Use of Large Language Models for Stance Classification

Iain J. Cruickshank, [1] **Lynnette Hui Xian Ng** [2]
[1,2]Army Cyber Institute
2101 New South Post Road
Highland Falls, NY 10996
[2]Carnegie Mellon University
5000 Forbes Road
Pittsburgh, PA
iain.cruickshank@westpoint.edu, huixiann@andrew.cmu.edu

Abstract

Stance detection, the task of predicting an author's viewpoint towards a subject of interest, has long been a focal point of research. Current stance detection methods predominantly rely on manual annotation of sentences, followed by training a supervised machine learning model. This manual annotation process, however, imposes limitations on the model's ability to fully comprehend the stances in the sentence and hampers its potential to generalize across different contexts. In this study, we investigate the use of Large Language Models (LLMs) for the task of stance classification, with an absolute minimum use of human labels. We scrutinize four distinct types of prompting schemes combined with LLMs, comparing their accuracies with manual stance determination. Our study reveals that while LLMs can match or sometimes even exceed the benchmark results in each dataset, their overall accuracy is not definitively better than what can be produced by supervised models. This suggests potential areas for improvement in the stance classification for LLMs. The application of LLMs, however, opens up promising avenues for unsupervised stance detection, thereby curtailing the need for manual collection and annotation of stances. This not only streamlines the process but also paves the way for expanding stance detection capabilities across languages. Through this paper, we shed light on the stance classification abilities of LLMs, thereby contributing valuable insights that can guide future advancements in this domain. The code used in this study is made available at https://anonymous.4open.science/r/LLM-Stance-Labeling/README.md.

Introduction

Identifying and classifying an individual's stance towards a particular entity is a pivotal challenge in the realm of computational social science research. Stance detection entails the automated prediction of an author's viewpoint or stance towards a subject of interest, often referred to as the "target" (Alturayeif, Luqman, and Ahmed 2023). Typically, a stance towards a subject is categorized as "Agree", "Disagree", or "Neutral". However, the labels representing stance can vary based on the specific target or context. Essentially, a stance mirrors an individual's perspective toward a specific topic or entity. Stance detection is used in downstream tasks like fake news detection, opinion surveys, and rumor detection (Küçük and Can 2022).

While the concept of stance might seem straightforward, detecting and classifying it involves unique challenges. First, the definitions of stance for labeling purposes can be ambiguous. For instance, previous studies have indicated discrepancies in stance definitions across various benchmark stance detection data sets. This inconsistency raises questions about the transferability of models trained on these data sets (Ng and Carley 2022; Allaway and McKeown 2023). Additionally, understanding stance is inherently context-dependent, as it represents an opinion about a specific entity. Without the appropriate context, comprehending the stance becomes nearly impossible. Consequently, these challenges hamper the broad applicability of any stance detection model, making stance classification an enduring challenge.

At the same time, recent developments in Large Language Models (LLM) have enabled breakthroughs in complex language understanding. In particular, through *prompting* LLMs, researchers have been able to use LLMs to solve several complex language tasks (Brown et al. 2020; Schmidt et al.). This paradigm of prompting a large pre-trained model even works in settings with few or no labeled data to solve classification problems (Liu et al. 2023; Brown et al. 2020; Zhao et al. 2021; Wei et al. 2023). Thus, it is possible for an LLM, with a suitable prompt to classify language by an ambiguous and complex label. As such, a few recent works have investigated the applicability of ChatGPT to classify stances on a couple of benchmark data sets (Zhang, Ding, and Jing 2022; Mets et al. 2023; Aiyappa et al. 2023). These works produced mixed results for the performance of ChatGPT and only considered task-based prompting, thus it is not clear if LLMs and prompt engineering could be used for stance classification more broadly.

In this paper, we investigate the following question: How well do Large Language Models with prompt engineering perform at stance classification, without fine-tuning? We utilize five publicly available data sets in this study and query different LLMs with four different prompting schemes of increasing contextual information. We also perform our testing with minimal use of labels (only few-shot prompting) to explore how useful these models and prompting are in a more real-world setting where one does not have labels for their stance classification problem already. Despite LLMs and

prompt engineering showing promise for some data sets, our surprising conclusion is that the underlying task still remains a challenge. Indeed, we find that LLMs do not perform significantly better than current supervised machine learning stance classification models, and present artifacts like inconsistent outputs, opening up areas for future research on improving the stance detection capabilities of LLMs.

Related Research

Previous work on stance detection has focused on the construction of supervised machine-learning models for the task. A commonly used machine learning classifier is the Support Vector Machine (Lai et al. 2018; Elfardy and Diab 2016), which has performed well in the SemEval-2016 stance detection competition (Mohammad et al. 2016). Supervised models that use neural network architectures are also popular. These include the use of convolutional neural networks (Wei et al. 2016), or recurrent neural networks (Zarrella and Marsh 2016), sometimes enhanced with textual entailment (Zhao and Yang 2020) and data augmentation for improved accuracies (Kawintiranon and Singh 2021). Most recent work, however, has focused on multi-task learning objectives and transfer learning from transformer-based neural networks (Alturayeif, Luqman, and Ahmed 2023; Yang et al. 2019; Zhao and Yang 2020). Despite the typically stronger in-domain performance of these models, they often struggle to generalize to new data or other targets of stances, and are often of little use to real-world practitioners due to these generalizability shortcomings (Ng and Carley 2022; Alturayeif, Luqman, and Ahmed 2023).

While most work in stance detection has focused on supervised machine learning techniques, there are also some unsupervised techniques. Unsupervised learning methods make use of the idea of language homogenity for label classification (Zhang et al. 2023). In that aspect, graph networks are a popular technique. Zhang et al. (2023) used graph neural networks to formulate homogeneous and heterogenous information for a user on Twitter, inferring the stance based on past tweets and tweets of neighbors. Another technique is label propagation based on the user interaction network or relationships (Weber, Garimella, and Batayneh 2013), to propagate stance labels based on existing knowledge. The interaction network can also be divided into partitions, with the stance of each partition then interpreted (Pick et al. 2022). Darwish et al. (2020) first projected a large set of Twitter user data to a low-dimensional space before clustering the interaction network into multiple partitions. While these methods do not rely on having a set of stance labels to train a model, they often rely on very specific scenarios for their use (i.e., a social media website that has explicit behavioral links between users, which can be used to create a network) and strong assumptions about the language or behavior of users to infer stance (i.e., use of a certain word or hashtag always conveys a certain stance).

Recently, there has been an increase in research focusing on the concept of zero-shot stance detection. Notably, (Allaway and McKeown 2023) discuss a variety of techniques for zero-shot stance detection and present an adaptation of the SemEval2016 dataset (Zarrella and Marsh 2016), along with their own VAST dataset, specifically designed for this purpose. They highlight the multitude of ways stance characterization can be achieved in a zero-shot manner. According to (Allaway and McKeown 2023), there are three primary paradigms of zero-shot stance detection: topic, language, and genre. For each paradigm, a model is typically trained on all data, except for one element of the paradigm which is reserved for the zero-shot test. For instance, in zero-shot topic stance detection, a model is trained on data from all topics except one, which is then used for evaluation. (Allaway and McKeown 2023) found that all models perform less optimally in the zero-shot setting compared to a fully supervised setting.

In terms of using LLMs for the stance detection task, current work has focused on just ChatGPT, with mixed results. (Zhang, Ding, and Jing 2022) found that with just an instruction-based prompt (although they hint at the use of reasoning in a prompt in their paper), ChatGPT could produce better results on the SemEval2016 benchmark data set than supervised models. However, Aiyappa et al. (2023) examined the stance detection task using the ChatGPT model, observing that while there is a boost in performance, there could be potential contamination of data through its massive trained dataset, therefore making the evaluation unreliable. Additionally, Mets et al. (2023) probed the usability of ChatGPT as a zero-shot classifier, using only a task-based prompt, for stance detection on a custom data set on the topic of immigration in various language news articles. They found that ChatGPT performed close to the best supervised model, but was ultimately inferior for the stance classification task. Finally, given the similarities between stance and sentiment, it is also worth mentioning that (Kheiri and Karimi 2023) investigated all of the OpenAI model offerings on the task of sentiment classification with a couple of different prompts and different benchmark data sets and found that GPT models, especially if fine-tuned, can significantly outperform any other models for sentiment classification. Overall, it is still unclear if LLMs, especially with prompt engineering and without the use of fine-tuning on labeled data, can perform the task of stance classification.

With the advent of LLMs for natural language tasks, the new discipline of prompt engineering has also come into being. Prompt engineering is the discipline of finding the right ways to provide input to models — or 'prompt' them — to get the best outputs (Schmidt et al.; White et al. 2023; Ramlochan 2023). While this is a fast-changing discipline, with new findings about prompt engineering constantly emerging, there are a couple of techniques that have emerged for use with LLMs in particular. The first is few-shot prompting, which is when you give a few examples of what you want the LLM to do as part of the prompt (White et al. 2023; Brown et al. 2020; Wei et al. 2023). This is different from fine-tuning the LLM or low-shot learning, as no training (i.e., adjusting of model weights) is performed when the examples are given; the examples are only given as part of the context of the task (Brown et al. 2020). While this prompting technique does consistently produce improved outputs, there are still possible instabilities in the technique which

can be caused by things like the ordering of the examples (Zhao et al. 2021; Lu et al. 2021). Another prompting technique that has consistently improved the output of LLMs is Chain-of-Thought Reasoning (Wei et al. 2022; Chen et al. 2023). In this prompting scheme, the LLM is typically asked to explain its reasoning and to work through answering a prompt step-by-step. Answering prompts in such a process usually improves outcomes and helps to prevent undesirable behaviors like hallucination, where the model produces a plausible-looking answer that is incorrect (Wei et al. 2022; Chen et al. 2023). This technique has been used in an iterative, chatting format to improve implicit sentiment classification in previous research (Fei et al. 2023). Thus, while the best means of interacting with LLMs is an open research question, there are certain prompting techniques, like few-shot prompting, that can elicit better outcomes from LLMs.

Methodology

In this section, we review the benchmark data sets for the stance classification task and describe the prompting techniques with an LLM to produce stance classification results.

Data Sets

We used a total of five publicly available, benchmark data sets, which have been manually annotated: covid-lies, election2016, phemerumors, semeval2016, and wtwt. These five data sets are of similar properties: they are constructed out of sentences which are Twitter posts and are written in the English language. The targets that are in the data sets range from misconceptions to elections to tragedies, which means the definition of stance varies between them. For example, in covid-lies and phemerumors the stance is about whether the statement supports or denies a rumor, while in semeval2016 and election2016 the stance is about an opinion of the target. Table 1 lists the data sets that are used, the targets that are present in the data sets, and the highest reported accuracy as found with the original papers. We followed the same data set handling procedures described in (Ng and Carley 2022) in terms of standardizing labels for evaluation (but not for prompting).

Prompting for Stance Classification

In order to investigate the use of LLMs and prompting for the task of stance classification, we used four different prompting schemes. The prompting schemes are hierarchical in nature such that each prompting scheme incorporates more information from the previous scheme. The following figure, Figure 1, displays the general, overarching prompting scheme and what elements are available to the LLM within each, individual prompting scheme, while actual examples of each prompt scheme are provided in the Appendix.

The following list details each of the prompting schemes we used to classify stance for each of the data sets.

1. **Task-only:** In the task-only prompt, we adopt a zero-shot learning prompting method, providing only the task (e.g. ``Classify the following statement ...''). for this we provide different classification outcomes depending on the data set, but they generally follow a 'AGREE', 'DISAGREE', or 'NEUTRAL' format.
2. **Context:** In this prompting scheme we add in contextual information about what the statement is and the target of the stance classification to the task of classifying the stance of the statement (e.g. ``The following statement is a [context]. Classify the following statement toward [target]''). This prompting scheme is what the previous works that investigated using ChatGPT for stance labeling used (Zhang, Ding, and Jing 2022; Aiyappa et al. 2023; Mets et al. 2023).
3. **Context + FSP:** For this prompting scheme, we now utilize few-shot prompting (Brown et al. 2020) and provide some examples of the stance being classified for a statement and entity. For this prompting scheme, we keep the context provided in the context scheme, to include the target for each of the few-shot examples.
4. **Context + FSP + Reasoning:** Lastly, we further enhance the prompts with reasoning. In this prompting scheme, we provide a reason for why each few-shot example was classified as the stance that it was classified as and we further prompt the LLM to provide its reasoning for why it classifies a statement with a particular stance by '... and the reasoning for the classification in the form of: 'stance: STANCE, reason: REASON''. With this prompt, we seek to leverage some of the benefits of Chain-of-Thought prompting (Wei et al. 2022) by forcing the LLM to consider its reasoning as part of performing the task.

Results

In this section, we present the results of using an LLM with prompting to perform stance classification. We begin by describing our test setup and then go on to present results for each of the benchmark data sets.

Experimental Set Up

In this section, we detail the LLMs we investigated, how the prompting schemes were constructed for each data set, and the hardware used for testing.

LLMs Given the results in (Aiyappa et al. 2023) and (Chen et al. 2023), we opted to use only local, open-sourced LLMs for this investigation, as it is possible closed LLMs, like those from OpenAI, may have data contamination issues with the benchmark data sets. In particular, we used a number of encoder-decoder and decoder-only models available on HuggingFace (Wolf et al. 2019). For the decoder-only models, we attempted to use GPT-NeoX (Black et al. 2022), Falcon -7B and -40B (Almazrouei et al. 2023), MPT-7b (Team 2023), and Llama-2 -7B and -13B (Touvron et al. 2023). Unfortunately, for the decoder-only models, we found they could not reliably produce stance classifications under any of the prompting schemes. In many cases, they would return nonsensical responses such as elements of the prompt, blank spaces, or an attempt to explain the task. As

Data Set	Event	Unweighted F1-score	Number of Examples
covid-lies	misconceptions towards COVID-19 pandemic	50.2	3,196
election2016	2016 US Presidential elections	0.55	2,378
phemerumors	tragedies (unrest, disasters, hostage, plane crash)	0.33	2,859
semeval2016	atheism, climate change, feminism, Hillary Clinton, abortion	0.69	2,814
wtwt	Company mergers and acquisitions	0.62	32,409

Table 1: Summary of data sets used with our descriptions of the events and best-reported unweighted, macro F1-score from the original data set.

```
The following social media posts are [context]. Each statement can [stance options] towards its associated
[target] and Each statement has the reason for its stance toward the target.

[target]: example target
statement: example statement
stance: [stance option]
reason: example reason
...

Now, classify the following statement as to whether it [stance options] toward the [target] [target context], and
give your reason for the classification. Only return the classification for the statement towards the [target] and
the reasoning for the classification in the form of: 'stance: STANCE, reason: REASON'

[target]: instance target
statement: instance statement
```

Figure 1: Overarching Prompting Scheme for Stance Classification Text highlights indicate the information available for each of the prompting schemes. Purple is the prompt with the task, green provides the addition of context, blue provides few-shot examples, and red adds reasoning to the examples. Each of the bracketed words indicates verbiage that would vary depending on the data set.

such, we do not present results from the decoder-only models. For the encoder-decoder models, we experimented with Flan-UL2 and Flan-Alpaca-GPT4-T5, which are both T5-based models and currently state of the art for these types of models (Tay et al. 2022; Chia et al. 2023).

For each of these models, we employ HuggingFace's AutoTokenizer[1] and pipeline classes to run the models (Wolf et al. 2019). We set each of the models to only provide the most probable output (e.g., by setting `temperature=0`) and a `max_length` of 1000 tokens.

Prompt Details For each of the data sets, we altered the prompts slightly, based on the context of the data set. The following table, Table 2, summarizes how the prompting schemes were adjusted in the methodology for each data set. For the few-shot prompting (FSP) scheme, we included 5 examples (shots) that were taken from the data set at random. These examples were the same for every statement of that data set. We used the Python Package Langchain to programmatically construct these prompts for the tests [2].

Hardware All of the tests were run on a computer with Ubuntu 22.04 Linux with x64 CPU with 40 cores, 376 GB of RAM and two NVIDIA A6000 GPUs.

[1]https://huggingface.co/docs/transformers/v4.33.0/en/model_doc/auto#transformers.AutoTokenizer

[2]https://www.langchain.com/

Evaluation For evaluation, we report the unweighted, macro-F1 accuracy metric following previous work (Mohammad et al. 2016). This macro-F1 score adjusts for the proportion of each class label type, for there is an imbalance of class labels in some of our data sets.

Experimental Results

The results for different prompting schemes and LLMs for the different benchmark data sets are presented in Table 3. For each of the combinations, we ran the test three times and report the average result; we found that the LLM outputs could vary slightly between runs, especially when there was no context provided as part of the prompt.

From the testing, we only found that an LLM with prompting could outperform the benchmark, supervised models on only two of five of the benchmark data sets. That said, the LLM plus prompting results do come close, often to within 0.05 or less of the supervised benchmark results. Additionally, when compared to previous zero-shot stance detection results from (Allaway and McKeown 2023), the LLMs perform significantly better on the semeval2016 data set than the zero-shot models (which still were able to train on some of the topics as part of the zero-shot topic evaluation).

We also note that the inclusion of context into the prompt was the single factor that most increased the performance of the LLMs; including context into the prompt always increased performance in stance classification. This result

Data Set	Context	Stance Options	Target	Target Context
covid-lies	about COVID or Coronavirus	supports, denies, neutral, unrelated	belief	is true
election2016	about politics	for, against, neutral	politician	N/A
phemerumors	commenting on whether a rumor is true	supports, denies, neutral	rumor	is true
semeval2016	expressing an opinion about an entity	for, against, neutral	entity	N/A
wtwt	that may be commenting on a corporate merger	for, against, neutral, unrelated	event	happening

Table 2: Summary of the prompt differences between each of the benchmark data sets. For each data set, we used slightly different target, context, and stance labels options in order to accommodate the different purposes of stance classification between the data sets.

		Data Set				
Prompting Scheme	**LLM**	**Phemerumors**	**covid-lies**	**semeval2016**	**election2016**	**wtwt**
Task Only	Flan-Alpaca-GPT4-T5-3B	0.27	0.29	0.41	0.42	0.44
Context	Flan-Alpaca-GPT4-T5-3B	0.35	0.37	0.57	0.49	0.47
Context and FSP	Flan-Alpaca-GPT4-T5-3B	0.41	0.45	0.57	0.46	0.5
Context and FSP and Reason	Flan-Alpaca-GPT4-T5-3B	0.35	0.49	0.61	0.58	0.49
Task Only	Flan-UL2	0.31	0.31	0.42	0.45	0.42
Context	Flan-UL2	0.37	0.35	0.66	0.55	0.57
Context and FSP	Flan-UL2	**0.44**	0.36	0.65	0.57	0.56
Context and FSP and Reason	Flan-UL2	0.32	0.41	0.64	**0.59**	0.53
Benchmark		0.33	**0.5**	**0.69**	0.55	**0.62**

Table 3: Average unweighted F1-scores for each prompting scheme, LLM, and data set combination. The benchmark results are given in the final row. The highest scoring result for each data set is bolded.

makes sense, as the definition of stance relies on context, such as the target of the stance. Whereas, the inclusion of few-shot examples and reasoning did not always improve performance across all of the benchmark data sets. This may be a result of which examples were selected for the few shots as it is known that example selection can affect few-shot prompting performance (Zhao et al. 2021; Lu et al. 2021). Finally, we note that the larger T-5 model generally performed better than the smaller one, despite the smaller one being trained on newer data sets (i.e., Alpaca-GPT4 which are data generated by GPT-4 in response to Alpaca prompts), which argues in favor of the general consensus that larger models are more capable.

Additionally, during testing, we noted that while the LLM was given explicit instructions about the output to return, there were occasionally variances in this output. For example, the model would occasionally return responses like `'For'`, `'for'`, `"'FOR'"`, or `'The stance is FOR'` for the stance label of `'FOR'`. While this can be addressed by a relatively simple post-processing script, and while we inspected the outcomes of all of the runs of the models for inconsistencies in outputs and did not find anything that could not be easily addressed, this is still an issue that needs to be accounted for when using LLMs for stance classification.

Along with the inconsistencies in output formats, we also found that the LLMs did not always provide meaningful reasons when confronted with the context + FSP + reason prompt. The models would occasionally recycle a reason from the few-shot examples or would not output a reason at all. Once again, this can be handled by a simple post-processing script, but it is another issue with using LLMs for stance classification.

Stance vs Sentiment Classification

In order to see if we could improve the performance of the LLMs, we also investigated a change in the verbiage of the prompts. Instead of asking for the 'stance' of a statement toward a target, we instead asked for the 'sentiment' of a statement toward a target. Sentiment analysis is a close cousin of the stance detection task. It analyzes the attitudes towards the text, expressing the attitudes in the form of a polarity (e.g. positive, negative, neutral) and has been used to understand the belief of people through their online writing (i.e., news, blogs) (Godbole, Srinivasaiah, and Skiena 2007). It is also a task that has been more broadly researched and has many more tools and data sets for the task, including data sets that are included in the pre-training of the T5 models. Our prompt generation setup and evaluation is the same as the stance classification task, except for replacing the words "stance" with "sentiment", and we only looked at the two data sets that have stance definitions closest to the definition of sentiment: semeval2016 and election2016 and only with the Flan-Alpaca-GPT4-T5 model. The results of this investigation are presented in the table, Table 4

From these results, we can clearly see that using the term 'sentiment,' despite sentiment classification being more familiar to the LLMs, actually decreased performance. As a result, we did not attempt to use a directed sentiment classification prompt as a proxy for stance classification for the

	Data Set	
Prompt	**semeval2016**	**election2016**
Task Only	0.36	0.4
Context	0.45	0.45
Context and FSP	0.5	0.44
Context and FSP and Reason	0.46	0.43

Table 4: Results of attempting to use sentiment toward a target as a proxy for stance classification, since the models had been pre-trained on a sentiment classification task. In each case, this prompting did not improve performance

other, less sentiment-like data sets. This result also indicates that LLMs perceive stance and sentiment differently and that the classification of stance is a different task for an LLM than the classification of sentiment.

Discussion

Large Language Models have become prominent and widely adopted due to their accessibility and capabilities to perform a plethora of natural language tasks. The use of LLMs has already become a mainstay for language-based tasks like summarization, question-answering, and translation. In our work, we examined the use of LLMs for stance prediction with different prompting schemes. We observed that the use of Large Language Models (LLMs) supplemented with adequate prompting produced results that were comparable to fully supervised models. These models also outperformed previous work looking at zero-shot models, and so in more demanding environments than previous zero-shot stance detection tests (previous zero-shot stance modeling would only hold out a particular topic or genre and train on the rest (Allaway and McKeown 2023), whereas the LLMs tested in this research did not train on any of the data whatsoever). This is a significant finding as it underscores the potential of LLMs in conjunction with effective prompting to achieve high performance in stance prediction tasks, often matching or even surpassing more resource-intensive supervised models.

Furthermore, while supervised machine learning models generally infer the target of the sentence themselves before deciding on a stance, in our investigation of LLMs, we provided contextual information that indicates things like the stance targets. The inclusion of contextual information in the prompts consistently improved the results. This information also increases the ability of the stance prediction to work across varied targets, removing the need to construct stance data sets pertaining to each target, and thereby improve generalizability. This technique can be enhanced with past work on Target-Stance-Extraction, which automatically extracts the target and corresponding stance from a sentence, reducing the need for large-scale target annotations (Li, Garg, and Caragea 2023). Overall, context-based prompts seemed to provide the models with a broader perspective of the information, enabling them to make more accurate stance predictions. This finding aligns with the notion that providing a richer context can aid in more nuanced understanding and better task performance as well as the definition of stance requiring context.

Further, LLMs are typically able to understand multiple languages, which facilitates multilingual stance classification, removing the need to collate and annotate datasets across multiple languages and find native speakers to do so. This opens up opportunities to analyze large sets of data with varied languages, such as opinion expression on social media.

Interestingly, we also found that different types of LLMs performed differently on the task. Specifically, encoder-decoder models, such as the T-5 model, demonstrated successful performance on the stance prediction task, in a zero-shot setting and with no additional model training. On the other hand, decoder-only models were not able to perform the task as effectively under the same scenario. This discrepancy in performance might be due to the inherent architecture of these models. Encoder-decoder models are inherently designed to understand the context and generate relevant output, making them well-suited for tasks like stance detection. Decoder-only models, however, might require additional fine-tuning on stance detection tasks before they can effectively carry out the task.

Our research also highlighted that few-shot prompting did not always improve performance. The reason for this is not entirely clear, but it's likely that the selection of samples for the few-shots may not have been optimal. These samples were selected at random and remained the same for every statement classified, which may have affected the effectiveness of few-shot prompting.

In conclusion, our findings underscore the versatility of LLMs in stance prediction tasks, particularly when used with context-informed prompting and an encoder-decoder architecture. Further research could explore fine-tuning strategies for decoder-only models and optimal sample selection for few-shot prompting in stance detection tasks, potentially unlocking new avenues for their application.

Stance detection as an LLM Benchmark Large Language Models (LLMs) are built and assessed on a range of benchmarks, from common sense reasoning, reading comprehension, to mathematical reasoning (Zheng et al. 2023). However, one language task not represented in current LLM benchmarking is complex language classification, such as stance classification. The task of stance classification is a crucial benchmark to investigate, as many policy formulations depend on understanding public opinion, for instance, opinions towards climate change policies (Upadhyaya, Fisichella, and Nejdl 2023). Misclassification of sentence stances can lead to incorrect interpretations of opinion slant. For downstream analyses that study aggregated stances towards topics to understand public reaction and formulate policy, incorrect classification can result in erroneous interpretations and policy mismatches (Alturayeif, Luqman, and Ahmed 2023).

Given the importance of stance classification to the wider society and the results of this study, we would like to propose stance classification be considered as a future benchmarking task for LLMs.

Limitations As in all studies, several limitations nuance our work. Our data sets are premised on manual annotations, which could be subjected to inconsistent annotations and varied sentence interpretations (Ng and Carley 2022). For example, there is a sentence about Michael Essien having Ebola, "@xx no he hasn't. The man himself confirmed not true @MichaelEssien" that was annotated as a neutral stance whereas it should be a stance *against* the claim that Michael Essien had contracted Ebola.

Additionally, while we attempted to consider a wide range of open-source models and possible configurations for those models, we were constrained by computational resources and time from using every possible permutation of open-source models for the task of stance classification. We believe that we have tested a representative sample of offerings, however, it is possible that a certain LLM, perhaps due to pre-training data or even architectural differences, may actually perform uncharacteristically in regards to the models tested in this study.

Broader Perspectives & Ethical Considerations In all research involving Large Language Models (LLMs), it is crucial to acknowledge and address potential ethical implications. A primary concern arises from the datasets used to pre-train LLMs, from which they acquire their language capabilities and knowledge base. These datasets may harbor inherent biases or offensive content (Schaul, Chen, and Tiku 2023). Although this study makes no attempt to exploit or introduce any form of bias, it is conceivable that these biases might inadvertently permeate the analysis performed using LLMs. This underlines the importance of diligently scrutinizing the data used to train LLMs, especially when they are employed in socially significant tasks such as stance classification, where bias can have profound implications.

Another ethical consideration pertains to the environmental impact of running these computationally intensive models. It's indisputable that LLMs consume more energy compared to their smaller counterparts. Thus, their use in tasks like stance classification is associated with a tangible energy cost. However, it is also crucial to balance this against the alternative scenario, which involves continuous human effort for labeling data, a process that is both labor-intensive and time-consuming due to the generalizability problem inherent in stance classification.

Looking ahead, as we strive to create more sustainable and efficient computational models, one potential avenue could be leveraging LLMs to distill smaller, more energy-efficient models for production purposes. This could significantly decrease the energy demand, making stance classification tasks more environmentally sustainable, while still benefiting from the superior performance of LLMs. As our understanding of LLMs continues to evolve, it is paramount to remain vigilant about these ethical considerations and strive towards more responsible and sustainable practices.

Finally, as with any classification effort of text, such efforts could be used for text-based censorship. For example, it's possible that our research could be used to identify, at scale, comments and users that a presenting a certain stance toward a target and then remove those users or comments. We believe, however, that the benefits of being able to more correctly classify stances of text comments outweigh its potential misuse. and that the same precautions used to prevent misuse with the classification of texts more broadly can also be applied to this work.

Conclusion

Stance classification is a crucial task that contributes significantly to discerning the author's perspective towards a particular event. Despite the demonstrated proficiency of Large Language Models (LLMs) in numerous natural language tasks, their performance varies and does not consistently surpass state-of-the-art models (Kocoń et al. 2023). In this study, we have illuminated the potential of LLMs, particularly when combined with effective prompting, in the realm of stance classification. However, it's important to note that they do not definitively outperform existing supervised methods.

Stance classification, due to the intricacies of language expression and the context-dependent nature of stance, continues to pose a formidable challenge. Yet, the utilization of LLMs for stance classification offers promising opportunities. Notably, it permits the adaptation of stance classification outputs without requiring extensive human annotation, thereby enabling the application of stance classification techniques in a diverse array of contexts beyond those that the original datasets were designed for.

This study serves to enhance our understanding of the stance classification capabilities of LLMs and proposes a pathway for future advancements. The findings underscore the need for improving both prompting schemes and LLM models, using stance classification as a benchmark. As we navigate this path, we anticipate pushing the boundaries of what's possible in stance classification, ultimately contributing to more nuanced and effective natural language processing applications.

Acknowledgments. The research for this paper was supported in part by the Center for Informed Democracy and Social-cybersecurity (IDeaS) and the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. This work was also conducted within the Cognitive Security Research Lab at the Army Cyber Institute at West Point and supported in part by the Office of Naval Research (ONR) under Support Agreement No. USMA 20057. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense, the U.S. Army, or the U.S. Government.

References

Aiyappa, R.; An, J.; Kwak, H.; and Ahn, Y.-Y. 2023. Can we trust the evaluation on ChatGPT? *arXiv preprint arXiv:2303.12767*.

Allaway, E.; and McKeown, K. 2023. Zero-shot stance detection: Paradigms and challenges. *Frontiers in Artificial Intelligence*, 5: 1070429.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Goffinet, E.; Heslow, D.; Launay, J.; Malartic, Q.; Noune, B.; Pannier, B.; and Penedo, G. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

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

Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonell, K.; Phang, J.; Pieler, M.; Prashanth, U. S.; Purohit, S.; Reynolds, L.; Tow, J.; Wang, B.; and Weinbach, S. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, J.; Chen, L.; Huang, H.; and Zhou, T. 2023. When do you need Chain-of-Thought Prompting for ChatGPT? *arXiv preprint arXiv:2304.03262*.

Chia, Y. K.; Hong, P.; Bing, L.; and Poria, S. 2023. INSTRUCTEVAL: Towards Holistic Evaluation of Instruction-Tuned Large Language Models. *arXiv preprint arXiv:2306.04757*.

Darwish, K.; Stefanov, P.; Aupetit, M.; and Nakov, P. 2020. Unsupervised user stance detection on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 141–152.

Elfardy, H.; and Diab, M. 2016. Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 434–439.

Fei, H.; Li, B.; Liu, Q.; Bing, L.; Li, F.; and Chua, T.-S. 2023. Reasoning Implicit Sentiment with Chain-of-Thought Prompting. *arXiv preprint arXiv:2305.11255*.

Godbole, N.; Srinivasaiah, M.; and Skiena, S. 2007. Large-Scale Sentiment Analysis for News and Blogs. *Icwsm*, 7(21): 219–222.

Kawintiranon, K.; and Singh, L. 2021. Knowledge Enhanced Masked Language Model for Stance Detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4725–4735.

Kheiri, K.; and Karimi, H. 2023. SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning. *arXiv preprint arXiv:2307.10234*.

Kocoń, J.; Cichecki, I.; Kaszyca, O.; Kochanek, M.; Szydło, D.; Baran, J.; Bielaniewicz, J.; Gruza, M.; Janz, A.; Kanclerz, K.; et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 101861.

Küçük, D.; and Can, F. 2022. A Tutorial on Stance Detection. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, 1626–1628. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391320.

Lai, M.; Patti, V.; Ruffo, G.; and Rosso, P. 2018. Stance evolution and twitter interactions in an italian political debate. In *International Conference on Applications of Natural Language to Information Systems*, 15–27. Springer.

Li, Y.; Garg, K.; and Caragea, C. 2023. 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)*, 10071–10085.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; and Stenetorp, P. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Mets, M.; Karjus, A.; Ibrus, I.; and Schich, M. 2023. Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media. *arXiv preprint arXiv:2305.13047*.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 31–41.

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

Pick, R. K.; Kozhukhov, V.; Vilenchik, D.; and Tsur, O. 2022. STEM: unsupervised structural embedding for stance detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11174–11182.

Ramlochan, S. 2023. What is Prompt Engineering? https://www.promptengineering.org/what-is-prompt-engineering/. Accessed: 2023-09-14.

Schaul, K.; Chen, S. Y.; and Tiku, N. 2023. Inside the Secret List of Websites that make AI like ChatGPT Sound Smart.

Schmidt, D. C.; Spencer-Smith, J.; Fu, Q.; and White, J. ???? Cataloging Prompt Patterns to Enhance the Discipline of Prompt Engineering.

Tay, Y.; Dehghani, M.; Tran, V. Q.; Garcia, X.; Bahri, D.; Schuster, T.; Zheng, H. S.; Houlsby, N.; and Metzler, D. 2022. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Team, M. N. 2023. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs. Accessed: 2023-05-05.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Upadhyaya, A.; Fisichella, M.; and Nejdl, W. 2023. A Multi-task Model for Sentiment Aided Stance Detection of Climate Change Tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 854–865.

Weber, I.; Garimella, V. R. K.; and Batayneh, A. 2013. Secular vs. Islamist Polarization in Egypt on Twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, 290–297. New York, NY, USA: Association for Computing Machinery. ISBN 9781450322409.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wei, J.; Wei, J.; Tay, Y.; Tran, D.; Webson, A.; Lu, Y.; Chen, X.; Liu, H.; Huang, D.; Zhou, D.; et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Wei, W.; Zhang, X.; Liu, X.; Chen, W.; and Wang, T. 2016. pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 384–388.

White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D. C. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yang, R.; Xie, W.; Liu, C.; and Yu, D. 2019. BLCU_NLP at SemEval-2019 Task 7: An Inference Chain-based GPT Model for Rumour Evaluation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 1090–1096. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Zarrella, G.; and Marsh, A. 2016. MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 458–463.

Zhang, B.; Ding, D.; and Jing, L. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Zhang, C.; Zhou, Z.; Peng, X.; and Xu, K. 2023. DoubleH: Twitter User Stance Detection via Bipartite Graph Neural Networks. *arXiv preprint arXiv:2301.08774*.

Zhao, G.; and Yang, P. 2020. Pretrained embeddings for stance detection with hierarchical capsule network on social media. *ACM Transactions on Information Systems (TOIS)*, 39(1): 1–32.

Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 12697–12706. PMLR.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Appendix 1: Examples of Prompts

This section provides examples of the prompts that are used in this work as an input to the LLMs to predict the stance of a sentence.

No Context

Template Classify the statement as to whether it is 'SUPPORTS', 'DENIES', 'NEUTRAL', or 'UNRELATED'. Only return the classification label for the statement, and no other text.

statement: {statement}

Sample Prompt Classify the statement as to whether it is 'SUPPORTS', 'DENIES', 'NEUTRAL', or 'UNRELATED'. Only return the classification label for the statement, and no other text.

statement: Their is NO "5G radiation poisoning" going on, so stop spreading this bullsh*t. The coronavirus is real, it's making people sick, and killing some, not 5G.

Context only

Template The following statement is social media post about COVID or Coronavirus. Classify the statement as to whether it 'SUPPORTS', 'DENIES', is 'NEUTRAL', or is 'UNRELATED' to the belief below being true. Only return the classification label for the statement toward the belief, and no other text.

belief: {belief}
statement: {statement}

Sample Prompt The following statement is social media post about COVID or Coronavirus. Classify the statement as to whether it 'SUPPORTS', 'DENIES', is 'NEUTRAL', or is 'UNRELATED' to the belief below being true. Only return the classification label for the statement toward the belief, and no other text.

belief: Coronavirus is caused by 5G.
statement: Their is NO "5G radiation poisoning" going on, so stop spreading this bullsh*t. The coronavirus is real, it's making people sick, and killing some, not 5G.

Context + Few Shot Prompting

Template examples = [
{'belief': {belief},
'statement': {statement},
'stance': {stance} },
{'belief': {belief},
'statement': {statement},
'stance': {stance} },
{'belief': {belief},
'statement': {statement},
'stance': {stance} }
]

prefix = '''
The following statements are social media posts about COVID or Coronavirus. The statements can support, deny, be neutral, or be unrelated toward its associated COVID belief.
'''

suffix = '''
Now, classify the following statement as to whether 'SUPPORTS', 'DENIES', is 'NEUTRAL', or is 'UNRELATED' toward the belief below being true. Only return the classification for the statement toward the belief, and no other text.

belief: {belief}
statement: {statement}
'''

Sample Prompt The following statements are social media posts about COVID or Coronavirus. The statements can support, deny, be neutral, or be unrelated toward its associated COVID belief.
belief: Africans are more resistant to coronavirus.
statement: Happen now Blacks are Immune to the coronavirus ' there is a GOD https://t.co/LRq7SZYK0G
stance: SUPPORTS

belief: Alex Jones' silver-infused toothpaste kills COVID-19
statement: #China #COVID-19 As work resumes in outbreak, brand-new 'normal' emerges https://t.co/VENOSSOnx5 https://t.co/RQoeSWoaHH
stance: UNRELATED

belief: COVID-19 is only as deadly as the seasonal flu.
statement: @islandmonk @Stonekettle Closer to 650,000 people will die of the flu this year. The figure of 30,000 is just in the U.S.
But the flu has approximately 0.1% mortality vs 2% for COVID-19. Do the math.
stance: DENIES

belief: Coronavirus is genetically engineered.
statement: @TheMadKiwi3 @goodfoodgal nah. A biological warfare agent would kill 99% of its victims, not 2% like the corona virus. This is a naturally occurring virus.
stance: DENIES

belief: SARS-CoV-2 can survive for weeks on surfaces.
statement: Coronavirus could survive up to 9 days outside the body, study says https://t.co/JUzdJgc5Dz
stance: SUPPORTS

Now, classify the following statement as to whether 'SUPPORTS', 'DENIES', is 'NEUTRAL', or is 'UNRELATED' toward the belief below being true. Only return the classification for the statement toward the belief, and no other text.
belief: Coronavirus is caused by 5G.

statement: Their is NO “5G radiation poisoning” going on, so stop spreading this bullsh*t. The coronavirus is real, itś making people sick, and killing some, not 5G.

Context + Few Shot Prompting + Reasoning

Template examples = [
{‘belief’: {belief},
‘statement’: {statement},
‘stance’: {stance},
‘reason’: {reason}
},
‘belief’: {belief},
‘statement’: {statement},
‘stance’: {stance},
‘reason’: {reason}
},
{‘belief’: {belief},
‘statement’: {statement},
‘stance’: {stance},
‘reason’: {reason}
}
]

The following statements are social media posts about COVID or Coronavirus. Each statement can support, deny, be neutral, or be unrelated toward its associated COVID belief and Each statement has the reason for that stance.

Now, classify the following statement as to whether ’SUPPORTS’, ’DENIES’, is ’NEUTRAL’, or is ’UNRELATED’ toward the belief below being true, and give your reason for the classification. Only return the classification for the statement toward the belief and the reasoning for the classification in the form of: ’stance: STANCE, reason: REASON’

belief: {event}
statement: {statement}

Sample Prompt The following statements are social media posts about COVID or Coronavirus. Each statement can support, deny, be neutral, or be unrelated toward its associated COVID belief and Each statement has the reason for that stance.

belief: Africans are more resistant to coronavirus.
statement: Happen now Blacks are Immune to the coronavirus ’ there is a GOD https://t.co/LRq7SZYK0G
stance: SUPPORTS
reason: The statement supports the belief that Africans are more resistant to COVID, as it claims black people are immune to COVID, and most people in Africa are black.

belief: Alex Jones’ silver-infused toothpaste kills COVID-19
statement: #China #COVID-19 As work resumes in outbreak, brand-new ‘normal’ emerges https://t.co/VENOSSOnx5 https://t.co/RQoeSWoaHH
stance: UNRELATED
reason: The statement is unrelated to the belief Alex Jones toothpaste as the statement does not talk about it.

belief: COVID-19 is only as deadly as the seasonal flu.
statement: @islandmonk @Stonekettle Closer to 650,000 people will die of the flu this year. The figure of 30,000 is just in the U.S.
But the flu has approximately 0.1% mortality vs 2% for COVID-19. Do the math.
stance: DENIES
reason: The statement denies COVID being only as deadly as the flu as it cites numbers that refute this belief.

belief: Coronavirus is genetically engineered.
statement: @TheMadKiwi3 @goodfoodgal nah. A biological warfare agent would kill 99% of its victims, not 2% like the corona virus. This is a naturally occurring virus.
stance: DENIES
reason: The statement denies COVID being genetically engineered as it claims if COVID was an engineered bioweapon, it would have killed much more people than it actually did.

belief: SARS-CoV-2 can survive for weeks on surfaces.
statement: Coronavirus could survive up to 9 days outside the body, study says https://t.co/JUzdJgc5Dz
stance: SUPPORTS
reason: The statement supports the belief that COVID can survive for weeks on surfaces because it claims COVID can survive outside of a body, which implies on a surface, for over a week.

Now, classify the following statement as to whether ‘SUPPORTS’, ‘DENIES’, is ‘NEUTRAL’, or is ‘UNRELATED’ toward the belief below being true, and give your reason for the classification. Only return the classification for the statement toward the belief and the reasoning for the classification in the form of: ‘stance: STANCE, reason: REASON’
belief: Coronavirus is caused by 5G.
statement: Their is NO “5G radiation poisoning” going on, so stop spreading this bullsh*t. The coronavirus is real, it’s making people sick, and killing some, not 5G.