Stance Reasoner: Zero-Shot Stance Detection on Social Media with Explicit Reasoning

Maksym Taranukhin[1] **Vered Shwartz**[2,3] **Evangelos Milios**[1]
[1] Faculty of Computer Science, Dalhousie University
[2] Department of Computer Science, University of British Columbia
[3] Vector Institute for AI
{m.t,eem}@dal.ca, vshwartz@cs.ubc.ca

Abstract

Social media platforms are rich sources of opinionated content. Stance detection allows the automatic extraction of users' opinions on various topics from such content. We focus on zero-shot stance detection, where the model's success relies on (a) having knowledge about the target topic; and (b) learning general reasoning strategies that can be employed for new topics. We present Stance Reasoner, an approach to zero-shot stance detection on social media that leverages explicit reasoning over background knowledge to guide the model's inference about the document's stance on a target. Specifically, our method uses a pre-trained language model as a source of world knowledge, with the chain-of-thought in-context learning approach to generate intermediate reasoning steps. Stance Reasoner outperforms the current state-of-the-art models on 3 Twitter datasets, including fully supervised models. It can better generalize across targets, while at the same time providing explicit and interpretable explanations for its predictions.

Keywords: stance detection, reasoning, social media

1. Introduction

With an abundance of opinions expressed on the Internet every day, *stance detection*, which aims at identifying the stance of a text towards a target of interest (an entity, claim, topic, etc.), has attracted much attention from the NLP community, as a testbed for automatically extracting opinionated information from massive amounts of text (Alturayeif et al., 2023). In this paper, we focus on the challenging variant of the task, *zero-shot stance detection*,[1] where the model is applied to new stance targets, unseen during training (Allaway and McKeown, 2020).

The concept of model generalization in zero-shot stance detection refers to a model's capacity to correctly identify the stance on new targets that it has not encountered before. The generalization ability depends on two factors. First, the model must capture background knowledge about the target. Second, the model needs to have general-purpose reasoning strategies over the context and background knowledge that it can apply to new targets. Consider the example in Figure 1. To make a prediction, a model should understand the context (*"She wants to be @POTUS"*) and the background knowledge about the target (@POTUS is the Twitter handle for the president of the US, and Hillary Clinton is a presidential candidate). The model should reason that since the author is implying that Hillary Clinton is not qualified to be president because of her bad managing abilities, the stance is `against`.

Target: *Hillary Clinton* **Stance:** *Against*

"She can't even manage her husband and she wants to be @POTUS"

Background knowledge:

- *@POTUS* is the Twitter handle for the president of the US.
- *Hillary Clinton* is the Democratic party nominee for president in the 2016 presidential election.

Reasoning:

- **Premise**: Hillary Clinton is not qualified to be president because of her poor managing abilities.
- **Conclusion**: The author is against Hillary Clinton.

Figure 1: An example of stance detection involves reasoning over background knowledge.

Previous approaches to zero-shot stance detection typically involved fine-tuning a pre-trained language model (PLM) (Liu et al., 2021; Clark et al., 2021; He et al., 2022). These supervised methods suffer from several drawbacks. First, these models may be learning features specific to the training

[1] We use the term "zero-shot" to describe the evaluation of the model on targets not seen during training. This is distinct from the conventional use of "zero-shot" to denote unsupervised methods. While our approach employs in-context learning, commonly referred to as "few-shot learning", we opt to use "zero-shot" for consistency with previous literature on stance detection.

targets, which negatively affects their ability to generalize to new targets (Kaushal et al., 2021). Second, even models that incorporate knowledge from external knowledge bases (KBs) may struggle from missing, sparse, or irrelevant knowledge, leading to subpar performance (Ma et al., 2019). Lastly, these models only output the predicted stance label without explaining the reason behind their prediction. The lack of transparency makes it challenging to understand the models' decision-making processes and address their errors.

To address these problems, in this paper, we present *Stance Reasoner*, a framework for zero-shot stance detection on social media that leverages explicit reasoning over background knowledge to guide the model's inference about the document's stance on a target. To achieve this, *Stance Reasoner* employs the *in-context learning* approach (Brown et al., 2020). Unlike traditional methods that involve fine-tuning a PLM using a large training set, our approach involves providing the PLM with an optimized prompt. This approach, which avoids extensive training, enhances the model's capability to generalize effectively to new and unseen targets.

Specifically, our method utilizes a PLM as a source of world knowledge together with the *chain-of-thought* (CoT) approach (Wei et al., 2022) to generate intermediate reasoning steps that lead to a label prediction. Therefore, our method not only predicts the stance label but also generates the underlying reasoning that supports its prediction. The ability to produce such explanations can help in understanding and debugging the models' decision-making processes. We demonstrate how our method can be used to detect annotation errors and ambiguous or otherwise difficult examples.

We evaluate *Stance Reasoner* on three public Twitter stance detection datasets spanning a diverse range of targets. *Stance Reasoner* outperforms all the baseline methods including the fully supervised state-of-the-art models. In addition, the results demonstrate that *Stance Reasoner* can provide an interpretable and generalizable approach to zero-shot stance detection on social media.

Our contributions are as follows:

- We present *Stance Reasoner*, a framework for zero-shot stance detection on social media that leverages explicit reasoning over background knowledge to guide the model's inference about the document's stance on a target and is based on the chain-of-thought (CoT) in-context learning.

- We analyze the impact of CoT on stance detection and show that the *Stance Reasoner's* ability to reason using CoT depends on the diversity of reasoning strategies required for in-context examples.

- We demonstrate that our method outperforms the current state-of-the-art models on 3 Twitter datasets, including fully supervised models and it can better generalize across targets, while at the same time providing explicit and interpretable explanations for its predictions.

We make our code publicly available. [2]

2. Methodology

In this work, we focus on *zero-shot stance detection* (Allaway and McKeown, 2020), which means the model is evaluated on a test set containing new targets that were never observed during training.

We propose *Stance Reasoner*, a zero-shot stance detection approach. *Stance Reasoner* uses CoT (Wei et al., 2022) to explicitly reason over background knowledge in order to guide the model's prediction regarding the document's stance on the target. In particular, we use a PLM and the CoT with the self-consistency approach to generate multiple intermediate reasoning steps that lead to the final prediction. Intermediate reasoning serves two purposes. First, it guides the model inference, and second, it provides a way to gain insights into the model's decision-making process.

We present the prompt (Sec 2.1), motivate the choice of in-context examples (Sec 2.2), and describe the self-consistency approach that we use to further increase the model's accuracy (Sec 2.3).

2.1. Prompt Formulation

At the core of our approach lies an optimized prompt that is used to induce the model to generate intermediate reasoning steps. As our experiments show, choosing the right prompt is key to the method's success. We design a prompt that consists of (i) the task description; and (ii) a set of examples augmented with the intermediate reasoning steps. Both are described below.

Task description. To provide the model with the best description of the task, we select the prompt with the highest likelihood according to a PLM. Following Gonen et al. (2022), we first use a PLM to generate multiple paraphrases of manually defined seed task descriptions and then select the description that yields the lowest average perplexity on 100 random tweets.

[2] https://github.com/maksym-taranukhin/stance_reasoner

We use the following description of the stance detection task in the format of multiple-choice question answering, with the stance labels `against, favor, none` as answer candidates:

```
Question:  Consider the tweet in a
conversation about the target, what
could the tweet's point of view be
towards the target?
```

Examples. To guide the model in generating intermediate reasoning steps, we provide a set of in-context examples, each with its respective reasoning and label. We define reasoning as an argument: the premise interprets the tweet, and supports the conclusion, which is the author's stance on the target. Therefore, each in-context example has the following format:

```
tweet:  <tweet>
target:  <stance target>
reasoning:  <premise> -> <conclusion>
stance:  <label>
```

2.2. Choice of In-Context Examples

We argue that the context examples should cover a diverse set of reasoning strategies, both simple and more advanced, in order to help the model better generalize across documents and targets. Towards this end, we consider the reasoning strategies grouped based on two aspects: (1) target implicitness, i.e. whether the target is explicitly discussed in the document or whether it is implied, the latter requiring non-trivial reasoning strategy; and (2) the use of various rhetorical devices which might also require more complex reasoning. For example, whether the stance is expressed via sarcasm, jokes, aphorism, rhetorical question, etc. Ideally, we would like the prompt to cover examples from each group according to both aspects. In practice, exhaustively covering all possible reasoning strategies is not feasible in a short prompt. We thus limit the prompt to 6 examples, 2 examples for each label (`favor, against, none`). We include only examples with an implicit stance since they are more challenging for the model. Additionally, we include one example that uses sarcasm and another example that asks a rhetorical question. Finally, to comply with the zero-shot stance detection setup, the stance target of the in-context examples are distinct from the stance targets used in our experiments. Nonetheless, we observed that our prompt generalizes well across targets and datasets.

2.3. Self-Consistency

To further increase the model's prediction accuracy and its ability to generalize beyond the reasoning strategies present in the prompt, we employ the self-consistency approach (Wang et al., 2023). Specifically, we generate multiple completions of the same data point and take a majority vote on the predicted labels as the final prediction. Other than increasing the accuracy of the model, self-consistency can also be used to spot examples that are inherently hard to predict, either due to the ambiguity naturally present in a tweet without additional context, or due to annotation errors. We define the model's prediction confidence as the ratio between the number of runs that predicted the majority label and the total number of runs. By considering confidence, we can recognize and eliminate unreliable predictions.

3. Experiments

We conducted experiments to evaluate the effectiveness of the proposed approach in modelling background knowledge and its impact on zero-shot stance detection. Below we provide the description of the datasets (Sec 3.1), language models (Sec 3.2), experimental setup (Sec 3.3), evaluation metrics (Sec 3.4), and baselines (Sec 3.5).

3.1. Datasets

We conduct experiments on 3 Twitter datasets for stance detection, covering a wide range of domains. The SemEval-2016 Task 6a dataset (Mohammad et al., 2016) encompasses tweets pertaining to five targets across political, social, religious, and environmental domains. The WT-WT dataset (Conforti et al., 2020) focuses on tweets pertaining to corporate acquisition operations, along 5 targets, and with 4 labels (adding the `unrelated` label for tweets not discussing the target). Lastly, the COVID-19 Stance dataset (Glandt et al., 2021) contains tweets discussing the coronavirus pandemic, featuring 4 targets within the public health domain.

We use the Twitter API[3] to gather tweets from the WT-WT and COVID-19 datasets. Due to the limited accessibility of some tweets, the final dataset sizes are smaller than the originally collected. To form a testing split, we adopt different strategies for each dataset. For WT-WT, we randomly sample 100 data points for each target-label combination, yielding 2000 examples. Also, the `comment` and `unrelated` labels are merged into a single `none` label, ensuring that all datasets consist of 3

[3] https://developer.twitter.com/en/products/twitter-api

stance labels. For COVID-19, we fill in missing test data points using random samples from the training split with matching label-target combinations. Finally, we preprocess the SemEval-2016 Task 6a dataset to remove the #SemST hashtag, which is not stance-indicative.

3.2. Models

We evaluate our approach on a range of open-source auto-regressive language models due to their strong in-context learning abilities (Wang et al., 2022). Specifically, we use LLaMA 65B (Touvron et al., 2023) - an open-source model trained on publicly available datasets, and the Vicuna (13B) model (Chiang et al., 2023) - the LLaMA models that are finetuned to follow instructions with reinforcement learning (Ouyang et al., 2022).

3.3. Setup

All models are used for inference only. We utilize the HuggingFace Transformers library (Wolf et al., 2020) to load the LLaMA models in half-precision and run them using 4 A100 40GB GPUs. In all experiments, the maximum sequence length is set to 256 and the temperature is set to 0. When we sample multiple completions, the number of samples per tweet is set to 5 and the temperature is set to 0.7. We generate 50 paraphrases of each seed description using LLaMA 65B.

3.4. Evaluation Metric

Following prior work, we report the macro-averaged F_1 score across the `against` and `favor` labels for each of the targets in the dataset.

3.5. Baselines

We compare our approach to several baselines depending on the dataset. To simulate the zero-shot stance detection setting, we follow the leave-one-target-out evaluation setup. That is when the model is trained on all but one target which is held out for evaluation. However, since our method does not make use of the dataset's training split, we just measure the performance of each target in the test set individually.

3.5.1. Supervised Baselines

The supervised approaches are evaluated on SemEval-2016 Task 6a only.

BERT-base (Allaway et al., 2021). A vanilla BERT-base model with a classification head. The input is represented as `[CLS]<tweet>[SEP]<target>[SEP]`.

TGA Net (Allaway and McKeown, 2020). The model uses unsupervised clustering of BERT-embeddings together with attention to improve performance on new targets.

TOAD (Allaway et al., 2021). A BiLSTM model that uses adversarial learning to produce topic-invariant representations for better generalization to new targets.

BERT-GCN (Liu et al., 2021). A knowledge-infused model that uses conventional GCN to embed the nodes of a sub-graph consisting of entities extracted from ConceptNet (Speer et al., 2017).

JoinCL (Liang et al., 2022). A join contrastive learning framework for zero-shot stance detection that combines stance contrastive learning and target-aware prototypical graph contrastive learning.

3.5.2. Unsupervised Baselines

The unsupervised approaches are evaluated on all datasets.

Zero-Shot (Kojima et al., 2022). Unsupervised zero-shot stance detection via multiple-choice question answering. We provide a task description and a test example in the prompt and let the model generate the answer with greedy decoding. Similarly to our model, we chose the prompt that yielded the lowest average perplexity on 100 random examples. The prompt optimization is performed separately for each model and size.

Zero-Shot CoT (Kojima et al., 2022). We use the prompt from the zero-shot setup and engage the model in reasoning with zero-shot CoT using a two-step approach: 1) *reasoning generation* via appending `Let's think step by step` to the input, followed by 2) *answer prediction* by concatenating the generated reasoning to the input together with an answer trigger `Therefore, the answer is`. We use greedy decoding and parse the second step output to extract the prediction.

3.5.3. Few-Shot Baselines

Few-Shot (Brown et al., 2020). We modify the prompt from the zero-shot setup to include 6 in-context examples, two for each label. These examples have explicit and implicit stances towards the targets that are not present in the dataset.

Model	HC	FM	LA	AT	CC	Avg
Supervised Models						
BERT†	49.6	41.9	44.8	55.2	37.3	45.8
TOAD†	51.2	54.1	46.2	46.1	30.9	45.7
TGA Net‡	49.3	46.6	45.2	52.7	36.6	46.1
BERT-GCN‡	50.0	44.3	44.2	53.6	35.5	45.5
JointCL‡	54.8	53.8	49.5	54.5	39.7	50.5
Unsupervised Models§						
Zero-Shot						
Vicuna 13B	55.6	61.4	45.3	7.2	62.5	46.4
LLaMA 65B	26.6	31.8	66.3	57.9	46.2	45.8
Zero-Shot CoT						
Vicuna 13B	67.9	52.4	50.8	27.7	63.3	52.4
LLaMA 65B	25.1	76.2	47.2	70.8	47.2	53.3
Few-Shot Models§						
Few-Shot						
Vicuna 13B	41.4	63.2	44.7	50.3	63.1	52.5
LLaMA 65B	41.7	52.9	69.4	69.2	58.5	58.3
Few-Shot CoT						
Vicuna 13B	72.0	65.3	66.1	52.2	65.7	64.3
LLaMA 65B	69.1	67.7	72.9	71.2	61.8	68.5
Stance Reasoner (ours)						
Vicuna 13B	**74.4**	66.8	67.6	53.3	67.4	65.9
LLaMA 65B	73.7	**76.2**	**79.4**	**75.7**	**68.1**	**72.6**

† Results reported by (Allaway et al., 2021) ‡ Results reported by (Liang et al., 2022) § Our implementation

Table 1: Experimental results on the SemEval 2016 task 6a dataset. We report macro F_1 scores for each target in the test split, namely, HC - *Hillary Clinton*, FM - *Feminist Movement*, LA - *Legalization of Abortion*, AT - *Atheism*, CC - *Climate Change is a Real Concern* . The best results are highlighted in bold. The second-best results for each group of the baseline models are highlighted with underlining.

Few-Shot CoT (Wei et al., 2022). We use the same prompt and in-context examples as in the few-shot baseline, but augment the examples with manually-written reasoning chains.

4. Results

Table 1 displays the performance of *Stance Reasoner* and the baselines on the SemEval 2016 task 6a test set, in terms of the macro-F_1 score for each target. *Stance Reasoner* outperforms not only all the unsupervised baselines but also the supervised baselines, including a knowledge-infused model BERT-GCN, and across all targets—despite seeing only 6 examples. In fact, supervision seems to be detrimental in the leave-one-target-out setup, and our few-shot approach surpasses the best-supervised method by between 20 and 30 F_1 points.

While the supervised models differ in their base language model, we can see that even compared to the few-shot model, *Stance Reasoner* achieves better performance across targets, with very large gaps on some targets (e.g., 32 points on atheism). In general, few-shot methods performed better than zero-shot methods, and adding CoT to zero-shot degraded the performance. We attribute this to the fact that the few-shot approach has access to the set of supporting examples compared to the zero-shot CoT approach.

Overall, our method achieved the best average F_1 score of 72.6. The results suggest that the proposed approach is able to effectively use reasoning to infer the correct stance of a document on a target. This leave-one-target-out setup shows the method's ability to generalize its reasoning strategies across targets. We attribute this to the fact that the prompt contains diverse reasoning strategies that the model can learn to employ, and to the self-consistency strategy that is more robust to the randomness of the decoding strategy.

4.1. Ablation Tests

Impact of Diverse Examples. We analyze the impact of choosing examples with diverse reasoning strategies on the final performance of our model. To that end, we compare the performance of *Stance Reasoner* on the SemEval 2016 task 6a test set with the performance of an identical model that differs only by including only examples with explicit stances and without rhetorical devices

Model	SemEval	Covid-19	WT-WT
Unsupervised Models[§]			
Zero-Shot			
Vicuna 13B	46.6	48.5	65.7
LLaMA 65B	45.8	31.8	66.3
Zero-Shot CoT			
Vicuna 13B	52.4	53.1	35.1
LLaMA 65B	53.3	55.8	41.4
Few-Shot Models[§]			
Few-Shot			
Vicuna 13B	52.5	51.1	64.7
LLaMA 65B	58.3	52.9	69.4
Few-Shot CoT			
Vicuna 13B	64.3	74.9	69.8
LLaMA 65B	68.5	75.5	73.7
Stance Reasoner			
Vicuna 13B	65.9	74.6	73.6
LLaMA 65B	**72.6**	**76.2**	**78.3**

[§] Our implementation

Table 2: Generalization results across three datasets. We report average macro F_1 scores for all targets in the test split. The best results are highlighted in bold. The second-best results for each group of the baseline models are highlighted with underlining.

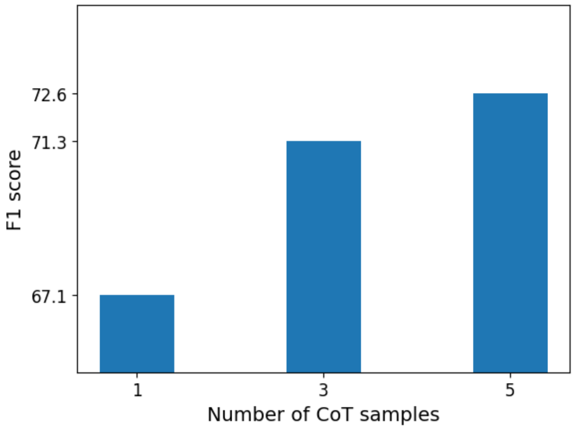

Figure 2: The impact of the number of sampled reasonings on the performance of Stance Reasoner. The performance increases with the number of samples.

(homogeneous in Table 3).

Stance Reasoner outperforms the homogeneous CoT prompt by a large margin. Furthermore, our approach outperforms the homogeneous CoT prompt even when the homogeneous model uses self-consistency and the diverse prompt doesn't. We conclude that including diverse reasoning strategies in the CoT prompt is beneficial for few-shot stance detection models.

Impact of Self-Consistency. We also evaluate the impact of the self-consistency strategy on the model performance. Table 3 shows that when we remove self-consistency (i.e. only generate one output), we get a substantial drop in performance. Fig 2 further shows the performance of our approach with self-consistency with a different number of sampled completions. The model performs better as we increase the number of samples. Self-consistency increases the model's robustness to noisy generations, leading to better performance.

Prompt generalization. We evaluate the generalization ability of our prompts by testing them on 3 Twitter datasets from different domains. Table 2 shows that our method based on the CoT prompt achieves state-of-the-art performance on all the datasets. This suggests that the CoT prompt is not domain-specific and can generalize to other domains without model fine-tuning.

Model size. We also analyze the effect of the model size on the performance of our approach. Table 1 and Table 2 show that the larger LLaMA 65B model consistently achieves better performance. However, the performance of Vacuna 13B is also competitive especially considering its size is 5 times smaller than LLaMA 65B. We hypothesize Vacuna's performance can be attributed to the ability of the model to follow instructions.

5. Qualitative Analysis

Table 4 shows the reasoning chains and labels predicted by *Stance Reasoner* for a few example tweets from the SemEval test set. We show how the model's confidence score (Sec 2.3) can be used to detect annotation errors, and ambiguous or difficult contexts.

① **Annotation Error.** The annotators marked this tweet as neutral instead of in favor of the stance "Climate change is a real concern". This is likely due to the absence of background knowledge that Urgenda is a pro-environment nonprofit organization. *Stance Reasoner* predicted the correct label and reasoning, thanks to its access to world knowledge from a PLM. Moreover, all self-consistency samples predicted the same label, as indicated by the confidence score. By employing this approach and manually inspecting tweets and corresponding reasoning we additionally found a total of 88 annotation errors in the dataset.

② **Ambiguous Context.** We also observe that some tweets are ambiguous without additional context. For example, whether the author of

Model	AT	CC	FM	HC	LA	Avg
Homogeneous CoT Prompt						
without Self-Consistency	71.0	63.8	65.6	69.1	64.7	66.9
with Self-Consistency	71.2	65.4	64.1	**76.4**	64.8	68.4
Stance Reasoner						
without Self-Consistency	72.2	65.5	75.1	75.7	66.8	70.6
with Self-Consistency	**73.7**	**66.2**	**79.4**	75.7	**68.1**	**72.6**

Table 3: Ablation study comparing the performance of Stance Reasoner to a less diverse, homogeneous CoT prompt. We report macro F_1 scores for each target in the SemEval 2016 task 6a test split, namely, HC - *Hillary Clinton*, FM - *Feminist Movement*, LA - *Legalization of Abortion*, AT - *Atheism*, CC - *Climate Change is a Real Concern*. The best results are highlighted in bold. The second-best results are highlighted with underlining.

#	**Example**	**Gold**	**Pred**	**Reasoning**	**Conf.**
①	**Tweet:** *It's most exciting to witness a major development! @urgenda* **Target:** *Climate Change is a Real Concern*	`none`	`favor`	**Premise:** *the author is excited about the major development of an organization that tackles climate change* **Conclusion:** *the author is in favor of climate change is a real concern*	1.0
②	**Tweet:** *One thing I learned from my job: doors to opportunity cover fee that only the privileged can afford. #privilege #truth* **Target:** *Feminist Movement*	`against`	`none`	**Premise:** *the author is in favor of truth* **Conclusion:** *the author is neutral towards the feminist movement*	0.4
③	**Tweet:** *@cbrangel so, you support the choice of wether or not you'd like to kill someone? Would you kill a born baby?* **Target:** *Legalization of Abortion*	`against`	`against`	**Premise:** *the author is against the idea of people choosing to kill other people* **Conclusion:** *the author is against the legalization of abortion.*	0.4

Table 4: Examples of tweets along with their gold label (**Gold**), and LLaMA 65B model-predicted label (**Pred**), reasoning chain (**Reasoning**), and self-consistency confidence as described in Sec 2.3 (**Conf.**). We show examples of: (1) annotation error, (2) ambiguous context, and (3) rhetorical devices.

the tweet is discussing male privilege specifically is unclear from the context. In those cases, the model's confidence tends to be lower. Despite the low confidence and the incorrect label predicted by the model for this example, the generated reasoning is logically sound.

③ **Rhetorical Device.** Finally, we observed low confidence predictions for tweets containing rhetorical devices such as rhetorical questions.

The results suggest that the proposed approach effectively uses reasoning over background knowledge to predict the correct label, and can even help identify annotation errors. Furthermore, the model is able to provide explicit and interpretable explanations for its predictions.

6. Prior Work

Target-specific stance detection. In this setup, the model is trained and evaluated on the same set of stance targets. Previous work on *target-specific* stance detection primarily employed additional knowledge sources of information to alleviate the problem of an implicit stance target, such as in Fig. 1 (Xu et al., 2019; Du et al., 2020; Sun and Li, 2021). Typically, a subgraph of relevant knowledge pertaining to the words in the stance document and target is extracted from KBs such as DBPedia (Auer et al., 2007) or ConceptNet (Speer et al., 2017), and incorporated into the stance detection model. Zhang et al. (2021) selected relevant concepts from multiple knowledge bases by measuring cosine-similarity between the BERT-embeddings of each concept and potential concepts (n-grams) in the document. Clark et al. (2021) employed Wikidata (Vrandečić and Krötzsch, 2014) to provide definition concepts to a language model as raw text.

Cross-target and zero-shot stance detection. Cross-target stance detection aims to predict stances for new targets related to the train targets, while zero- and few-shot stance detection aims to predict stance for entirely unrelated targets with no or little training data. In both setups, external knowledge may be used to help uncover implicit targets and generalize to new ones. Prior work incorporated relevant Wikipedia articles (Hanawa et al., 2019), semantic and emotion lexicons (Zhang et al., 2020), and knowledge from ConceptNet (Liu et al., 2021). However, as shown recently, such extracted knowledge is not consistently helpful for the model to make predictions (Chan et al., 2021; Raman et al., 2021). In addition, He et al. (2022) noted that retrieving relevant Wikipedia articles sometimes required manual work, and some targets lacked Wikipedia pages entirely (He et al., 2022).

Prompt-based and in-context learning approaches. Since all these methods are fully supervised, they tend to overfit the target-specific features and fall short when predicting the stance of new targets. In this work, we propose an in-context learning method that requires only a small number of labeled examples, preserving the model's generality which contributes to higher performance in zero-shot stance detection.

Recently, with the wide adoption of large LMs and in-context learning, there has been parallel work that also explored the use of prompts to perform stance detection. Zhang et al. (2023a) showed that CoT prompting with ChatGPT can outperform zero- and few-shot supervised learning approaches. However, the work didn't study the impact of the prompt selection on the performance, used a subset of the SemEval stance dataset targets and employed a closed-source model that limits the reproducibility of the work.

Our work is closely related to techniques for automatic CoT prompt construction, such as Auto-CoT (Zhang et al., 2023b), but stands out in several key aspects. Firstly, our approach employs a fixed prompt structure, which enables cross-dataset generalization and is specifically tailored for stance detection. Secondly, we not only assess the performance but also conduct a detailed analysis of the LM's reasoning abilities within the realm of stance detection. This is in contrast to methods like Auto-CoT, which primarily focus on enhancing performance without thoroughly examining the LM's reasoning structure and validity.

Chain-of-Thought Prompting *Chain-of-thought* (CoT; Wei et al., 2022) is an in-context learning approach that uses a language model to generate a sequence of intermediate reasoning steps that lead to the final prediction. *Few-shot* CoT builds a prompt to the model that consists of an optional task description T and a set of M examples. Each example $\{(x_i, r_i, y_i)\}_{i=1}^{M}$ consists of the input x_i, and the intermediate reasoning steps r_i that lead to label y_i. To obtain a prediction, the prompt is concatenated with a new data point x_i', for which the model needs to generate the reasoning steps and the predicted label. *Zero-shot* CoT (Kojima et al., 2022) excludes the set of M examples from the model's prompt and instead appends the text "Let's think step by step" to encourage the language model to generate the reasoning r in an unsupervised way. In a subsequent step, "Therefore, the answer is" is appended to the reasoning to predict the label. A further improvement, known as CoT with *self-consistency* can be achieved by sampling multiple completions for the same data point and taking a majority vote among them to obtain the final prediction (Wang et al., 2023).

Building on this foundation, our stance reasoner improves the traditional CoT method by including examples that demonstrate various reasoning strategies designed specifically for detecting stances in texts. It also uses a clear reasoning format that moves from the starting point to the conclusion, making it better suited for stance detection. These changes highlight what sets our work apart from regular CoT methods and emphasize its effectiveness in accurately analyzing and understanding social media texts with reasoning.

7. Conclusion

We presented Stance Reasoner, a zero-shot stance detection model. Stance Reasoner generates explicit reasoning over background knowledge to predict the stance of a given tweet regarding a target. Our empirical results show that Stance Reasoner outperforms the current state-of-the-art models on a Twitter dataset, including fully supervised models, and that it can better generalize to new targets, domains and datasets. We also presented a qualitative analysis of the model's performance, showing that it can accurately identify annotation errors and generate interpretable explanations for its predictions.

In the future, we plan to develop a model that can better handle tweets including rhetorical devices such as sarcasm and rhetorical questions, as well as tweets containing quotations. We will also further instigate the types of knowledge that are missing from language models and how to supplement the model with such knowledge. Finally, we also will aim to extend our method's application beyond concise texts like tweets to longer formats such as opinion pieces or blog posts. Given the more scattered and implicit presentation of information in these longer texts, adapting our method to accommodate the extended reasoning chains will pose a significant challenge.

8. Limitations

Language Models. The proposed approach relies on the knowledge encoded in a PLM. We expect the model's performance and generalization ability to degrade if tested on brand-new topics on which the PLM doesn't contain information. In addition, Stance Reasoner is most effective with larger models, which might be prohibitively expensive to run and geographically limited to some regions.

Social Media Text. We tested Stance Reasoner on datasets in the social media domain, where texts tend to be short and noisy. Although our approach is not designed specifically for this domain, the question of whether it can generalize to other domains or longer texts (e.g., news articles) is an interesting future research direction.

CoT Faithfulness and Task Definition. While CoT generates the intermediate reasoning steps leading to the prediction, there is no guarantee that the prediction causally depends on the reasoning steps (Creswell et al., 2023). While we were able to manually verify the correctness of a sample of the reasoning chains, we also note that judging some of the examples required substantial efforts and sometimes extra context. For example, a tweet such as "When women spend too much time out of the kitchen they get over opinionated and think they know everything #feminist" seems at face value to be against the feminist movement; however it may be interpreted in favor of it if it was written by a feminist user, as a sarcastic response to a misogynistic tweet. We thus advocate for future research to adapt the stance detection task from a classification task to a more flexible format where models can generate multiple interpretations along with their reasoning.

9. Ethical Considerations

We proposed a tool for automated stance detection on social media. As with any automated tool, it has the potential of being used in unintended ways and amplifying existing social issues such as political polarization. Thus, while the proposed approach can be used for various positive applications, such as identifying and addressing fake news, it is important to consider ethical implications and potential harms to individuals and society when deploying the proposed approach in real-world applications.

The datasets used in this paper are publicly available for research purposes on the owners' website.

Acknowledgements

We would like to express our sincere gratitude to the Digital Research Alliance of Canada for providing the computational resources that were instrumental in conducting the experiments and analysis presented in this paper. Furthermore, we would like to acknowledge the invaluable contributions of Diversio. The unique insights gained from this partnership greatly enriched the quality and impact of our research. Vered's research is supported by the Vector Institute for AI, the CIFAR AI Chair program, and the Natural Sciences and Engineering Research Council of Canada. Evangelos' research is supported by the Natural Sciences and Engineering Research Council of Canada, and the Boeing Company. Maksym was supported by Diversio via MITACS Accelerate grants.

10. Bibliographical References

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4).

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, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bin Bi, Chen Wu, Ming Yan, Wei Wang, Jiangnan Xia, and Chenliang Li. 2020. Generating well-formed answers by machine reading with stochastic selector networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7424–7431.

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. Curran Associates, Inc.

Aaron Chan, Jiashu Xu, Boyuan Long, Soumya Sanyal, Tanishq Gupta, and Xiang Ren. 2021. SalKG: Learning from knowledge graph explanations for commonsense reasoning. In *Advances in Neural Information Processing Systems*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Thomas Clark, Costanza Conforti, Fangyu Liu, Zaiqiao Meng, Ehsan Shareghi, and Nigel Collier. 2021. Integrating transformers and knowledge graphs for Twitter stance detection. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 304–312, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*.

Jiachen Du, Lin Gui, Ruifeng Xu, Yunqing Xia, and Xuan Wang. 2020. Commonsense knowledge enhanced memory network for stance classification. *IEEE Intelligent Systems*, 35(4):102–109.

John W Du Bois. 2007. The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 164(3):139–182.

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, Online. Association for Computational Linguistics.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. 2022. Demystifying Prompts in Language Models via Perplexity Estimation.

Kazuaki Hanawa, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. 2019. Stance Detection Attending External Knowledge from Wikipedia. *Journal of Information Processing*, 27(0):499–506.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from Wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77, Dublin, Ireland. Association for Computational Linguistics.

Ayush Kaushal, Avirup Saha, and Niloy Ganguly. 2021. tWT–WT: A dataset to assert the role of target entities for detecting stance of tweets. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3879–3889, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53(1).

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. 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, Dublin, Ireland. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. 2019. Towards generalizable neuro-symbolic systems for commonsense question answering. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 22–32, Hong Kong, China. Association for Computational Linguistics.

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, San Diego, California. Association for Computational Linguistics.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and Sentiment in Tweets. *ACM Transactions on Internet Technology*, 17(3):26:1–26:23.

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. Curran Associates, Inc.

Mrigank Raman, Aaron Chan, Siddhant Agarwal, PeiFeng Wang, Hansen Wang, Sungchul Kim, Ryan Rossi, Handong Zhao, Nedim Lipka, and Xiang Ren. 2021. Learning to deceive knowledge graph augmented models via targeted perturbation. In *International Conference on Learning Representations*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Yuqing Sun and Yang Li. 2021. Stance detection with knowledge enhanced bert. In *Artificial Intelligence*, pages 239–250. Springer International Publishing.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57:78–85.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022. What language model architecture and pretraining objective works best for zero-shot generalization? In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22964–22984. PMLR.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Michael Wojatzki and Torsten Zesch. 2016. Stance-based Argument Mining – Modeling Implicit Argumentation Using Stance. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, pages 313–322.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhenhui Xu, Qiang Li, Wei Chen, Yingbao Cui, Zhen Qiu, and Tengjiao Wang. 2019. Opinion-Aware Knowledge Embedding for Stance Detection. In Jie Shao, Man Lung Yiu, Masashi Toyoda, Dongxiang Zhang, Wei Wang, and Bin Cui, editors, *Web and Big Data*, volume 11642 of *Lecture Notes in Computer Science*, pages 337–348. Springer, Cham.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Yangyang Li, and Liwen Jing. 2023a. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.

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, Online. Association for Computational Linguistics.

Xin Zhang, Jianhua Yuan, Yanyan Zhao, and Bing Qin. 2021. Knowledge Enhanced Target-Aware Stance Detection on Tweets. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction*, pages 171–184. Springer Singapore.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

A. Dataset Details

Target	train split counts				test split counts			
	against	favor	none	total	against	favor	none	total
Atheism	304	92	117	513	160	32	28	220
Climate Change	15	212	168	395	11	123	35	169
Feminist Movement	328	210	126	664	183	58	44	285
Hillary Clinton	393	118	178	689	172	45	78	295
Legalization of Abortion	355	121	177	653	189	46	45	280
All Targets	1395	753	766	2914	715	304	230	1249

Table 5: The SemEval 2016 Task 6a dataset (Mohammad et al., 2016) count statistics. The splits are provided by the dataset authors.

Target	refute	support	comment	unrelated	total
Aetna → Humana	717 (1106)	728 (1038)	1937 (2804)	1925 (2949)	5307 (7897)
Anthem → Cigna	1286 (1969)	682 (970)	2068 (3098)	3293 (5007)	7329 (11622)
CVS Health → Aetna	294 (518)	1323 (2469)	3016 (5520)	1618 (3115)	6251 (11622)
Cigna → Express Scripts	140 (253)	408 (773)	506 (947)	306 (554)	1360 (2527)
Disney → 21st Century Fox	217 (378)	797 (1413)	4568 (8495)	4019 (7908)	9601 (18194)
All Targets	2654 (4224)	3938 (6663)	12095 (20864)	11161 (19533)	29848 (51284)

Table 6: The WT-WT dataset (Conforti et al., 2020) count statistics. The counts represent the number of data points accessible via Twitter API. The original counts provided by the dataset authors are enclosed in parentheses.

Target	train split counts				val split counts				test split counts			
	again.	favor	none	total	again.	favor	none	total	again.	favor	none	total
Anthony S. Fauci, M.D.	211 (480)	273 (388)	312 (596)	796 (1464)	28 (65)	39 (52)	39 (83)	106 (200)	27 (65)	43 (52)	38 (83)	108 (200)
Keeping Schools Closed	98 (166)	259 (409)	135 (215)	492 (790)	26 (42)	69 (103)	41 (55)	136 (200)	19 (42)	64 (103)	37 (55)	120 (200)
Stay at Home Orders	137 (284)	104 (136)	417 (552)	658 (972)	30 (58)	18 (27)	89 (115)	137 (200)	31 (58)	18 (27)	92 (115)	141 (200)
Wearing a Face Mask	206 (512)	366 (531)	170 (264)	742 (1307)	31 (78)	52 (81)	22 (41)	105 (200)	35 (78)	52 (81)	24 (41)	111 (200)
All targets	652 (1442)	1002 (1464)	1034 (1627)	2688 (4533)	115 (243)	178 (263)	191 (294)	484 (800)	112 (243)	177 (263)	191 (294)	480 (800)

Table 7: The COVID 19 Stance dataset (Glandt et al., 2021) count statistics. The counts represent the number of data points accessible via Twitter API. The original counts provided by the dataset authors are enclosed in parentheses. The splits are provided by the dataset authors.

B. Model Details

We used the following model checkpoints from HuggingFace Hub in our experiments:

- decapoda-research/llama-65b-hf
- TheBloke/vicuna-13B-1.1-HF

C. Prompts

C.1. Prompt Generation

We manually define the following seed task descriptions:

- *What is the tweet’s stance on “{target}”?*
- *In the context of a discussion about “{target}”, what could be the tweet’s stance on “{target}”?*

We used the following meta-prompt to paraphrase the seed task descriptions with LLaMA-65B and sampling temperature = 0.7:

```
Write 50 diverse paraphrases for the following sentence: <seed-prompt>.
Paraphrases:
```

The perplexity of a task description is measured on random 100 examples sampled from the train split of the SemEval-2016 dataset and format with the below prompt. Note: we did not use labels in the prompt therefore this procedure is unsupervised.

```
Question: {task description}
The options are:
- against
- favor
- none
tweet: <{text}>
Answer:
```

Utilizing this approach we found that the best prompt is the same among the models. We explain this observation due the fact Vicuna is a finetuned LLaMA model and therefore shares the same pre-training data.

C.2. Zero-Shot Prompt

```
Question: In a conversation about “{target}”, what could the tweet’s point of
view be towards “{target}”?
The options are:
- against
- favor
- none
tweet: <{text}>
Answer: The tweet could be
```

C.3. Zero-Shot CoT Details

The stance detection prompt used in zero-shot CoT:

```
Question: In a conversation about “{target}”, what could the tweet’s point of
view be towards “{target}”?
The options are:
1. against
2. favor
3. none
tweet: <{text}>
Answer: Let’s think step by step. {CoT}. Therefore, the answer is
```

We used the following regular expression to find the first occurrence of an option number concatenated with a dot and take the corresponding option word as the final prediction: `(1|2|3).`

C.4. Few-Shot Prompt and Few-Shot CoT Prompt

The Stance Reasoner few-shot prompt with reasoning chains is highlighted in blue.

```
Question:  Consider the tweet in a conversation about the target, what could the
tweet's point of view be towards the target?
The options are:
- against
- favor
- none

tweet:  <I'm sick of celebrities who think being a well known actor makes them
an authority on anything else.  #robertredford #UN>
target:  Liberal Values
reasoning:  the author is implying that celebrities should not be seen as
authorities on political issues, which is often associated with liberal values
such as Robert Redford who is a climate change activist -> the author is against
liberal values
stance:  against

tweet:  <I believe in a world where people are free to move and choose where
they want to live>
target:  Immigration
reasoning:  the author is expressing a belief in a world with more freedom of
movement -> the author is in favor of immigration.
stance:  favor

tweet:  <I love the way the sun sets every day.  #Nature #Beauty>
target:  Taxes
reasoning:  the author is in favor of nature and beauty -> the author is neutral
towards taxes
stance:  none

tweet:  <If a woman chooses to pursue a career instead of staying at home, is
she any less of a mother?>
target:  Conservative Party
reasoning:  the author is questioning traditional gender roles, which are often
supported by the conservative party -> the author is against the conservative
party
stance:  against

tweet:  <We need to make sure that mentally unstable people can't become killers
#protect #US>
target:  Gun Control
reasoning:  the author is advocating for measures to prevent mentally unstable
people from accessing guns -> the author is in favor of gun control.
stance:  favor

tweet:  <There is no shortcut to success, there's only hard work and dedication
#Success #SuccessMantra>
target:  Open Borders
reasoning:  the author is in favor of hard work and dedication -> the author is
neutral towards open borders
stance:  none
```

C.5. Homogeneous Few-Shot Prompt and Few-Shot CoT Prompt

Homogeneous few-shot prompt with reasoning chains is highlighted in blue.

```
Question:  Consider the tweet in a conversation about the target, what could the
tweet’s point of view be towards the target?
The options are:
- against
- favor
- none

tweet:  <RT @MyDailyMeat:  Real food is MEAT, not vegetables.  Humans were built
to eat meat, not vegan diets.  #meatlover #notvegan #realfood>
target:  Veganism
reasoning:  the author is against vegan diets -> the author is against veganism
stance:  against

tweet:  <The rainbow flag means more than just a pride symbol.  It’s a symbol
of our fight for EQUALITY. #LoveIsLove>
target:  LGBTQ Rights
reasoning:  the author is in favor of equal rights for the LGBTQ community ->
the author is in favor of LGBTQ rights
stance:  favor

tweet:  <I love the way the sun sets every day.  #Nature #Beauty>
target:  Taxes
reasoning:  the author is in favor of nature and beauty -> the author is neutral
towards taxes
stance:  none

tweet:  <@lifekingra @guardian The public can’t be trusted to be 100% honest
in their "truthful" interpretations and memories.>
target:  Police Body Camera Ban
reasoning:  the author is against relying on the public’s interpretations and
memories -> the author is against of police body camera ban
stance:  against

tweet:  <Veganism is not a restriction but rather an expansion of your love,
care and respect for all creatures.>
target:  Animal Rights
reasoning:  the author is in favor of veganism -> the author is in favor of
animals -> the author is in favor of animal rights
stance:  favor

tweet:  <There is no shortcut to success, there’s only hard work and dedication
#Success #SuccessMantra>
target:  Open Borders
reasoning:  the author is in favor of hard work and dedication -> the author is
neutral towards open borders
stance:  none
```