizations have been urging fashion designers and retailers to stop using fur in their products. The fur industry is notorious for its cruel treatment of animals, and not just for their fur. Many of these animals are bred and raised in cramped, unsanitary conditions before being beaten, electrocuted, or skinned alive. It's not just an ethical issue, either; the environmental impact of the fur industry is also significant.

Table 11: More Examples of generated texts given claims and stance labels with various prompts.

	#Against	#Favor	#Neutral
Train	7,188	7,189	7,007
Val	1,800	1,800	1,800

Table 12: Distribution of instances of *CHATStance*.

this number by the number of unique words in the lengthier text of the pair. Third, for each text sample within the D_1 dataset, we determine its maximum word overlap score by computing its word overlap score with each sample of the D_2 dataset. Finally, we calculate the average of these maximum scores to represent the overall word overlap similarity between D_1 and D_2.

In Table 15, we report the word overlap similarity between *CHATStance* and the test set of each human-annotated dataset. We also measure the similarity between the training set and the test set of each human-annotated dataset. We can observe that *CHATStance* exhibits the lowest word overlap similarity in four out of six human-annotated test sets. For instance, the overlap between the ibm30k training set and the sem16 test set exceeds the overlap between the *CHATStance* and the sem16 test set by a margin of 8%. This underscores the minimal risk of data contamination posed by the *CHATStance* dataset in our experiments.

F Prompt for ChatGPT Baseline

For datasets such as ibm30k and pstance, where the stance labels are limited to 'favor' and 'against', we use the following ChatGPT prompt: "Question: What is the stance of the text [TEXT] towards the claim [CLAIM]? The answer should just be selected from 'favor', or 'against'. Answer:" For datasets like vast, covid19, wtwt, and sem16, where the stance labels include 'favor', 'against', and 'neutral', our ChatGPT prompt is: "Question: What is the stance of the text [TEXT] towards the claim [CLAIM]? The answer should just be selected from 'favor', 'against', or 'neutral'. Answer:".

G Data Filtering with Various Ratios

In this section, we investigate the impact of removing instances with low *variability* of different ratios. Specifically, we remove instances with the lowest *variability* from the training set at rates of 1%, 3%, 5%, 10%, 20%, and 30%, respectively. These filtered training sets are then utilized to train the RoBERTa-large model. The hyper-parameter p (p=1) mentioned in §3.4 is selected based on the performance on the validation set. Results are shown in Table 13. First, we can observe that models trained with data filtering generally perform better than the model trained without data filtering, indicating the effectiveness of removing less informative and noisy data. Second, data filtering with the ratio of 1% shows the best performance in overall, which may suggest that an excessive data filtering of *CHATStance* dataset could remove informative data and skew the balance of the training data, thereby negatively affecting the model performance.

Model	pstance	ibm30k	sem16	vast	covid	wtwt	Avg
w/o *data filtering*	75.00	94.33	55.67	70.07	65.69	**32.48**	65.54
data filtering (3%)	73.25	93.90	57.19	70.98	66.40	32.01	65.62
data filtering (5%)	76.34	93.91	58.72	70.73	65.75	31.71	66.19
data filtering (10%)	76.45	93.89	56.18	71.87	66.01	31.59	66.00
data filtering (20%)	76.00	93.64	58.91	70.58	66.75	32.20	66.35
data filtering (30%)	74.17	**94.59**	58.00	69.26	62.86	31.85	65.12
data filtering (1%)	**77.63**	94.42	**59.28**	**72.24**	**67.54**	31.60	**67.12**

Table 13: Removing different ratios of low-*variability* instances for our data filtering method.

Text	Target	Stance
As a student, I have experienced firsthand the effects of budget cuts on education. It is discouraging to see programs like art and music being eliminated due to lack of funding. Teachers are being laid off and class sizes are increasing, which negatively impacts the quality of education.	The benefits of space exploration do not justify the cost.	Favor
Criticism of literature is a controversial issue. Writers have the right to express themselves. However, their work must be judged on its own merits. People should be free to choose what they want to read.	Urban fantasy novels featuring supernatural creatures are often racist and offensive.	Against
The world has become increasingly dependent on digital communication, we use it for everything from shopping to running a business. Governments and businesses alike rely on encryption tools to protect their sensitive data. Banning encryption tools would remove an essential layer of security. Without encryption, we would be vulnerable to cyberattacks, identity theft, and other criminal activities.	The problem of induction.	Neutral

Table 14: Examples of instances filtered by ZeroStance.

Dataset	pstance	ibm30k	sem16	vast	covid	wtwt
pstance	0.18	0.12	0.11	0.08	0.11	0.10
ibm30k	0.14	0.25	0.14	0.08	0.13	0.12
sem16	0.13	0.15	0.25	0.06	0.11	0.11
vast	0.09	**0.08**	0.08	0.10	0.09	**0.06**
covid	0.13	0.13	0.10	0.07	0.18	0.10
wtwt	0.11	0.10	0.09	**0.05**	0.08	0.39
CHAT	**0.08**	0.10	**0.06**	0.08	**0.07**	**0.06**

Table 15: Word overlap similarity between *CHAT-Stance* (*CHAT*) and the test set of each human-annotated dataset. We also calculate the similarity between the training set represented by each row and the test set by each column of each human-annotated dataset.

H Training Settings

In our work, we utilized the gpt-3.5-turbo-0301 version of ChatGPT for data collection. We performed all experiments on a single NVIDIA RTX A6000 GPU. RoBERTa-large[2] is used to evaluate our *CHATStance* dataset due to the effectiveness of RoBERTa on stance detection task (Li and Caragea, 2021; Hardalov et al., 2021; Li et al., 2021a; Liu et al., 2023). The total training time for fine-tuning the RoBERTa-large model on our dataset is less than 4 hours. The learning rate of RoBERTa-large is set to 1e-5. AdamW (Loshchilov and Hutter, 2019) is utilized as the optimizer. The mini-batch is set to 64. The model is trained for 4 epochs with early stopping and the patience is 5. Each result is the average of three runs with different initializations. Hyper-parameters are selected based on the model performance on the validation set.

[2] https://huggingface.co/vinai/bertweet-large

I Examples of Instances Filtered by Data Filtering

In this section, we perform human evaluations based on instances that are identified by our *data filtering* method. We observe that our *data filtering* method can effectively identify instances that are either noisy or less informative. We show some examples in Table 14. In the first two examples, the text and target are irrelevant, suggesting a *neutral* stance rather than the given stance labels. The third example presents a *neutral* stance toward the target. However, the target "The problem of induction" is too broad. This lack of specificity diminishes its utility for improving model generalization to new domains.

J Human Evaluation

To understand the effectiveness of *ZeroStance*, we perform a quality analysis on targets and texts generated by *ZeroStance*. For each prompt, we randomly selected 100 instances, ensuring a balanced distribution of *favor*, *against*, and *neutral* labels. We then examined the (text, target, stance) consistency. The analysis was conducted by two of the co-authors with expertise in stance detection, who had been trained using samples from existing stance datasets to ensure accurate relevancy assessment. Our findings indicate that 96% of instances demonstrate consistency in (texts, target, stance). Of the remaining instances, 83% exhibited

Text	Target	Stance
As a gardener, I know the value of planting more trees. In my backyard, I have planted dozens of trees over the years. They provide shade, oxygen, and a home for wildlife.	Planting trees is not an effective solution to reduce carbon emissions.	Favor
The warm sun beat down on my face as I walked along the beach, feeling the sand between my toes. I couldn't help but feel grateful for the abundance of beauty around me.	Corporate stock buybacks should be taxed as capital gains.	Against

Table 16: Examples of incorrect instances generated by *ZeroStance*.

incorrect stances towards the targets, while 17% contained texts that were irrelevant to their respective targets. This inconsistency may be attributed to the noise inherent in ChatGPT. We show some of our evaluated examples in Table 16. In the first example, the text appears to support tree planting, even though the target highlights the negative side of the practice and the stance label is *against*. The second example depicts a generated text that lacks relevance to the given target, where the stance label is *neutral*.

K Cost Comparison

The gpt-3.5-turbo-0301 API costs $0.0005 per 1k input tokens and $0.0015 per 1k output tokens. For our target generation, with the average of 30 input tokens and 12 output tokens across 204 categories in America, Europe, and Asia, the cost is $0.0005*30*204*3/1k + 0.0015*24k*12/1k ≈ $0.43. For our text generation task, we generate *favor* and *against* instances for 9k targets (including 3 types of prompts) with an average of 74 input tokens and 66 output tokens. The cost is $0.0005*74*9k*2/1k + $0.0015*66*9k*2/1k ≈ $2.45. The total cost of constructing *CHATStance* is only around $3. In comparison, online crowdsourcing platforms typically charge around $0.11 for 50 tokens (Wang et al., 2021), leading to a total expense of around $6.2k, which is over 2,000 times more expensive than our approach.