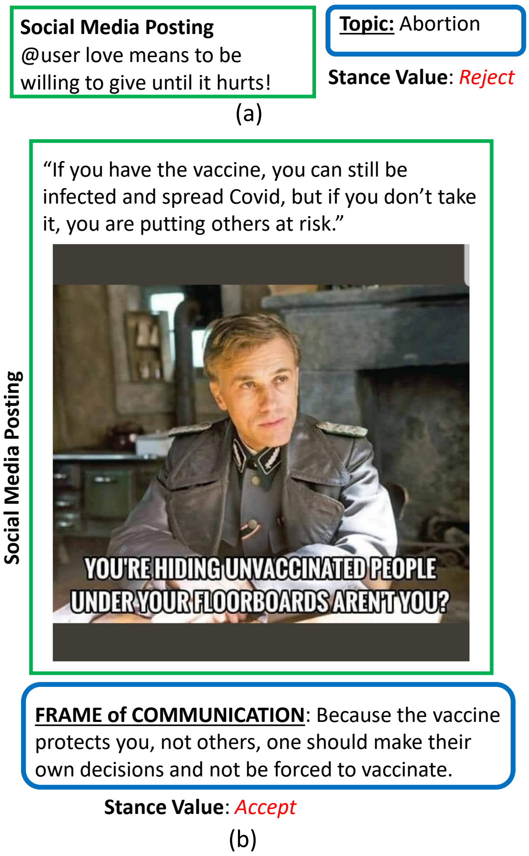
Tree-of-Counterfactual Prompting for Zero-Shot Stance Detection

Maxwell A. Weinzierl and **Sanda M. Harabagiu**
Human Language Technology Research Institute, The University of Texas at Dallas
`{maxwell.weinzierl, sanda}@utdallas.edu`

Abstract

Stance detection enables the inference of attitudes from human communications. Automatic stance identification was mostly cast as a classification problem. However, stance decisions involve complex judgments, which can be nowadays generated by prompting Large Language Models (LLMs). In this paper we present a new method for stance identification which (1) relies on a new prompting framework, called Tree-of-Counterfactual prompting; (2) operates not only on textual communications, but also on images; (3) allows more than one stance object type; and (4) requires no examples of stance attribution, thus it is a "Tabula Rasa" Zero-Shot Stance Detection (TR-ZSSD) method. Our experiments indicate surprisingly promising results, outperforming fine-tuned stance detection systems.



Social Media Posting
@user love means to be willing to give until it hurts!

Topic: Abortion

Stance Value: *Reject*

(a)

"If you have the vaccine, you can still be infected and spread Covid, but if you don't take it, you are putting others at risk."

YOU'RE HIDING UNVACCINATED PEOPLE UNDER YOUR FLOORBOARDS AREN'T YOU?

FRAME of COMMUNICATION: Because the vaccine protects you, not others, one should make their own decisions and not be forced to vaccinate.

Stance Value: *Accept*

(b)

Figure 1: (a) Example of the stance of a Social Media Posting (SMP) towards a topic; (b) example of the stance of a SMP towards a Frame of Communication (FoC).

1 Introduction

Stance was defined by Biber and Finegan (1988) as the expression of an author's standpoint and judgment towards a given proposition. As such, stance is used for discovering *attitudes* from various sources, ranging from social media, to debates or news articles, cf. Hardalov et al. (2021). Regardless of the context in which it is used, stance always has a subject and an object. The subject of stance can be the speaker in a conversation or the author of a Social Media Posting (SMP). The stance object, as reported in Hardalov et al. (2021, 2022); Liu et al. (2023c), can be sometimes a topic, while other times it can be a frame of communication. Topics typically correspond to the name of a controversial political figure, e.g. Hillary Clinton, Donald Trump, or a noun phrase, e.g. "gun control" or "abortion". Automatically inferring the stance towards a topic entails modeling complex judgments that sometimes use commonsense knowledge about the topic, as shown in the example illustrated in Figure 1(a). The SMP used in this example implies that it is best to give birth - even at the cost of pain. This is a direct quote from Mother Teresa, in a famous speech concerning abortion at a National Prayer Breakfast in Washington, D.C., on February 5, 1994. Therefore, the SMP's author has a stance of *Reject* towards the topic of "abortion".

Frames of Communication (FoCs), the other possible objects of stance that we consider in this paper, highlight problems of interest by addressing what causes those problems, and determining moral judgments, as defined in Entman (1993). The FoC

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 861–880
August 11-16, 2024 ©2024 Association for Computational Linguistics

illustrated in Figure 1(b) highlights the problem of collective responsibility during vaccination campaigns, explaining why people should make their own vaccination decisions. The SMP from this example contains both a text and an image. At first look, the FoC is contradicted by the text of the SMP, but interestingly, the image also contributes to the stance decision. The image represents the terrifying Nazi character of Hans Landa from the movie Inglourious Basterds, and the text superimposed on the image implies that he is looking for unvaccinated people to punish them. A closer inspection of the entire SMP indicates that the quoted text in fact is sarcastic, and combined with the image, it accounts for the inference of an *Accept* stance.

The examples illustrated in Figure 1 showcase some of the complex forms of reasoning required when inferring stance values. This motivated us to find whether existing Large Language Models (LLMs) and Large Multimedia Models (LMMs) can be used to access their knowledge, and, more importantly, to use their reasoning capabilities to infer the stance of communications involving both text and images. In this paper we focused on social media platforms where millions of users express their opinions and attitudes in conversations about issues of their interest, therefore we explored stance detection from SMPs.

Previous stance detection methods (Augenstein et al., 2016; Du et al., 2017; Sun et al., 2018; Siddiqua et al., 2019; Conforti et al., 2020; Miao et al., 2020) mostly used topics as stance objects, casting stance detection as a classification problem, trained on texts addressing a certain number of topics, and tested on texts addressing the same topics. However, as reported in Allaway and McKeown (2020); Allaway et al. (2021), stance detection methods designed in this way are limited in their ability to generalize to a wide variety of topics. To address this limitation, Zero-Shot Stance Detection (ZSSD) was considered in Allaway and McKeown (2020); Allaway et al. (2021); Liang et al. (2022a,b); Li and Yuan (2022), in which the stance topics from the test data are not seen in the training data. To take this a step further, in this paper we address the problem of "Tabula Rasa" Zero-Shot Stance Detection (TR-ZSSD), in which no example of stance inference towards any topic is needed, and moreover, the stance of SMPs towards topics as well as FoCs is identified.

Instead of classifying the stance value of an SMP towards a topic or an FoC, we considered a special form of counterfactual reasoning, in which we prompt an LLM or an LMM to generate an explanation if the stance s of an SMP p towards the stance object o is considered to take a value v. Since the value v is not known, we consider *counterfactuals*, i.e. all *alternative* possibilities of stance values, given the observed p and o. This entails that for each pair (p, o) we obtain (1) an explanation e_A, providing the rationale for assigning the value *Accept* to s; (2) an explanation e_R, providing the rationale of the *Reject* value; and (3) an explanation e_N, providing the rationale of the *No Stance* value. For each explanation, the LLM (or LMM) follows different reasoning paths, thus it generates a Tree-of-Thoughts, inspired by (Yao et al., 2023). Finally, the LLM (or LMM) verifies which of the three explanations is most likely, using Chain-of-Verification prompting (Dhuliawala et al., 2023), leading to the final inference of the stance value v. In all this reasoning process, based on multiple prompting of an LLM (or LMM), no need for prior examples of stance value assignments is required. We call this method *Tree-of-Counterfactual (ToC) prompting for Tabula Rasa Zero-Shot Stance Detection (ToC-TR-ZSSD)*, making the following contributions:

◁**1**▷ We introduce the first Tabula Rasa Zero-Shot Stance Detection (TR-ZSSD) method, that operates without any prior example of stance value assignment. TR-ZSSD is a step forward from previous ZSSD, in which examples of stance value assignment are provided, but new topics are not available at training time.

◁**2**▷ Our TR-ZSSD method operates not only on texts, but also on SMPs that contain images. Moreover, the method allows the stance to consider different objects, e.g. topics or FoCs.

◁**3**▷ We introduce the first stance detection method that prompts LLMs or LMMs to generate explanations of possible stance values and to verify the explanations, selecting the most rational stance value.

◁**4**▷ Our ToC prompting for stance detection refocuses the problem of stance identification on reasoning from current approaches relying on classification, enabling an explainable approach to stance identification.

We make available all code, prompts, and experiments on GitHub[1].

---

[1] `https://github.com/Supermaxman/gpt-gleam`

2 Related Work

Stance detection on social media attracted plenty of interest, spearheaded by Task 6 of SemEval-2016 (Mohammad et al., 2016), where multiple neural systems (Zarrella and Marsh, 2016; Wei et al., 2016; Tutek et al., 2016) were evaluated. The dataset used in SemEval-2016 considered tweets discussing several controversial topics (e.g., "abortion"), annotated with the stance of each tweet's text towards these topics. Continued interest was sparked by the release of TweetEval (Barbieri et al., 2020), a commonly utilized benchmark for stance detection (Nguyen et al., 2020; Loureiro et al., 2022; Du et al., 2023). Additional research interest formed around the Zero-Shot Stance Detection (ZSSD) setting, which focused on identifying the stance of novel topics, not present in the training data (Allaway et al., 2021; Liang et al., 2022a,b). Cross-topic ZSSD systems typically focus on stance datasets with hundreds or thousands of topics, available from VAST (Allaway and McKeown, 2020) and WT-WT (Conforti et al., 2020). This formulation of cross-topic ZSSD stands in contrast with the Tabula Rasa ZSSD employed in this work, in which no training examples are provided for the ToC-TR-ZSSD method. ToC-TR-ZSSD also contrasts with prior unsupervised stance detection research, which focused on identifying stance via the structural properties of the speaker graph (Murakami and Raymond, 2010; Pick et al., 2022). Furthermore, stance detection research has considered incorporating external knowledge, such as Li et al. (2023a), and In-Context Learning, such as Zhang et al. (2023a), which complement the introduction of ToC-TR-ZSSD.

Counterfactual reasoning considers alternatives to facts, events, or states, making inferences about what might have happened or what may have been possible. For example, if you trip and fall due to not tying your shoes, you can easily imagine a counterfactual reality where you tied your shoes and therefore did not trip. Counterfactual reasoning is known to play an important role in human learning and decision-making. In philosophical and AI literature (Lewis, 1973; Ortiz Jr., 1999; Pearl, 2009), counterfactuals were considered to give us reliable insights into the specifics of causality. However, recently, counterfactual reasoning has been integrated into NLP systems to enhance understanding and generation capabilities by considering altering narratives or decisions based on hypothetical scenarios. For example, Qin et al. (2019) utilize counterfactual reasoning for story re-writing, while Zeng et al. (2020) incorporated counterfactual reasoning to improve Named Entity Recognition (NER) by generating counterfactual examples. Counterfactual reasoning also yields insights into the explainability of NLP systems, with Jacovi et al. (2021) focusing on identifying and altering features in input data with counterfactuals that change model predictions, thereby providing insights into model behavior, and Robeer et al. (2021) where classification predictions are explained with counterfactuals. Recently, counterfactuals have been utilized to improve prompt design for LLMs (He et al., 2022; Chen et al., 2023), yielding improved few-shot performance on vision and language tasks by discovering more generalizable prompts. However, none of these LLM prompting techniques directly utilized LLMs to generate, compare, and explain counterfactual rationales, as we do in the ToC-TR-ZSSD method.

3 The Method

Tree-of-Counterfactual (ToC) prompting of Large Language Models (LLMs)[2] for stance detection is structured into three distinct phases.

□ Phase A, illustrated in Figure 2(A), involves constructing a tree of counterfactuals, considering each possible stance value. For instance, an *Accept* branch is constructed for the SMP p_i and the stance object o_j when we assume the stance value assigned to $s_{i,j}$ is *Accept*. Similarly, counterfactual *Reject* and *No Stance* branches are constructed. Note that each branch is always constructed, regardless of the content of p_i or o_j. However, only one of the possible stance values will hold - but we need to proceed to the next phases to find out which stance value is the most rational for p_i towards o_j. Therefore Phase A sets the stage for the following phases to perform reasoning that allows $s_{i,j}$ to be inferred.

□ Phase B illustrated in Figure 2(B), prompts the LLM to explain the reason why a certain stance value $s_{i,j}$ is attributed to the pair (p_i, o_j), for each of the possible counterfactual stance values. The prompting approach seeks to generate post hoc rationalizations for why each counterfactual stance value is held by SMP p_i towards stance object o_j.

---

[2] We use Large Language Models (LLMs) and Large Multimodal Models (LMMs) interchangeably here, depending on the modality of the SMPs under consideration.

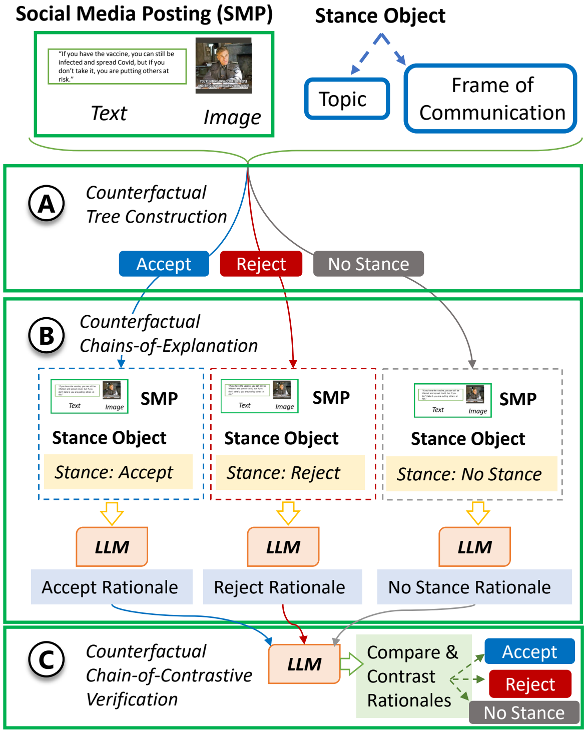

Social Media Posting (SMP)
"If you have the vaccine, you can still be infected and spread Covid, but if you don't take it, you are putting others at risk."
Text
Image
Stance Object
Topic
Frame of Communication
A
Counterfactual Tree Construction
Accept
Reject
No Stance
B
Counterfactual Chains-of-Explanation
SMP
Stance Object
Stance: Accept
Stance: Reject
Stance: No Stance
LLM
Accept Rationale
Reject Rationale
No Stance Rationale
C
Counterfactual Chain-of-Contrastive Verification
LLM
Compare & Contrast Rationales
Accept
Reject
No Stance

Figure 2: Tree-of-Counterfactual Prompting (ToC) for Zero-Shot Stance Detection.

This **Chain-of-Explanation (CoE)** prompting instructs the LLM to provide step-by-step rationales:

Chain-of-Explanation Prompting

System Prompt

You are an expert linguistic assistant.
You will be tasked with **explaining why** a tweet may have a stance towards a provided [topic|frame of communication].
(Make sure to explain how the image and text together contribute to the stance.)
You should discuss your reasoning in detail, **thinking step-by-step**.

User Prompt

Social Media Posting: {text}
[Topic|Frame of Communication]: {object}
Stance: {stance}
({image})

As can be seen, depending on the type of stance object, i.e. topic or FoC, the language of the CoE prompt is modified slightly, because of the square "[...|...]" brackets, with the pipe symbol "|" representing the possible exclusive options for the prompt. Furthermore, if multimodal stance detection is being performed with an LMM, additional sections are included in the prompt to address images, represented by round "(...)" brackets. Textual content from the SMPs is inserted into the "{text}" section, while the textual representation for the object of stance is included in the "{object}" section. Counterfactual stance values are inserted into the "{stance}" section. If multimodal stance detection is being performed, then the image context is also included in the "{image}" section.

With CoE prompting of the LLM, we obtain counterfactual rationales e_A, e_R, and e_N, one for each possible stance value. However, the final decision of the value of $s_{i,j}$ is produced in phase C.

☐ Phase C, illustrated in Figure 2(C), takes inspiration from Chain-of-Verification (CoVe) prompting, proposed by Dhuliawala et al. (2023). CoVe prompting provides LLMs the capability to deliberate on the responses they give, such that they can correct their mistakes through introspection. We employ the counterfactual rationales e_A, e_R, and e_N, produced in phase B, as the baseline responses to consider for verification, and we alter the verification step to consider a form of *contrastive verification*. In this way we create **Counterfactual Chain-of-Contrastive Verification (C-CoCV)** prompting to introspect upon the counterfactual rationales generated in Phase B, and compares their arguments with the contents of the post p_i and stance object o_j to contrast each stance value hypothesis:

Counterfactual Chain-of-Contrastive Verification Prompting

System Prompt

You are an expert linguistic assistant.
You will be tasked with **judging** which stance value a tweet has towards a provided [topic|frame of communication].
Thorough rationales will be provided for each stance value.
You should discuss your reasoning in detail, **thinking step-by-step**.
Discuss the strengths and weaknesses for each rationale, providing a **final judgement** for the stance value of the tweet towards the provided [topic|frame of communication].
Your final sentence should include only one possible stance value: Accept, Reject, or No Stance

User Prompt

Social Media Posting: {text}
[Topic|Frame of Communication]: {object}
Stance: Accept Rationale: {accept_rationale}
Stance: Reject Rationale: {reject_rationale}
Stance: No Stance Rationale: {no_stance_rationale}
({image})

As can be seen, the semantics of C-CoCV prompting uses the same semantics as CoE prompting, to allow for the various stance objects to be considered as well as for using the text and image content of the SMP or the text content of the stance objects. In addition, the counterfactual rationales generated from CoE prompting are inserted into the "{accept_rationale}", "{reject_rationale}", and "{no_stance_rationale}" sections.

The three phases of ToC prompting exploit the capacity of Large Language Models (LLMs) to generate and evaluate post hoc rationalizations for different stance hypotheses, facilitating a more explainable inference of stance, wherein each stance hypothesis is rigorously scrutinized for its validity.

Dataset	Stance Object Type	Text Content	Image Content	*Accept*	*Reject*	*No Stance*
SemEval-2016	Topic	✓	✗	304	715	230
CoVaxFrames	Frame of Communication	✓	✗	1,461	448	376
MMVax-Stance	Frame of Communication	✓	✓	578	332	642

Table 1: Dataset details and distribution of stance values for test collections from SemEval-2016 Task 6 A, CoVaxFrames, and MMVax-Stance.

This process echoes the natural human cognitive process of considering and comparing alternatives before arriving at a conclusion (Van Hoeck et al., 2015), allowing for a stance detection mechanism that is both robust and transparent. Because ToC prompting requires no use of annotated stance data or demonstrations, it can be considered a TR-ZSSD method.

4 Experimental Results

4.1 Experiments

We consider three distinct stance detection settings within which to compare and evaluate ToC-TR-ZSSD. These settings are determined by three different stance-annotated datasets, detailed in Table 1. Furthermore, as our method is entirely zero-shot, we discuss any possibilities of test collection contamination in the training data of the considered LLMs in Appendix B.

SemEval-2016 (Mohammad et al., 2016) introduced one of the first widely used stance datasets on Twitter. It addressed five controversial topics: *"Abortion"*, *"Atheism"*, *"Climate Change"*, *"Feminism"*, and *"Hillary Clinton."* An ongoing leaderboard on the SemEval 2016 task was organized by TweetEval (Barbieri et al., 2020), which fostered continued competition on the dataset. This resource has been widely used in computational linguistics and natural language processing research to develop and benchmark methods capable of detecting stance from text, contributing significantly to advancements in automatic stance detection. Statistics for the distribution of stance values in the test collection of SemEval-2016 Task 6 A are provided in Table 1.

CoVaxFrames (Weinzierl and Harabagiu, 2022b) made available stance annotations on SMPs originating from Twitter, discussing COVID-19 vaccine hesitancy. CoVaxFrames includes 113 Frames of Communication (FoCs), discovered through Question Answering as answers to questions from the Vaccine Confidence Repository (Rossen et al., 2019). Stance was annotated towards these 113 FoCs for a collection of text-only SMPs. FoCs have recently been found to be superior stance object types (Weinzierl and Harabagiu, 2024) when compared to topics, requiring less implicit knowledge to infer the stance of a text towards an FoC than a topic. CoVaxFrames includes FoCs such as *"The COVID vaccine renders pregnancies risky, and it is unsafe for unborn babies"* and *"Preference for getting COVID-19 and fighting it off than getting vaccinated."* Statistics for the distribution of stance values in the test collection of CoVaxFrames are provided in Table 1.

MMVax-Stance (Weinzierl and Harabagiu, 2023) is the only multimodal dataset that contains stance annotations, with the SMPs originating from Twitter. The objects of the stance annotations are the same 113 COVID-19 vaccine hesitancy FoCs as CoVaxFrames. The complexity of multimodal stance annotation involving the interpretation of texts as well as memes, images, and videos, presents a significant challenge, making this dataset very valuable. Statistics for the distribution of stance values in the test collection of MMVax-Stance are provided in Table 1.

4.2 Results

To evaluate the quality of stance detection on the test collections from SemEval-2016, CoVaxFrames, and MMVax-Stance we use the Precision (P), Recall (R), and F_1 metrics for detecting the *Accept* and *Reject* values of stance. We also compute a Macro averaged Precision, Recall, and F_1 score over the *Accept* and *Reject* stance values. The text-only stance detection evaluation results for SemEval-2016 are listed in Table 2, while the evaluation results for CoVaxFrames are listed in Table 3. The multimodal stance detection results for MMVax-Stance are presented in Table 4. The bolded numbers represent the best results obtained for each setting: stance detection with fine-tuning and TR-ZSSD. All hyperparameters are included in the linked GitHub repository and discussed in Appendix A.

Two LLMs were considered in our text-only TR-

System	Macro F_1	Macro P	Macro R	Accept F_1	Accept P	Accept R	Reject F_1	Reject P	Reject R
pkudblab	67.3	-	-	62.0	-	-	72.7	-	-
MITRE	67.8	-	-	59.3	-	-	76.3	-	-
SVM-ngrams	69.0	-	-	63.0	-	-	75.0	-	-
RoB-RT	69.3	-	-	-	-	-	-	-	-
RoBERTa + Enriched	70.1	69.1	73.6	-	-	-	-	-	-
BERTweet	71.2	-	-	-	-	-	-	-	-
MoLE - DANN	72.1	-	-	-	-	-	-	-	-
PNEM	72.1	-	-	66.6	-	-	77.7	-	-
TimeLMs	**72.9**	-	-	-	-	-	-	-	-
GPT-3.5-Direct	56.2	87.8	41.6	51.9	89.6	36.6	60.4	86.0	46.5
GPT-3.5-CoT	62.8	80.2	52.6	68.9	77.1	62.2	56.6	83.3	42.9
GPT-3.5-ToC	69.4	80.6	**68.7**	78.7	68.2	**93.1**	60.0	93.0	44.3
GPT-4-Direct	67.9	93.0	56.7	52.0	94.8	35.9	**83.8**	91.3	**77.5**
GPT-4-CoT	70.5	**96.8**	55.6	73.2	95.5	59.4	67.8	**98.1**	51.7
GPT-4-ToC	**77.1**	**96.8**	64.6	**83.1**	**95.7**	73.4	71.2	97.8	55.9

Table 2: Results from prior stance detection fine-tuning experiments and tabula rasa zero-shot stance detection experiments on the test collection from SemEval-2016 Task 6 A. Missing values represent unreported precision, recall, and F_1 values.

System	Macro F_1	Macro P	Macro R	Accept F_1	Accept P	Accept R	Reject F_1	Reject P	Reject R
SBERT-NLI	39.9	42.6	40.5	50.3	61.3	42.7	29.4	23.9	38.2
DS-BERT	69.1	68.8	69.5	81.0	**79.3**	82.8	57.2	58.2	56.2
LES-GAT	72.4	69.6	75.4	80.6	77.1	84.5	64.1	62.1	66.3
LES-GAT-MF	75.2	**73.0**	77.9	**83.6**	77.8	90.5	66.8	68.3	65.4
LACRScore	**76.2**	72.7	**80.7**	83.4	75.3	**93.5**	**68.9**	**70.1**	**67.8**
GPT-3.5-Direct	55.8	72.8	45.8	56.0	81.5	42.6	55.5	64.1	49.0
GPT-3.5-CoT	59.5	75.0	49.9	60.3	83.0	47.4	58.7	66.9	52.3
GPT-3.5-ToC	64.4	75.7	56.1	73.8	84.0	65.8	55.0	67.3	46.4
GPT-4-Direct	62.5	79.2	62.6	47.5	91.9	32.0	77.5	66.4	**93.1**
GPT-4-CoT	71.3	83.2	62.9	69.5	87.4	57.7	73.1	78.9	68.1
GPT-4-ToC	**79.1**	**90.2**	**70.8**	**77.5**	**93.4**	**66.2**	**80.8**	**86.9**	75.4

Table 3: Results from prior stance detection fine-tuning experiments and tabula rasa zero-shot stance detection experiments on the test collection from CoVaxFrames.

ZSSD experiments on SemEval-2016 and CoVaxFrames: GPT-3.5 (Ouyang et al., 2022), and GPT-4 (OpenAI, 2023). Both these LLMs have been shown to have very high performance in complex zero-shot and few-shot reasoning tasks (Espejel et al., 2023; Ouyang et al., 2022; OpenAI, 2023). Furthermore, two baseline prompting methods were employed with GPT-3.5 and GPT-4: Direct prompting and Chain-of-Thought (CoT) prompting (Wei et al., 2022).

Next, we discuss the results on each dataset.

SemEval-2016 Results: We compare against multiple fine-tuned systems from the original SemEval-2016 competition, including SVM-ngrams (Mohammad et al., 2016), MITRE (Zarrella and Marsh, 2016), and pkudblab's system (Wei et al., 2016). Additional fine-tuned systems were considered which participated in the TweetEval benchmark, including RoB-RT (Barbieri et al., 2020), BERTweet (Nguyen et al., 2020), RoBERTa + Enriched (Liu et al., 2023c), MoLE - DANN (Hardalov et al., 2021), PNEM (Siddiqua et al., 2019), and TimeLMs (Loureiro et al., 2022). Among these systems, the TimeLMs system achieved the highest Macro F_1 score, 72.9, demonstrating the importance of taking into account the diachronic nature of social media. Table 2 illustrates that GPT-3.5 and GPT-4 underperform nearly all the fine-tuned systems when performing TR-ZSSD with direct prompting, and only come close to fine-tuned systems when utilizing CoT prompting, demonstrating the importance of different prompting strategies. However, significant performance gains are achieved when utilizing ToC prompting for TR-ZSSD, yielding competitive zero-shot performance for GPT-3.5 and surpassing fine-tuned baselines with GPT-4. ToC prompting with GPT-4 was able to achieve a Macro F_1 score of 77.1, gaining over 4 points of F_1 score over TimeLMs, the best fine-tuned system.

CoVaxFrames Results: We considered multiple baseline systems utilized by prior work on COVAXFRAMES, including SBERT-NLI (Hossain et al., 2020), DS-BERT (Weinzierl et al., 2021), LES-GAT (Weinzierl et al., 2021), LES-GAT-MF (Weinzierl and Harabagiu, 2022b), and LACRScore (Weinzierl and Harabagiu, 2022a). The LACRScore system achieved the highest Macro F_1 score, 76.2, of the fine-tuned systems on COVAXFRAMES due to the benefits of considering attitude consistency. Table 3 illustrates that GPT-3.5 and GPT-4 similarly underperform nearly all the fine-tuned systems when performing TR-ZSSD with direct prompting, and again, come close to the fine-tuned systems when utilizing CoT prompting. However, ToC prompting again yields massive F_1 score increases, with GPT-4 yet again surpassing the fine-tuned baselines. GPT-4 with ToC prompting was able to achieve a Macro F_1 score of 79.1, gaining nearly 3 points of F_1 score over the LACRScore system with zero training examples from COVAXFRAMES.

MMVax-Stance Results: For our multimodal TR-ZSSD experiments on MMVAX-STANCE, two LMMs were considered: LLaVA-1.5 (Liu et al., 2023b,a), and GPT-4V (OpenAI, 2024). LLaVA connects a vision encoder and an LLM to form a large multimodal model designed for general-purpose visual and language understanding, with additional instruction tuning from GPT-4. LLaVA-1.5 expands upon the LLaVA architecture by incorporating a Multilayer Perceptron (MLP) as a vision-language connector and adding academic-task-oriented Visual Question Answering (VQA) data to the pre-training tasks. GPT-4V has been shown to demonstrate a promising ability to align with human judgments across numerous multimodal tasks (Zhang et al., 2023b), with GPT-4V scoring top of the leaderboards for Cognition subtasks on the MME leaderboards (Fu et al., 2023). Two baseline prompting methods were employed with LLaVA-1.5 and GPT-4V: Direct prompting and CoT prompting.

DS-BERT, LES-GAT-MF, and LACRscore systems, employed previously on COVAXFRAMES, were utilized as baseline fine-tuned systems on MMVAX-STANCE, along with image captioning performed by BLIP-2 (Li et al., 2023b), as per Weinzierl and Harabagiu (2023). Additionally, CLIP-Joint utilized the text and the image encoder of CLIP (Radford et al., 2021) to perform stance classification, while VILT (Kim et al., 2021), FLAVA (Singh et al., 2022), and BridgeTower (BT) (Xu et al., 2023) were also fine-tuned for stance classification. Additionally, BT + S_{All} (Weinzierl and Harabagiu, 2023) was included, a BridgeTower system employing significant training data augmentation. The BT system achieved a Macro F_1 score of 52.9 on MMVAX-STANCE, struggling to achieve performance parity with text-based stance detection systems due to the increased difficulty of multimodal stance detection. Table 4 illustrates that LLaVA-1.5 and GPT-4V underperform the fine-tuned systems in TR-ZSSD with direct prompting, and become very competitive with the fine-tuned systems with CoT prompting. Furthermore, ToC prompting again yields massive F_1 score gains, with ToC prompting with GPT-4V achieving a Macro F_1 score of 60.6, an increase of nearly 8 points of F_1 score over BT with no multimodal training examples from MMVAX-STANCE. However, GPT-4V with ToC prompting failed to outperform the best fine-tuned system, which employed massive additional synthetic data. BT + S_{All} achieved a Macro F_1 score of 71.3 with 46,606 additional synthetic multimodal stance examples upon which the system was trained - representing neatly a 600% increase in training data (Weinzierl and Harabagiu, 2023).

4.3 Discussion

Achieving a Macro F_1 score of 77.1 on SemEval-2016, 79.1 on COVAXFRAMES, and 60.6 on MMVAX-STANCE in TR-ZSSD demonstrate that ToC-TR-ZSSD performs surprisingly well, as it outperforms fine-tuned supervised systems on text-only stance detection and achieves competitive performance for stance detection from multimedia SMPs. We believe that this is in large part due to ToC prompting. Similar F_1 scores were unattainable with zero training data only a few years ago, with systems typically requiring at least hundreds, if not thousands, of training examples. Furthermore, accessing the reasoning capabilities of LLMs with improved prompting methodologies continued to yield improvements in stance detection, with ToC prompting generating the most impressive results. We hypothesize that the superiority of ToC prompting, compared to CoT prompting, is due to the fact that CoT prompting requires LLMs to immediately commit to a rationale that supports the final answer, which provides little space for

System	Macro F_1	Macro P	Macro R	Accept F_1	Accept P	Accept R	Reject F_1	Reject P	Reject R
DS-BERT + Caps	44.1	44.9	43.5	50.6	49.1	52.1	37.6	40.7	34.9
LES-GAT-MF + Caps	42.7	42.1	43.2	48.2	47.0	49.5	37.1	37.3	36.9
LACRscore + Caps	44.4	43.3	45.5	50.1	49.3	50.9	38.7	37.4	40.1
CLIP-Joint	46.8	43.4	53.8	56.3	45.6	73.5	37.3	41.2	34.0
VILT	44.4	38.6	52.5	53.5	48.4	59.9	35.2	28.9	45.2
FLAVA	47.6	42.9	53.5	55.8	49.6	63.7	39.5	36.3	43.4
BT	52.9	46.8	60.9	61.7	53.9	72.3	44.1	39.8	49.4
BT + S_{All}	**71.3**	**71.5**	**71.2**	**79.5**	**78.6**	**80.3**	**63.2**	**64.4**	**62.1**
LLaVA-1.5-Direct	33.8	48.3	26.2	36.3	56.2	26.8	31.3	40.4	25.6
LLaVA-1.5-CoT	41.5	48.7	36.5	47.1	59.2	39.1	35.9	38.2	33.9
LLaVA-1.5-ToC	46.8	54.6	42.5	53.8	53.6	**54.0**	39.8	55.7	31.0
GPT-4V-Direct	48.8	60.7	41.9	45.6	66.1	34.8	51.9	55.3	48.9
GPT-4V-CoT	52.6	62.8	45.9	50.7	67.3	40.7	54.5	58.3	51.1
GPT-4V-ToC	**60.6**	**71.6**	**54.0**	**56.8**	**77.5**	44.8	**64.4**	**65.6**	**63.3**

Table 4: Results from prior multimodal stance detection fine-tuning experiments and tabula rasa multimodal zero-shot stance detection experiments on the test collection from MMVAX-STANCE.

LLMs to consider all possible alternative hypotheses. Improvements to CoT prompting have been proposed to alleviate this issue, such as Self Consistency with CoT (CoT-SC) prompting (Wang et al., 2023), which samples multiple Chains-of-Thought from an LLM and considers the majority voted final answer from all chains, and Tree-of-Thought (ToT) prompting (Yao et al., 2023), which maintains a tree of thoughts by sampling many thought continuations from an LLM and evaluating early whether each of those should continue or end. However, fundamentally these CoT-based approaches provide little space for the LLM to explore alternative hypotheses, as each chain or tree path must still commit early to an accurate final answer. Furthermore, sampling from the next-word prediction probabilities produced by LLMs is unlikely to yield varied rationales and conclusions, which greatly limits the considered hypothesis space. This highlights a general problem with sampling from LLMs, which can result in bias in the generated text that tends to lack diversity when generating text from underrepresented groups (Lahoti et al., 2023). ToC prompting alleviates these concerns of hypotheses exploration by explicitly considering each of the counterfactual stance values.

Exploring examples of counterfactual rationales generated by GPT-4 provides insights into why ToC prompting improves upon baseline prompting methods. ToC prompting consists of first performing CoE prompting with each counterfactual stance value, and then finally performing C-CoCV prompting to determine a final stance value. Figure 3 illustrates an example from the SemEval-2016 Task 6 A test set, employing CoE prompting with GPT-4 with a counterfactual stance value of *Accept*. We find that GPT-4 generates an interesting analysis of how the SMP could hold an *Accept* stance towards the topic of *"Climate Change"*, possibly due to the ambiguity and context in which the SMP was made. Figure 4 illustrates the same SMP, but a different counterfactual stance value of *Reject* towards *"Climate Change"*. GPT-4 generates a convincing rationale, with how the language of the SMP implies a *"...rejection of the mainstream understanding and urgency of climate change."* Finally, Figure 5 illustrates the same SMP, but with the final counterfactual stance value of *No Stance* towards *"Climate Change"*. Herein, GPT-4 rejects the *No Stance* value entirely, only being able to find evidence to the contrary. This ToC prompting example ends with C-CoCV prompting on this same SMP, with the response by GPT-4 provided in Figure 6. GPT-4 compares and contrasts the counterfactual hypotheses and evidence gathered, and comes to the ultimate conclusion that the stance of the SMP towards *"Climate Change"* is *Reject*. This conclusion aligns with the annotated stance value of *Reject* in the SemEval-2016 test set. An additional exploration of counterfactual rationales that considers diverse hypotheses is provided in Appendix C.

The ToC results when prompting GPT-4V on MMVAX-STANCE demonstrate significant advancement in multimodal reasoning. Multimodal stance detection is extremely challenging, demonstrated by the performance of the fine-tuned baseline systems, as it requires the integration of textual and visual cues to accurately interpret nuanced and often sarcastic or ironic expressions - an extremely challenging linguistic phenomenon, as reported in Joshi et al. (2017). This complexity demands advanced computational techniques capable of understanding not only the literal content but also the contextual and cultural nuances that influence interpretation. We believe that the results of using ToC prompting with GPT-4V clearly demonstrate that LMMs are capable of counterfactual reasoning, probably capturing cultural knowledge and handling well sarcastic multimodal content.

A quantitative and qualitative analysis of the mistakes made with ToC prompting was performed by researchers at the University of Texas at Dallas. On SemEval-2016, 412 incorrect stance values were predicted for the test set by GPT-4. A majority of mistakes (76%) were made on SMPs labeled with the *Reject* stance, with a majority of these examples (74%) being incorrectly identified as *No Stance* by GPT-4. Similarly, of the mistakes made on SMPs labeled with an *Accept* stance, 81% were misidentified as having *No Stance* by GPT-4. Furthermore, GPT-4 outright refused to process 1% of the SMPs from SemEval-2016 due to the controversial nature of the SMPs. These SMPs were assigned a default stance value of *No Stance* for evaluation, but this assignment does not account for the significant number of false negatives produced by GPT-4.

An analysis of the 723 mistakes by GPT-4 made on the test set from CoVaxFrames reveals a similar pattern: A majority of mistakes (68%) were made on SMPs labeled with the *Accept* stance, with a majority of these examples (61%) being incorrectly identified as *No Stance* by GPT-4. Of the mistakes made on SMPs labeled with a *Reject* stance, 81% were again misidentified as having *No Stance* by GPT-4. Additionally, GPT-4 outright refused to process 1% of the SMPs from CoVaxFrames due to the controversial nature of the SMPs. Analyzing the 599 mistakes made by GPT-4V on the test set from MMVax-Stance emphasized the challenge of utilizing OpenAI models for sensitive tasks: GPT-4V outright refused to process 2% of the SMPs from MMVax-Stance, primarily due to the controversial nature of the images and memes utilized by these SMPs. A majority of mistakes (50%) were made on SMPs labeled with the *Accept* stance, with a majority of these examples (69%) being incorrectly identified as *No Stance* by GPT-4V. Inspecting the SMPs labeled with the *Reject* stance, 84% of the mistakes were again misidentified as having *No Stance* by GPT-4V.

Manually inspecting the counterfactual rationales and verification rationales elucidated why many examples were misidentified as having *No Stance*: GPT-4 & GPT-4V were extremely conservative in assigning *Accept* or *Reject* stance values to SMPs discussing sensitive topics such as *"Abortion"*, or sensitive FoCs such as *"The COVID-19 Vaccine is a satanic plan to microchip people."* Mitigating this conservative bias in favor of improving stance detection performance could yield significant improvements in TR-ZSSD. However, this bias is likely present due to the safety tuning performed by OpenAI, and therefore caution should be taken when attempting to bypass these safeguards.

The ToC-TR-ZSSD results support the recommendation that additional research should focus on more advanced prompting methods to get the most out of knowledge available in LLMs and to unleash their reasoning capabilities. Significant attention is currently being paid to in-context learning (Dong et al., 2023) and selecting good demonstrations for few-shot learning (Brown et al., 2020; Zhao et al., 2021). However, we believe the results from this paper demonstrate that prompting methods that creatively exploit the problem domain can yield similar performance improvements with LLMs, without requiring demonstrations.

5 Conclusion

In this paper we introduce the task of Tabula Rasa Zero-Shot Stance Detection (TR-ZSSD). We demonstrate that Large Language Models (LLMs) and Large Multimodal Models (LMMs) are capable of performing high-quality TR-ZSSD with Tree-of-Counterfactual (ToC) prompting, a novel prompting methodology relying on counterfactual reasoning. Stance detection experiments were performed over several varied stance datasets, and numerous strong baseline systems were considered for comparison. ToC prompting with LLMs, such as GPT-4, outperformed fine-tuned baselines and alternative prompting methodologies for text-based TR-ZSSD. Furthermore, ToC prompting with LMMs, such as GPT-4V, demonstrated competitive performance with fine-tuned baselines for multimodal TR-ZSSD. Our ToC-TR-ZSSD methodology required zero labeled examples, successfully generalized across different objects of stance (i.e. topics or frames of communication), improved performance across both text-only and multimodal stance detection, and generated explanatory rationales, enabling LLM and LMM interpretability.

6 Limitations

The method introduced in this work was tested only on SMPs from Twitter / X. Our approach may not work as well on SMPs originating from other social media platforms, particularly platforms such as Reddit, where longer textual content is typical. Furthermore, our method relies on the textual content of SMPs along with included images. Many SMPs use multiple images, GIFs, videos, and other multimedia content that our current method would not utilize. In future work, we plan to extend our methods by enabling them to operate across more modalities and longer-form content on additional social media platforms.

An important limitation of our approach derives from the need to have Large Language Models (LLMs) or Large Multimodal Models (LMMs) infused with significant cultural and moral knowledge. While current LLMs demonstrate strong cultural and moral understanding in majority English-speaking countries, LLMs have been found to be less accurate in predicting empirical moral norms and cultural knowledge across non-majority English-speaking different countries (Ramezani and Xu, 2023). Furthermore, while a limited form of reasoning appears as an emergent ability in LLMs at scale, it remains unclear to what extent LLMs are truly capable of reasoning (Huang and Chang, 2023). In future work, we plan to measure and address these biases by discovering and employing more diverse Frames of Communication on a wide swath of social media platforms.

7 Ethical Statement

We respected the privacy and honored the confidentiality of the users that have produced the posts in SemEval-2016, CoVaxFrames, and MMVax-Stance. We received approval from the Institutional Review Board at the University of Texas at Dallas for working with these Twitter social media datasets. IRB-21-515 stipulated that our research met the criteria for exemption #8(iii) of the Chapter 45 of Federal Regulations Part 46.101.(b). Experiments adhered to rigorous professional standards, with test collection evaluations deferred until a final method was chosen based on training and validation performance. All experimental settings, configurations, and procedures were clearly laid out in this work, the supplemental material, and the linked GitHub repository. We believe our research carries minimal risks, given its purpose to enhance comprehension of the adoption or rejection of COVID-19 vaccine hesitancy frames on social media. The central focus of the conducted research was the public good, aiming to benefit both natural language processing and public health research.

References

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.

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

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in english. *Discourse Processes*, 11(1):1–34.

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.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models.

In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. 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.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Chunhui Du, Jidong Tian, Haoran Liao, Jindou Chen, Hao He, and Yaohui Jin. 2023. Task-level thinking steps help large language models for challenging classification task. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2454–2470, Singapore. Association for Computational Linguistics.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3988–3994.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. 2023. Gpt-3.5 vs gpt-4: Evaluating chatgpt's reasoning performance in zero-shot learning.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models.

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

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Xuehai He, Diji Yang, Weixi Feng, Tsu-Jui Fu, Arjun Akula, Varun Jampani, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Wang. 2022. CPL: Counterfactual prompt learning for vision and language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3407–3418, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5).

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

David K. Lewis. 1973. *Counterfactuals*. Blackwell, Malden, Mass.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023a. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717, Singapore. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022a. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2738–2747, New York, NY, USA. Association for Computing Machinery.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Lin Miao, Mark Last, and Marina Litvak. 2020. Twitter data augmentation for monitoring public opinion on COVID-19 intervention measures. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. 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.

Akiko Murakami and Rudy Raymond. 2010. Support or oppose? classifying positions in online debates from reply activities and opinion expressions. In *Coling 2010: Posters*, pages 869–875, Beijing, China. Coling 2010 Organizing Committee.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report.

OpenAI. 2024. GPT-4V(ision) system card.

Charles L. Ortiz Jr. 1999. Explanatory update theory: Applications of counterfactual reasoning to causation. *Artificial Intelligence*, 108(1):125–178.

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.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, USA.

Ron Korenblum Pick, Vladyslav Kozhukhov, Dan Vilenchik, and Oren Tsur. 2022. Stem: Unsupervised structural embedding for stance detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11174–11182.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Marcel Robeer, Floris Bex, and Ad Feelders. 2021. Generating realistic natural language counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3611–3625, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Isabel Rossen, Mark J. Hurlstone, Patrick D. Dunlop, and Carmen Lawrence. 2019. Accepters, fence sitters, or rejecters: Moral profiles of vaccination attitudes. *Social Science & Medicine*, 224:23–27.

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

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Martin Tutek, Ivan Sekulić, Paula Gombar, Ivan Paljak, Filip Čulinović, Filip Boltužić, Mladen Karan, Domagoj Alagić, and Jan Šnajder. 2016. TakeLab at SemEval-2016 task 6: Stance classification in tweets using a genetic algorithm based ensemble. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 464–468, San Diego, California. Association for Computational Linguistics.

Nicole Van Hoeck, Patrick D. Watson, and Aron K. Barbey. 2015. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience*, 9.

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 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*, volume 35, pages 24824–24837. Curran Associates, Inc.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 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)*, pages 384–388, San Diego, California. Association for Computational Linguistics.

Maxwell Weinzierl and Sanda Harabagiu. 2022a. Identifying the adoption or rejection of misinformation targeting covid-19 vaccines in twitter discourse. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3196–3205, New York, NY, USA. Association for Computing Machinery.

Maxwell Weinzierl and Sanda Harabagiu. 2023. Identification of multimodal stance towards frames of communication. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Maxwell Weinzierl, Suellen Hopfer, and Sanda M. Harabagiu. 2021. Misinformation adoption or rejection in the era of covid-19. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):787–795.

Maxwell A. Weinzierl and Sanda M. Harabagiu. 2022b. From hesitancy framings to vaccine hesitancy profiles: A journey of stance, ontological commitments and moral foundations. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1087–1097.

Maxwell A. Weinzierl and Sanda M. Harabagiu. 2024. The impact of stance object type on the quality of stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15942–15954, Torino, Italia. ELRA and ICCL.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023. Bridgetower: Building bridges between encoders in vision-language representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10637–10647.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Guido Zarrella and Amy Marsh. 2016. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California. Association for Computational Linguistics.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280, Online. Association for Computational Linguistics.

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.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023b. Gpt-4v(ision) as a generalist evaluator for vision-language tasks.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Prompting Hyperparameters

All prompting experiments employed through the OpenAI API utilized the same hyperparameters: 1024 max generated tokens, a temperature of 1.0, and 0.7 top-p sampling. GPT-4 utilized the gpt-4-1106-preview model, while GPT-3.5 employed the gpt-3.5-turbo-1106 model. GPT-4V made use of the gpt-4-vision-preview model, while LLaVA-1.5 employed the official LLaVA GitHub code[3], with minor modifications. LLaVA hyperparameters also mimic those utilized with the OpenAI models, with max generated tokens of 1024, a temperature of 1.0, and 0.7 top-p sampling. Complete details are provided in the linked GitHub repository, including configuration files with all hyperparameters utilized and prompts employed.

B Data Contamination Discussion

As ToC-TR-ZSSD is an entirely zero-shot method, it is paramount that examples from the test collections of SemEval-2016 Task 6 A, COVAXFRAMES, and MMVAX-STANCE were not included in the training of LLMs, such as GPT-3.5, GPT-4, and LLaVA-1.5. GPT-3.5 and GPT-4 may have seen some parts of SemEval-2016 Task 6 A, as we are not privy to the secret datasets infused into these methods by OpenAI. However, we believe this is unlikely, due to the low performance of direct prompting for both models on this dataset. Regardless, we believe the experiments on SemEval-2016 Task 6 A still demonstrate that ToC prompting improves upon other zero-shot prompting methods for this dataset.

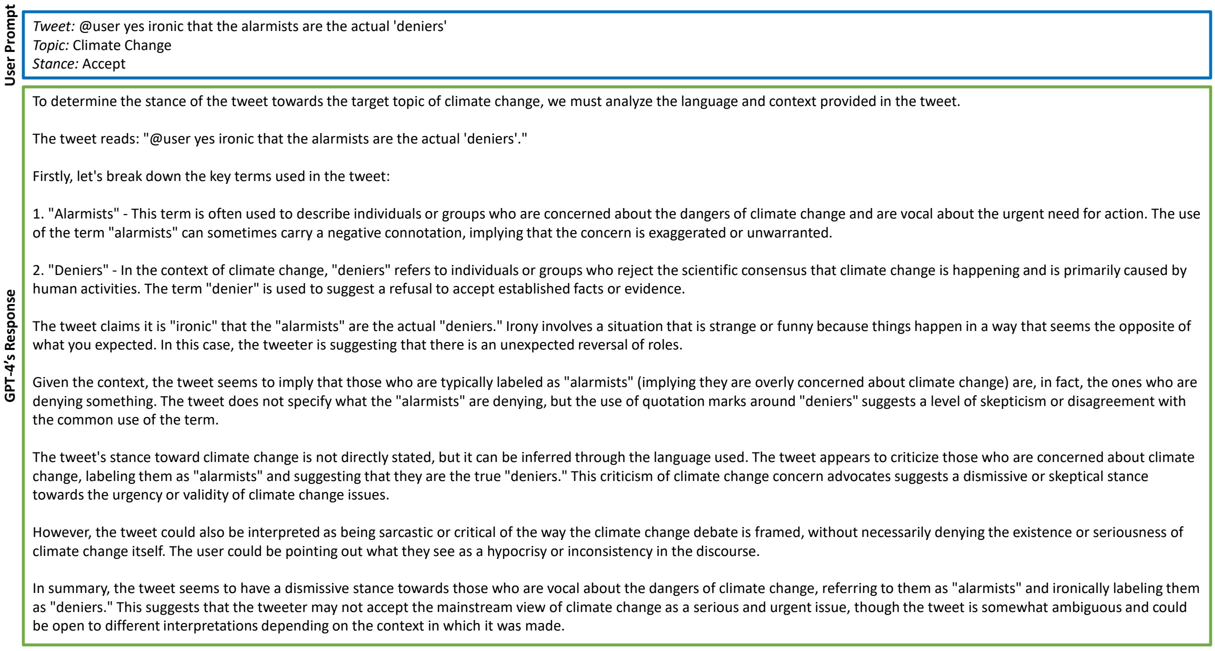
COVAXFRAMES was published in late May, 2022, while MMVAX-STANCE was published in early October, 2023. Both datasets do not have publicly available raw tweets at the time of publication to Twitter / X Developer TOS / IRB concerns. Access to the raw data requires utilizing the Twitter / X API to “hydrate" these tweet IDs. GPT-3.5 is stated to have a knowledge cutoff of September 2021[4], and therefore cannot have access to COVAXFRAMES or MMVAX-STANCE. GPT-4 is stated to have a knowledge cutoff of April 2023[5], and therefore cannot have access to MMVAX-STANCE, and is unlikely to have access to the raw Twitter

---

[3] https://github.com/haotian-liu/LLaVA

[4] https://platform.openai.com/docs/models/gpt-3-5-turbo

[5] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo


User Prompt
GPT-4's Response

Tweet: @user yes ironic that the alarmists are the actual 'deniers'
Topic: Climate Change
Stance: Accept

To determine the stance of the tweet towards the target topic of climate change, we must analyze the language and context provided in the tweet.

The tweet reads: "@user yes ironic that the alarmists are the actual 'deniers'."

Firstly, let's break down the key terms used in the tweet:

1. "Alarmists" - This term is often used to describe individuals or groups who are concerned about the dangers of climate change and are vocal about the urgent need for action. The use of the term "alarmists" can sometimes carry a negative connotation, implying that the concern is exaggerated or unwarranted.

2. "Deniers" - In the context of climate change, "deniers" refers to individuals or groups who reject the scientific consensus that climate change is happening and is primarily caused by human activities. The term "denier" is used to suggest a refusal to accept established facts or evidence.

The tweet claims it is "ironic" that the "alarmists" are the actual "deniers." Irony involves a situation that is strange or funny because things happen in a way that seems the opposite of what you expected. In this case, the tweeter is suggesting that there is an unexpected reversal of roles.

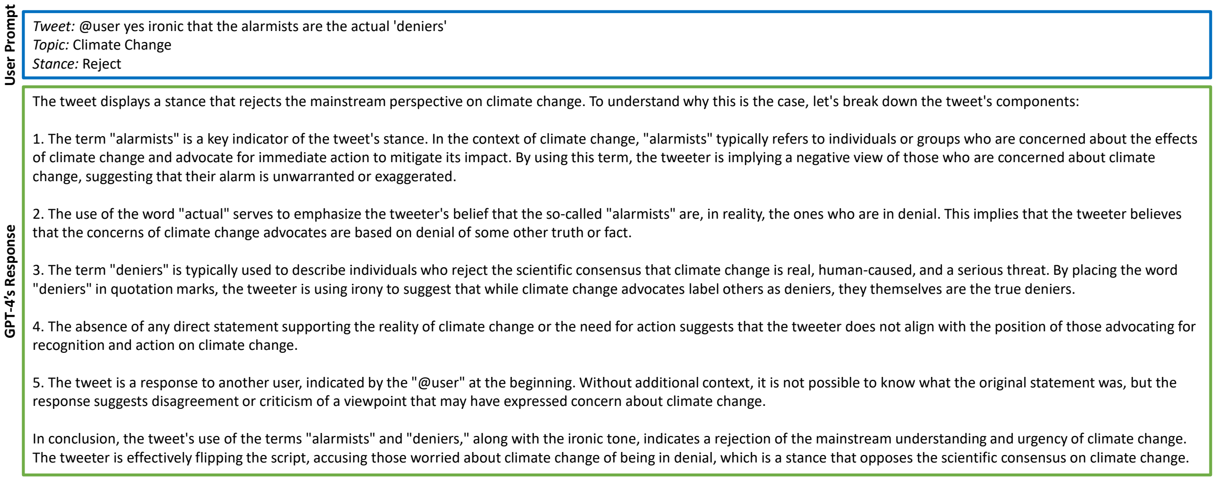
Given the context, the tweet seems to imply that those who are typically labeled as "alarmists" (implying they are overly concerned about climate change) are, in fact, the ones who are denying something. The tweet does not specify what the "alarmists" are denying, but the use of quotation marks around "deniers" suggests a level of skepticism or disagreement with the common use of the term.

The tweet's stance toward climate change is not directly stated, but it can be inferred through the language used. The tweet appears to criticize those who are concerned about climate change, labeling them as "alarmists" and suggesting that they are the true "deniers." This criticism of climate change concern advocates suggests a dismissive or skeptical stance towards the urgency or validity of climate change issues.

However, the tweet could also be interpreted as being sarcastic or critical of the way the climate change debate is framed, without necessarily denying the existence or seriousness of climate change itself. The user could be pointing out what they see as a hypocrisy or inconsistency in the discourse.

In summary, the tweet seems to have a dismissive stance towards those who are vocal about the dangers of climate change, referring to them as "alarmists" and ironically labeling them as "deniers." This suggests that the tweeter may not accept the mainstream view of climate change as a serious and urgent issue, though the tweet is somewhat ambiguous and could be open to different interpretations depending on the context in which it was made.

Figure 3: Example of a counterfactual rationale generated with Chain-of-Explanation prompting with GPT-4 for the Accept stance from SemEval-2016.


User Prompt
GPT-4's Response

Tweet: @user yes ironic that the alarmists are the actual 'deniers'
Topic: Climate Change
Stance: Reject

The tweet displays a stance that rejects the mainstream perspective on climate change. To understand why this is the case, let's break down the tweet's components:

1. The term "alarmists" is a key indicator of the tweet's stance. In the context of climate change, "alarmists" typically refers to individuals or groups who are concerned about the effects of climate change and advocate for immediate action to mitigate its impact. By using this term, the tweeter is implying a negative view of those who are concerned about climate change, suggesting that their alarm is unwarranted or exaggerated.

2. The use of the word "actual" serves to emphasize the tweeter's belief that the so-called "alarmists" are, in reality, the ones who are in denial. This implies that the tweeter believes that the concerns of climate change advocates are based on denial of some other truth or fact.

3. The term "deniers" is typically used to describe individuals who reject the scientific consensus that climate change is real, human-caused, and a serious threat. By placing the word "deniers" in quotation marks, the tweeter is using irony to suggest that while climate change advocates label others as deniers, they themselves are the true deniers.

4. The absence of any direct statement supporting the reality of climate change or the need for action suggests that the tweeter does not align with the position of those advocating for recognition and action on climate change.

5. The tweet is a response to another user, indicated by the "@user" at the beginning. Without additional context, it is not possible to know what the original statement was, but the response suggests disagreement or criticism of a viewpoint that may have expressed concern about climate change.

In conclusion, the tweet's use of the terms "alarmists" and "deniers," along with the ironic tone, indicates a rejection of the mainstream understanding and urgency of climate change. The tweeter is effectively flipping the script, accusing those worried about climate change of being in denial, which is a stance that opposes the scientific consensus on climate change.

Figure 4: Example of a counterfactual rationale generated with Chain-of-Explanation prompting with GPT-4 for the Reject stance from SemEval-2016.

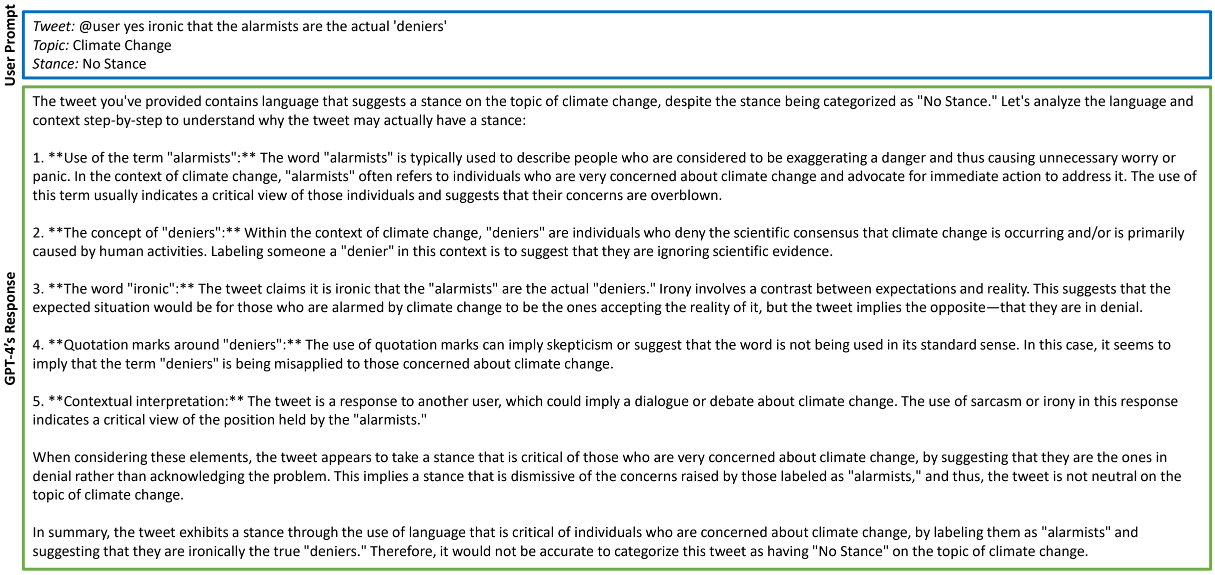
/ X data for CoVaxFrames. Finally, LLaVA-1.5 is stated to have been trained on: 558K filtered image-text pairs from LAION/CC/SBU, captioned by BLIP, 158K GPT-generated multimodal instruction-following data, 450K academic-task-oriented VQA data mixture, and 40K ShareGPT data[6]. Therefore, we believe it is highly unlikely that LLaVA-1.5 contains any of the data from these three datasets.

C Examples of Counterfactual Rationales

Figure 7 illustrates an example from the CoVaxFrames test set, employing CoE prompting with GPT-4 with a counterfactual stance value of *Accept*. GPT-4 thoroughly outlines the concerns of the SMP relating to the FoC under question: "Homeopathic / Oriental medicine / alternative medicine / naturopath may Be Alternative to COVID-19 Vaccine." A strong argument is made that the SMP is in support of alternative medicine options when compared to the COVID-19 vaccines, by "... acknowledging and defending alternative medicine as a legitimate option..." Upon considera-

---

[6] `https://huggingface.co/liuhaotian/llava-v1.5-7b`


User Prompt
GPT-4's Response

Tweet: @user yes ironic that the alarmists are the actual 'deniers'
Topic: Climate Change
Stance: No Stance

The tweet you've provided contains language that suggests a stance on the topic of climate change, despite the stance being categorized as "No Stance." Let's analyze the language and context step-by-step to understand why the tweet may actually have a stance:

1. **Use of the term "alarmists":** The word "alarmists" is typically used to describe people who are considered to be exaggerating a danger and thus causing unnecessary worry or panic. In the context of climate change, "alarmists" often refers to individuals who are very concerned about climate change and advocate for immediate action to address it. The use of this term usually indicates a critical view of those individuals and suggests that their concerns are overblown.

2. **The concept of "deniers":** Within the context of climate change, "deniers" are individuals who deny the scientific consensus that climate change is occurring and/or is primarily caused by human activities. Labeling someone a "denier" in this context is to suggest that they are ignoring scientific evidence.

3. **The word "ironic":** The tweet claims it is ironic that the "alarmists" are the actual "deniers." Irony involves a contrast between expectations and reality. This suggests that the expected situation would be for those who are alarmed by climate change to be the ones accepting the reality of it, but the tweet implies the opposite—that they are in denial.

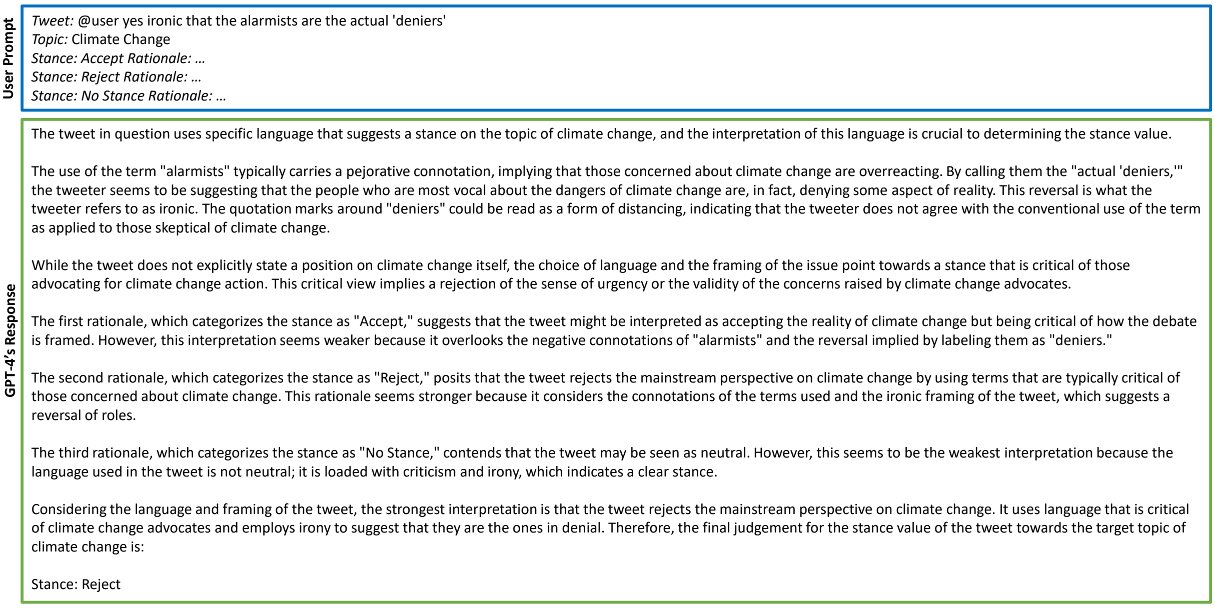
4. **Quotation marks around "deniers":** The use of quotation marks can imply skepticism or suggest that the word is not being used in its standard sense. In this case, it seems to imply that the term "deniers" is being misapplied to those concerned about climate change.

5. **Contextual interpretation:** The tweet is a response to another user, which could imply a dialogue or debate about climate change. The use of sarcasm or irony in this response indicates a critical view of the position held by the "alarmists."

When considering these elements, the tweet appears to take a stance that is critical of those who are very concerned about climate change, by suggesting that they are the ones in denial rather than acknowledging the problem. This implies a stance that is dismissive of the concerns raised by those labeled as "alarmists," and thus, the tweet is not neutral on the topic of climate change.

In summary, the tweet exhibits a stance through the use of language that is critical of individuals who are concerned about climate change, by labeling them as "alarmists" and suggesting that they are ironically the true "deniers." Therefore, it would not be accurate to categorize this tweet as having "No Stance" on the topic of climate change.

Figure 5: Example of a counterfactual rationale generated with Chain-of-Explanation prompting with GPT-4 for No Stance from SemEval-2016.


User Prompt
GPT-4's Response

Tweet: @user yes ironic that the alarmists are the actual 'deniers'
Topic: Climate Change
Stance: Accept Rationale: ...
Stance: Reject Rationale: ...
Stance: No Stance Rationale: ...

The tweet in question uses specific language that suggests a stance on the topic of climate change, and the interpretation of this language is crucial to determining the stance value.

The use of the term "alarmists" typically carries a pejorative connotation, implying that those concerned about climate change are overreacting. By calling them the "actual 'deniers,'" the tweeter seems to be suggesting that the people who are most vocal about the dangers of climate change are, in fact, denying some aspect of reality. This reversal is what the tweeter refers to as ironic. The quotation marks around "deniers" could be read as a form of distancing, indicating that the tweeter does not agree with the conventional use of the term as applied to those skeptical of climate change.

While the tweet does not explicitly state a position on climate change itself, the choice of language and the framing of the issue point towards a stance that is critical of those advocating for climate change action. This critical view implies a rejection of the sense of urgency or the validity of the concerns raised by climate change advocates.

The first rationale, which categorizes the stance as "Accept," suggests that the tweet might be interpreted as accepting the reality of climate change but being critical of how the debate is framed. However, this interpretation seems weaker because it overlooks the negative connotations of "alarmists" and the reversal implied by labeling them as "deniers."

The second rationale, which categorizes the stance as "Reject," posits that the tweet rejects the mainstream perspective on climate change by using terms that are typically critical of those concerned about climate change. This rationale seems stronger because it considers the connotations of the terms used and the ironic framing of the tweet, which suggests a reversal of roles.

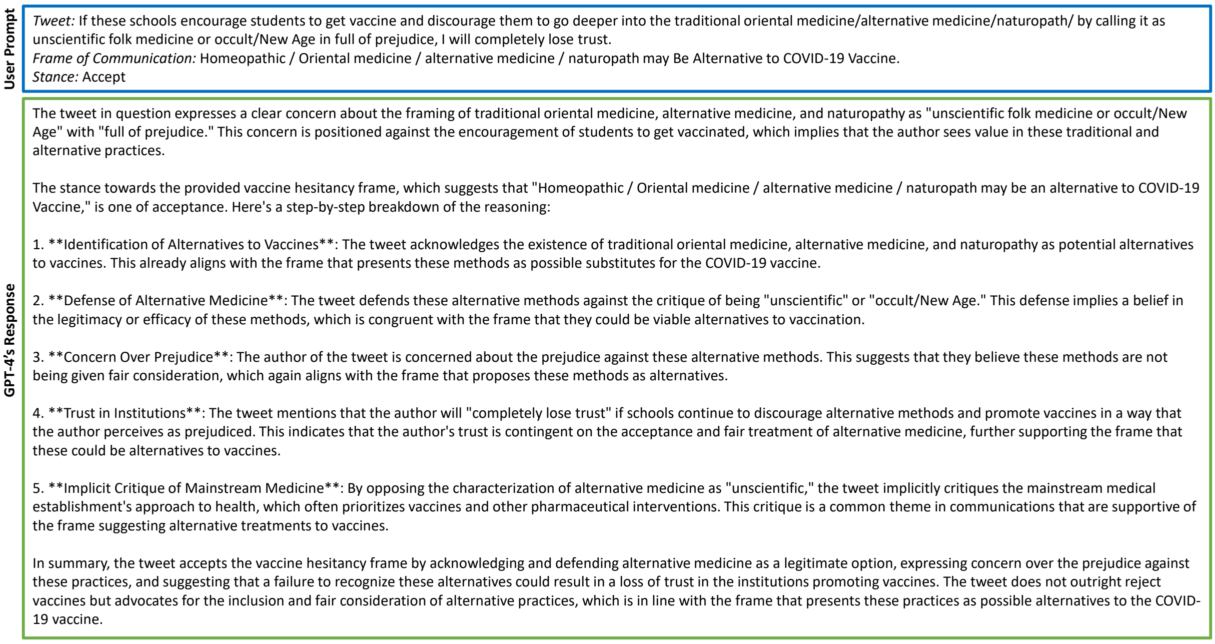
The third rationale, which categorizes the stance as "No Stance," contends that the tweet may be seen as neutral. However, this seems to be the weakest interpretation because the language used in the tweet is not neutral; it is loaded with criticism and irony, which indicates a clear stance.

Considering the language and framing of the tweet, the strongest interpretation is that the tweet rejects the mainstream perspective on climate change. It uses language that is critical of climate change advocates and employs irony to suggest that they are the ones in denial. Therefore, the final judgement for the stance value of the tweet towards the target topic of climate change is:

Stance: Reject

Figure 6: Example of a verification rationale generated with Counterfactual Chain-of-Contrastive Verification prompting with GPT-4 from SemEval-2016.

tion of the *Reject* counterfactual, Figure 8 demonstrates that GPT-4 possesses the capability to reason counterfactually, recognizing that "...the SMP's author is advocating for a more inclusive and less biased consideration of different medical traditions rather than outright stating that these traditions should replace the vaccine." This hypothesis would be extremely unlikely to arise through direct or CoT prompting approaches, as it requires one to assume that the post has a *Reject* stance towards the "homeopathic alternatives to the COVID-19 vaccine" FoC to make such nuanced distinctions between "advocating" for homeopathic alternatives versus seeking to "replace the vaccine." Figure 9 further explores the *No Stance* counterfactual, demonstrating again a nuanced consideration of the SMP concerning the FoC. Finally, Figure 10 demonstrates how GPT-4 is able to take each of these nuanced hypotheses and weigh the evidence. GPT-4 assesses that the *Accept* rationale is "...strong because it highlights the SMP's alignment with the frame's suggestion...", while the *Reject* rationale "...seems weaker" and the *No Stance* rationale "...fails to recognize the implicit stance the SMP takes by defending alter-


User Prompt
GPT-4's Response

Tweet: If these schools encourage students to get vaccine and discourage them to go deeper into the traditional oriental medicine/alternative medicine/naturopath/ by calling it as unscientific folk medicine or occult/New Age in full of prejudice, I will completely lose trust.
Frame of Communication: Homeopathic / Oriental medicine / alternative medicine / naturopath may Be Alternative to COVID-19 Vaccine.
Stance: Accept

The tweet in question expresses a clear concern about the framing of traditional oriental medicine, alternative medicine, and naturopathy as "unscientific folk medicine or occult/New Age" with "full of prejudice." This concern is positioned against the encouragement of students to get vaccinated, which implies that the author sees value in these traditional and alternative practices.

The stance towards the provided vaccine hesitancy frame, which suggests that "Homeopathic / Oriental medicine / alternative medicine / naturopath may be an alternative to COVID-19 Vaccine," is one of acceptance. Here's a step-by-step breakdown of the reasoning:

1. **Identification of Alternatives to Vaccines**: The tweet acknowledges the existence of traditional oriental medicine, alternative medicine, and naturopathy as potential alternatives to vaccines. This already aligns with the frame that presents these methods as possible substitutes for the COVID-19 vaccine.

2. **Defense of Alternative Medicine**: The tweet defends these alternative methods against the critique of being "unscientific" or "occult/New Age." This defense implies a belief in the legitimacy or efficacy of these methods, which is congruent with the frame that they could be viable alternatives to vaccination.

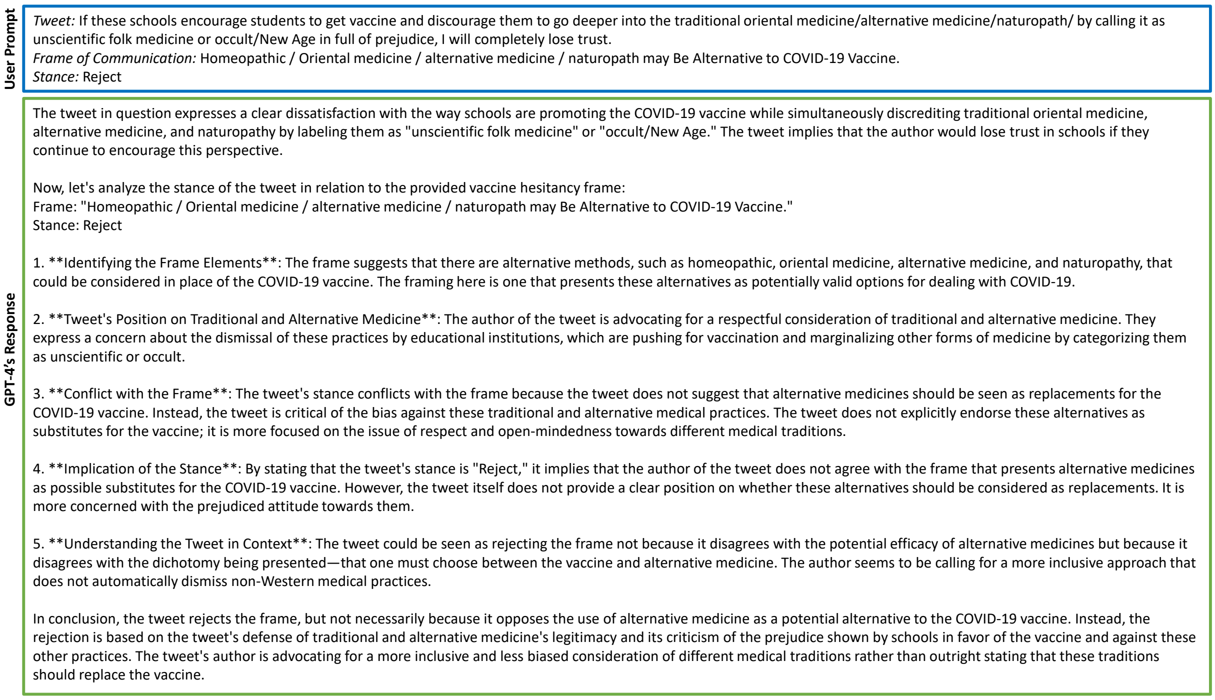
3. **Concern Over Prejudice**: The author of the tweet is concerned about the prejudice against these alternative methods. This suggests that they believe these methods are not being given fair consideration, which again aligns with the frame that proposes these methods as alternatives.

4. **Trust in Institutions**: The tweet mentions that the author will "completely lose trust" if schools continue to discourage alternative methods and promote vaccines in a way that the author perceives as prejudiced. This indicates that the author's trust is contingent on the acceptance and fair treatment of alternative medicine, further supporting the frame that these could be alternatives to vaccines.

5. **Implicit Critique of Mainstream Medicine**: By opposing the characterization of alternative medicine as "unscientific," the tweet implicitly critiques the mainstream medical establishment's approach to health, which often prioritizes vaccines and other pharmaceutical interventions. This critique is a common theme in communications that are supportive of the frame suggesting alternative treatments to vaccines.

In summary, the tweet accepts the vaccine hesitancy frame by acknowledging and defending alternative medicine as a legitimate option, expressing concern over the prejudice against these practices, and suggesting that a failure to recognize these alternatives could result in a loss of trust in the institutions promoting vaccines. The tweet does not outright reject vaccines but advocates for the inclusion and fair consideration of alternative practices, which is in line with the frame that presents these practices as possible alternatives to the COVID-19 vaccine.

Figure 7: Example of a counterfactual rationale generated with Chain-of-Explanation prompting with GPT-4 for the Accept stance from COVAXFRAMES.


User Prompt
GPT-4's Response

Tweet: If these schools encourage students to get vaccine and discourage them to go deeper into the traditional oriental medicine/alternative medicine/naturopath/ by calling it as unscientific folk medicine or occult/New Age in full of prejudice, I will completely lose trust.
Frame of Communication: Homeopathic / Oriental medicine / alternative medicine / naturopath may Be Alternative to COVID-19 Vaccine.
Stance: Reject

The tweet in question expresses a clear dissatisfaction with the way schools are promoting the COVID-19 vaccine while simultaneously discrediting traditional oriental medicine, alternative medicine, and naturopathy by labeling them as "unscientific folk medicine" or "occult/New Age." The tweet implies that the author would lose trust in schools if they continue to encourage this perspective.

Now, let's analyze the stance of the tweet in relation to the provided vaccine hesitancy frame:
Frame: "Homeopathic / Oriental medicine / alternative medicine / naturopath may Be Alternative to COVID-19 Vaccine."
Stance: Reject

1. **Identifying the Frame Elements**: The frame suggests that there are alternative methods, such as homeopathic, oriental medicine, alternative medicine, and naturopathy, that could be considered in place of the COVID-19 vaccine. The framing here is one that presents these alternatives as potentially valid options for dealing with COVID-19.

2. **Tweet's Position on Traditional and Alternative Medicine**: The author of the tweet is advocating for a respectful consideration of traditional and alternative medicine. They express a concern about the dismissal of these practices by educational institutions, which are pushing for vaccination and marginalizing other forms of medicine by categorizing them as unscientific or occult.

3. **Conflict with the Frame**: The tweet's stance conflicts with the frame because the tweet does not suggest that alternative medicines should be seen as replacements for the COVID-19 vaccine. Instead, the tweet is critical of the bias against these traditional and alternative medical practices. The tweet does not explicitly endorse these alternatives as substitutes for the vaccine; it is more focused on the issue of respect and open-mindedness towards different medical traditions.

4. **Implication of the Stance**: By stating that the tweet's stance is "Reject," it implies that the author of the tweet does not agree with the frame that presents alternative medicines as possible substitutes for the COVID-19 vaccine. However, the tweet itself does not provide a clear position on whether these alternatives should be considered as replacements. It is more concerned with the prejudiced attitude towards them.

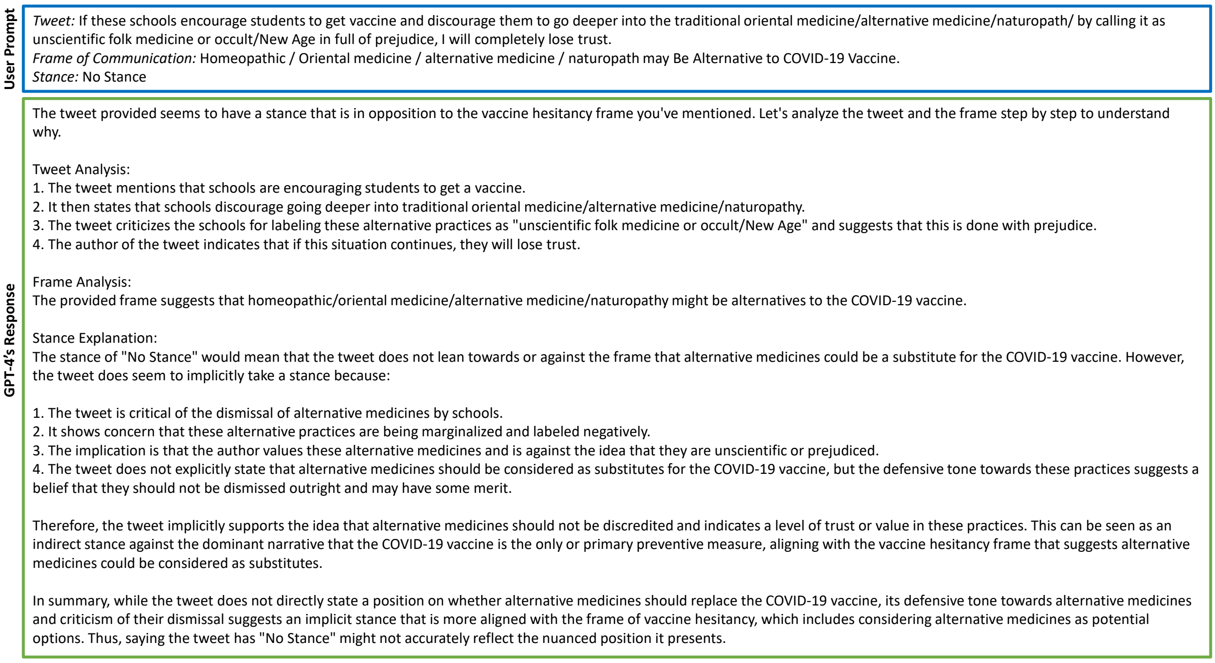
5. **Understanding the Tweet in Context**: The tweet could be seen as rejecting the frame not because it disagrees with the potential efficacy of alternative medicines but because it disagrees with the dichotomy being presented—that one must choose between the vaccine and alternative medicine. The author seems to be calling for a more inclusive approach that does not automatically dismiss non-Western medical practices.

In conclusion, the tweet rejects the frame, but not necessarily because it opposes the use of alternative medicine as a potential alternative to the COVID-19 vaccine. Instead, the rejection is based on the tweet's defense of traditional and alternative medicine's legitimacy and its criticism of the prejudice shown by schools in favor of the vaccine and against these other practices. The tweet's author is advocating for a more inclusive and less biased consideration of different medical traditions rather than outright stating that these traditions should replace the vaccine.

Figure 8: Example of a counterfactual rationale generated with Chain-of-Explanation prompting with GPT-4 for the Reject stance from COVAXFRAMES.

native medicines." Ultimately, GPT-4 judges correctly that the stance of this SMP is *Accept* towards the FoC under consideration, aligning again with the human judgments present in COVAXFRAMES.

Figure 11 illustrates a multimodal SMP from the test set of MMVAX-STANCE. The SMP is short, stating in text only "The covid19 vaccine story" and providing a meme of Jim Halpert, a character from The Office. In this meme, Jim points to a flip chart, and in the first section, the text on the flip chart says, "The COVID vaccine should be tested on politicians and bureaucrats." In the second section, where Jim is smiling, the flip chart says, "If they survive, the vaccine is safe. If they don't


User Prompt
GPT-4's Response

Tweet: If these schools encourage students to get vaccine and discourage them to go deeper into the traditional oriental medicine/alternative medicine/naturopath/ by calling it as unscientific folk medicine or occult/New Age in full of prejudice, I will completely lose trust.
Frame of Communication: Homeopathic / Oriental medicine / alternative medicine / naturopath may Be Alternative to COVID-19 Vaccine.
Stance: No Stance

The tweet provided seems to have a stance that is in opposition to the vaccine hesitancy frame you've mentioned. Let's analyze the tweet and the frame step by step to understand why.

Tweet Analysis:
1. The tweet mentions that schools are encouraging students to get a vaccine.
2. It then states that schools discourage going deeper into traditional oriental medicine/alternative medicine/naturopathy.
3. The tweet criticizes the schools for labeling these alternative practices as "unscientific folk medicine or occult/New Age" and suggests that this is done with prejudice.
4. The author of the tweet indicates that if this situation continues, they will lose trust.

Frame Analysis:
The provided frame suggests that homeopathic/oriental medicine/alternative medicine/naturopathy might be alternatives to the COVID-19 vaccine.

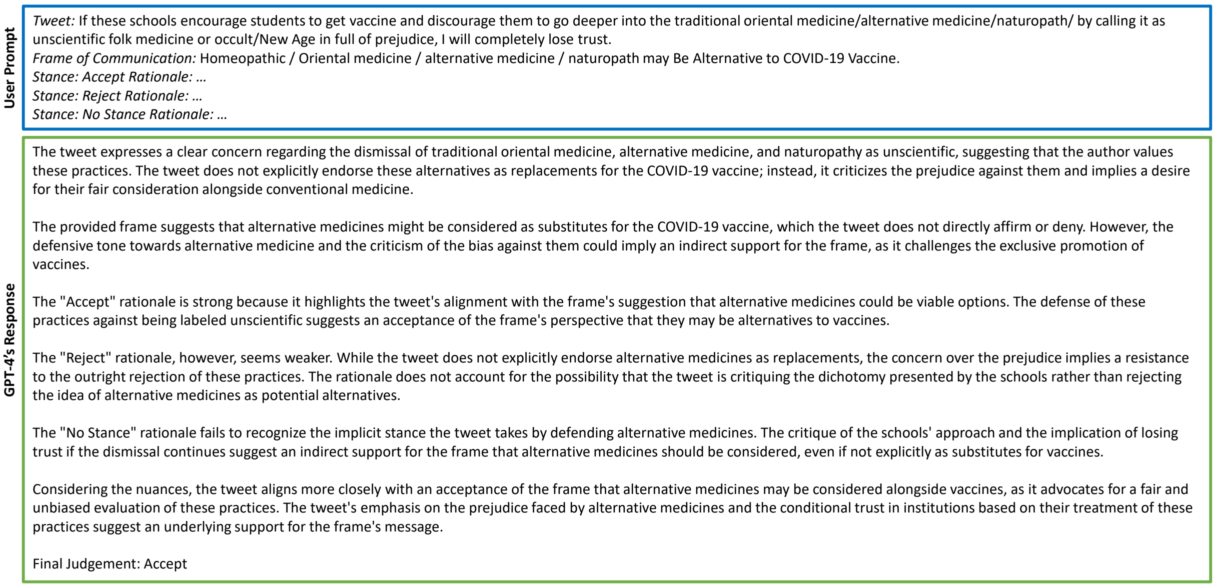
Stance Explanation:
The stance of "No Stance" would mean that the tweet does not lean towards or against the frame that alternative medicines could be a substitute for the COVID-19 vaccine. However, the tweet does seem to implicitly take a stance because:

1. The tweet is critical of the dismissal of alternative medicines by schools.
2. It shows concern that these alternative practices are being marginalized and labeled negatively.
3. The implication is that the author values these alternative medicines and is against the idea that they are unscientific or prejudiced.
4. The tweet does not explicitly state that alternative medicines should be considered as substitutes for the COVID-19 vaccine, but the defensive tone towards these practices suggests a belief that they should not be dismissed outright and may have some merit.

Therefore, the tweet implicitly supports the idea that alternative medicines should not be discredited and indicates a level of trust or value in these practices. This can be seen as an indirect stance against the dominant narrative that the COVID-19 vaccine is the only or primary preventive measure, aligning with the vaccine hesitancy frame that suggests alternative medicines could be considered as substitutes.

In summary, while the tweet does not directly state a position on whether alternative medicines should replace the COVID-19 vaccine, its defensive tone towards alternative medicines and criticism of their dismissal suggests an implicit stance that is more aligned with the frame of vaccine hesitancy, which includes considering alternative medicines as potential options. Thus, saying the tweet has "No Stance" might not accurately reflect the nuanced position it presents.

Figure 9: Example of a counterfactual rationale generated with Chain-of-Explanation prompting with GPT-4 for No Stance from CoVaxFrames.


User Prompt
GPT-4's Response

Tweet: If these schools encourage students to get vaccine and discourage them to go deeper into the traditional oriental medicine/alternative medicine/naturopath/ by calling it as unscientific folk medicine or occult/New Age in full of prejudice, I will completely lose trust.
Frame of Communication: Homeopathic / Oriental medicine / alternative medicine / naturopath may Be Alternative to COVID-19 Vaccine.
Stance: Accept Rationale: …
Stance: Reject Rationale: …
Stance: No Stance Rationale: …

The tweet expresses a clear concern regarding the dismissal of traditional oriental medicine, alternative medicine, and naturopathy as unscientific, suggesting that the author values these practices. The tweet does not explicitly endorse these alternatives as replacements for the COVID-19 vaccine; instead, it criticizes the prejudice against them and implies a desire for their fair consideration alongside conventional medicine.

The provided frame suggests that alternative medicines might be considered as substitutes for the COVID-19 vaccine, which the tweet does not directly affirm or deny. However, the defensive tone towards alternative medicine and the criticism of the bias against them could imply an indirect support for the frame, as it challenges the exclusive promotion of vaccines.

The "Accept" rationale is strong because it highlights the tweet's alignment with the frame's suggestion that alternative medicines could be viable options. The defense of these practices against being labeled unscientific suggests an acceptance of the frame's perspective that they may be alternatives to vaccines.

The "Reject" rationale, however, seems weaker. While the tweet does not explicitly endorse alternative medicines as replacements, the concern over the prejudice implies a resistance to the outright rejection of these practices. The rationale does not account for the possibility that the tweet is critiquing the dichotomy presented by the schools rather than rejecting the idea of alternative medicines as potential alternatives.

The "No Stance" rationale fails to recognize the implicit stance the tweet takes by defending alternative medicines. The critique of the schools' approach and the implication of losing trust if the dismissal continues suggest an indirect support for the frame that alternative medicines should be considered, even if not explicitly as substitutes for vaccines.

Considering the nuances, the tweet aligns more closely with an acceptance of the frame that alternative medicines may be considered alongside vaccines, as it advocates for a fair and unbiased evaluation of these practices. The tweet's emphasis on the prejudice faced by alternative medicines and the conditional trust in institutions based on their treatment of these practices suggest an underlying support for the frame's message.

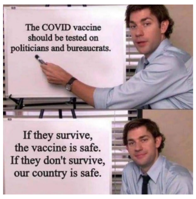
Final Judgement: Accept

Figure 10: Example of a verification rationale generated with Counterfactual Chain-of-Contrastive Verification prompting with GPT-4 from CoVaxFrames.

survive, our country is safe." The FoC under consideration as the object of stance is "The Federal government lied about the COVID-19 vaccines to reduce the information about effective treatments." Figure 11 also includes the response generated by GPT-4V with CoE prompting with a counterfactual stance value of *Accept*. GPT-4V describes the content of the meme, and continues to provide a 3-point analysis of how the meme and the text contribute to the counterfactual stance value of *Accept*, recognizing that "...the SMP indirectly supports the frame..." However, Figure 12 illustrates how a counterfactual stance value of *Reject* can shift the perspective on this SMP. GPT-4V identifies that the SMP is satirical and humorous in nature, concluding that the SMP "...does not directly support the idea that the government lied about the vaccines..." A consideration of the *No Stance* counterfactual is provided in Figure 13, yielding a similar rationale to that of the *Reject* counterfactual: that "...the

User Prompt

Tweet: The covid19 vaccine story
Frame of Communication: The Federal government lied about the COVID-19 vaccines to reduce the information about effective treatments.
Stance: Accept


The COVID vaccine should be tested on politicians and bureaucrats.
If they survive, the vaccine is safe. If they don't survive, our country is safe.

GPT-4V's Response

The image accompanying the tweet shows a man presenting two statements on a flip chart. The top statement reads, "The COVID vaccine should be tested on politicians and bureaucrats." The bottom statement reads, "If they survive, the vaccine is safe. If they don't survive, our country is safe."

The image and the text it contains contribute to the stance in a few ways:

1. Suggesting that politicians and bureaucrats should be the test subjects for the COVID-19 vaccine implies a lack of trust in the safety or efficacy of the vaccine. It insinuates that the vaccine might not be safe and thus should be tested on those who are deemed expendable or less trustworthy by the author of the tweet.

2. The conditional statements about survival indicate a cynical view of the political and bureaucratic class, implying that their well-being is not the primary concern and that there might be a positive outcome for the country regardless of the vaccine's safety.

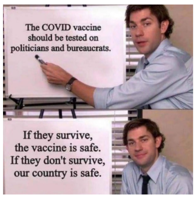
3. The overall tone of the image is sarcastic and critical, which aligns with the provided vaccine hesitancy frame that suggests the Federal government might have lied about the vaccines. The image seems to question the honesty of the information disseminated by the government and those in power.

By using humor and sarcasm, the tweet indirectly supports the frame that there is a reason to doubt the information provided by the government about COVID-19 vaccines, which could include lying about the vaccines to downplay other effective treatments. The acceptance stance is therefore reinforced by the image's insinuation that there is a lack of transparency and trustworthiness among politicians and bureaucrats regarding the vaccine.

Figure 11: Example of a counterfactual rationale generated with Chain-of-Explanation prompting with GPT-4V for the Accept stance from MMVAX-STANCE.

User Prompt

Tweet: The covid19 vaccine story
Frame of Communication: The Federal government lied about the COVID-19 vaccines to reduce the information about effective treatments.
Stance: Reject


The COVID vaccine should be tested on politicians and bureaucrats.
If they survive, the vaccine is safe. If they don't survive, our country is safe.

GPT-4V's Response

The image accompanying the tweet shows a man presenting two statements on a flip chart. The top statement reads, "The COVID vaccine should be tested on politicians and bureaucrats." The bottom statement reads, "If they survive, the vaccine is safe. If they don't survive, our country is safe."

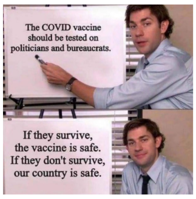
The stance of the tweet is rejecting the frame that the Federal government lied about the COVID-19 vaccines to reduce information about effective treatments. The image and text together convey a satirical message that suggests distrust in politicians and bureaucrats rather than addressing the effectiveness of the vaccine or the veracity of government statements about it. The humor in the image implies that the safety of the vaccine could be tested on those who are often distrusted by the public (politicians and bureaucrats), rather than seriously engaging with the notion of vaccine safety or government honesty.

This stance can be seen as rejecting the provided frame because it does not directly support the idea that the government lied about the vaccines; instead, it seems to be poking fun at the situation by suggesting an extreme and unethical method of testing vaccine safety. The underlying tone of the image may imply skepticism towards authority figures, but it does not provide any evidence or arguments that the government actively misled the public about COVID-19 treatments. Therefore, the tweet and the image, while expressing distrust in government officials, do not align with the specific claim that the government lied about the vaccines to suppress information about effective treatments.

Figure 12: Example of a counterfactual rationale generated with Chain-of-Explanation prompting with GPT-4V for the Reject stance from MMVAX-STANCE.

User Prompt

Tweet: The covid19 vaccine story
Frame of Communication: The Federal government lied about the COVID-19 vaccines to reduce the information about effective treatments.
Stance: No Stance


The COVID vaccine should be tested on politicians and bureaucrats.
If they survive, the vaccine is safe. If they don't survive, our country is safe.

GPT-4V's Response

The image shows a man pointing at a flip chart with two statements written on it. The top panel reads, "The COVID vaccine should be tested on politicians and bureaucrats." The bottom panel reads, "If they survive, the vaccine is safe. If they don't survive, our country is safe." This image is a meme format that is often used to express a point through humor or sarcasm.

In relation to the provided vaccine hesitancy frame – that the Federal government lied about the COVID-19 vaccines to reduce information about effective treatments – the meme in the image does not directly address this frame. The meme does not make any claims about the government lying or withholding information about effective treatments. Instead, it uses a hypothetical scenario about testing vaccines on politicians and bureaucrats to make a satirical point, which could be interpreted as a commentary on distrust in political leaders or the safety of the vaccine, depending on the viewer's perspective.

The stance of the tweet as "No Stance" regarding the provided frame is accurate because the meme does not engage with the specific claim that the government lied about the vaccines to suppress information about effective treatments. The meme does suggest skepticism about the vaccine or the motives of politicians, but it does not provide a clear stance on the specific issue of government dishonesty about vaccine efficacy or alternative treatments. Therefore, it does not affirm or deny the frame, but rather exists independently as a piece of satirical commentary.

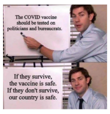
Figure 13: Example of a counterfactual rationale generated with Chain-of-Explanation prompting with GPT-4V for No Stance from MMVAX-STANCE.

meme does not engage with the specific claim that

the government lied about the vaccines...” and con-

User Prompt

Tweet: The covid19 vaccine story
Frame of Communication: The Federal government lied about the COVID-19 vaccines to reduce the information about effective treatments.
Stance: Accept Rationale: …
Stance: Reject Rationale: …
Stance: No Stance Rationale: …


The COVID vaccine should be tested on politicians and bureaucrats.
If they survive, the vaccine is safe. If they don't survive, our country is safe.

GPT-4V's Response

The image presented is a meme featuring a man pointing to a flip chart with two panels. The top panel reads, "The COVID vaccine should be tested on politicians and bureaucrats." The bottom panel reads, "If they survive, the vaccine is safe. If they don't survive, our country is safe." This meme uses humor and sarcasm to make a statement about the COVID-19 vaccine and politicians.

The "Accept" rationale posits that the image suggests a lack of trust in the safety or efficacy of the vaccine and cynicism towards politicians and bureaucrats. It interprets the sarcasm as aligning with the provided frame that the Federal government might have lied about the vaccines. The strengths of this rationale lie in recognizing the distrustful tone towards authority figures and the implication that the vaccine's safety is questionable. However, it may be overreaching to conclude that the meme supports the idea that the government has lied about other effective treatments based solely on this image.

The "Reject" rationale contends that the tweet rejects the frame because it does not directly support the idea that the government lied about the vaccines. It argues that the meme is a satirical commentary on the distrust of politicians and bureaucrats, not a commentary on vaccine safety or government honesty. The strength of this rationale is that it acknowledges the meme's lack of direct engagement with the specific claim of government dishonesty about the vaccines. However, the weakness is that it may overlook the broader implications of vaccine skepticism and distrust in government, which could be interpreted as indirectly supporting the frame.

The "No Stance" rationale suggests that the tweet does not take a stance on the provided frame because the meme does not directly address the claim of government lying or withholding information about effective treatments. It emphasizes that the meme is satirical and does not affirm or deny the frame, but serves as an independent commentary. The strength of this rationale is that it accurately recognizes that the meme does not make any explicit claims about government misinformation. Its weakness might be in not acknowledging that the meme could still resonate with sentiments of vaccine hesitancy and skepticism towards government authority, which might be seen as indirectly related to the frame.

Considering the content of the meme and the nature of the frame, the meme's satirical take on vaccine testing on politicians does not directly address the claim of government dishonesty about the efficacy of vaccines or the suppression of information about other treatments. It expresses a general skepticism towards politicians and humorously suggests an extreme method of vaccine testing without making specific claims about government lies. Therefore, the stance of the tweet towards the provided vaccine hesitancy frame is best described as:

No Stance.

Figure 14: Example of a verification rationale generated with Counterfactual Chain-of-Contrastive Verification prompting with GPT-4V from MMVAX-STANCE.

cludes that "...it does not affirm or deny the frame, but rather exists independently as a piece of satirical commentary." Figure 14 provides the response produced by GPT-4V with C-CoCV prompting on the same SMP, with all counterfactual rationales included. GPT-4V compares and contrasts each of the counterfactual hypotheses and evidence present, and concludes that "...the meme's satirical take on vaccine testing on politicians does not directly address the claim of government dishonesty about the efficacy of vaccines..." and that "...the stance of the SMP towards the provided vaccine hesitancy frame is best described as: No Stance." Again, ToC prompting yields an aligned response to the human judgments present in MMVAX-STANCE, as this post is annotated as having *No Stance* towards the FoC under consideration.