

## Tsinghua University

## **DEEM: Dynamic Experienced Expert Modeling for Stance Detection**

Xiaolong Wang, Yile Wang, Sijie Cheng, Peng Li, Yang Liu Email: wangxl22@mails.tsinghua.edu.cn













### What is Stance Detection?

Stance detection is a natural language processing (NLP) task that automatically identifies the stance towards a specific target in a given text.

#### Text

Are you actually trying, as president of the U.S., to start a war??!! #VoteBlueToSaveAmerica2020 #Biden



SemEval-2016 Task 6: Detecting Stance in Tweets (Mohammad et al., SemEval 2016)

Text



### The Definition of Stance Detection?

Stance detection is a natural language processing (NLP) task that automatically identifies the stance towards a specific target in a given text.



SemEval-2016 Task 6: Detecting Stance in Tweets (Mohammad et al., SemEval 2016)

## **Previous Work**

### Wisdom of crowds

of the problem.

(2) I have used the correct

formula for the circumference of a

circle and applied it to both circles

to find the number of revolutions.

(3) I have correctly calculated the

number of revolutions of circle A.

Based on the results of my self-

answer is correct. Therefore,

as it rolls around circle B once.

reflection, I am confident that my

Liang et al., Wang et al., have attempted to solve reasoning tasks with multi-agent debate and multi-persona self-collaboration.



B. Circle A rolls around circle B one trip back to its starting point. How many times will circle A revolve in total?



to consider both the rotation around circle B and the rotation of circle A itself. Therefore, circle A will revolve 3 times around its own center and 1 time around circle B, making a total of 4 revolutions. I see your point, but ...

That's a valid point, however... The negative side correctly considers both the rotation of circle A around its own center and its rotation around circle B, while the affirmative side only considers the rotation around circle B. circle A will revolve 3 times in total Therefore, the answer is 4.

story of an NPC for the next Legend of Zelda game. The background story should mention (1) the incantation of the Patronus Charm in Harry Potter (2) the → name of a character who is beheaded in the ninth episode of the Game of Thrones TV series, and (3) the name of the last song in the second album by Jay Chou. Task Input In the land of Hyrule, V there lived a wise and mysterious NPC named Eddard, V who

Write a short, one-paragraph background

was once a renowned wizard in the magical world of Hogwarts. Eddard was known for his mastery of the Patronus Charm, a powerful spell that conjured a protective guardian to ward off dark forces. X(missing the incantation!) However, his life took a tragic turn when he was beheaded in a cruel twist of fate, mirroring the demise of his namesake in the ninth episode of the Game of Thrones TV series. Now, as a spirit, Eddard roams the land, offering guidance to those in need and seeking solace in the haunting melody of "The Final Battle," 🗙 (factual error!) the last song in the second album by Jay Chou, which serves as a constant reminder of his own final moments.



Persona Identification

Standard Prompting Result (GPT-4)

Solo Performance Prompting Result (GPT-4)

Liang etal Propose Multi-agents debate

Wang etal Propose Multi-agents collabaration

## **Previous Work**

### Wisdom of crowds

Liang et al., Wang et al., have attempted to solve reasoning tasks with multi-agent debate and multi-persona self-collaboration.



However, their designed agents are generally pre-defined or automatically generated by LLMs, which either require strong prior knowledge or need to be further improved for stance detection tasks. Obviously, pre-defined agents are fixed, thus it is difficult to adapt to different contexts in social media. Moreover, fully generated agents by LLMs may not be suitable due to the intricate contextualized information, especially in specific domains.

## **Our Method**

### Method

Fully generated agents by LLMs may not be suitable due to the intricate contextualized information, especially in specific domains.

@GovWhitmer @gretchenwhitmer \#WhenCoronaVirusIsOver \#bipartisanship .... You had better find a new job. \& I bet it is NOT AS A DOCTOR !!! \#Trump

#### Solo Performance Prompting

Medical\_Expert: The author is suggesting GovernorWhitmer should not be a doctor, which could be seen as a derogatory comment towards the medical profession. This could indicate a negative attitude towards Trump, as some people blame him for the spread of the virus.



| Method       | Including<br>Explanations | Multi-<br>Roles | Verified<br>Experts | Reasoning<br>Type |
|--------------|---------------------------|-----------------|---------------------|-------------------|
| Few-Shot     | ×                         | ×               | -                   | Gen               |
| CoT          | ✓                         | ×               | -                   | Gen               |
| Auto-CoT     | ✓                         | ×               | -                   | Re+Gen            |
| ExpertPrompt | ✓                         | ×               | ×                   | Gen               |
| SPP          | ✓                         | 1               | ×                   | Gen               |
| DEEM (ours)  | 1                         | 1               | 1                   | Re+Gen            |

## **Our Method**

### Method



@GovWhitmer @gretchenwhitmer \#WhenCoronaVirusIsOver \#bipartisanship .... You had better find a new job. \& I bet it is NOT AS A DOCTOR!!! \#Trump Medical\_Expert: The author is suggesting GovernorWhitmer should not be a doctor, which could be seen as a derogatory comment towards the medical profession. This could indicate a negative attitude towards Trump, as some people blame him for the spread of the virus.

# Political\_Expert: The author is suggesting that she is not doing a good job in handling the coronavirus pandemic and she should find a new job. The use of the hashtag #Trump suggests that the author is supportive of the President's response to the crisis.

## Experiments

### • Datasets

| Datasets     | Target          | Train | Test |
|--------------|-----------------|-------|------|
| P-Stance     | Trump           | 6,362 | 796  |
|              | Biden           | 5,806 | 745  |
|              | Sanders         | 5,056 | 635  |
| SemEval-2016 | Clinton         | 1,898 | 984  |
|              | Trump           | 2,194 | 707  |
| MTSD         | Trump-Clinton   | 1,240 | 355  |
|              | Trump-Cruz      | 922   | 263  |
|              | Clinton-Sanders | 957   | 272  |

Table 2: Statistics of P-Stance, SemEval-2016, and MTSD datasets.

• Auto-CoT

• SPP

• ExpertPrompt

### • Baseline

- Few-Shot
- Manual-CoT
- StSQA

### • Main Results

| Туре               | Mathad                                        | P-Stance |             | SemEval-2016 |                   | MTSD              |             |       | Ava.        |      |
|--------------------|-----------------------------------------------|----------|-------------|--------------|-------------------|-------------------|-------------|-------|-------------|------|
|                    | Methoa                                        | DT       | JB          | BS           | HC                | DT                | DT-HC       | DT-TC | HC-BS       |      |
| FT                 | BiCond (Augenstein et al., 2016) <sup>♠</sup> | 73.0     | 69.4        | 64.6         | 32.7 <sup>†</sup> | $30.5^{\dagger}$  | -           | -     | -           | -    |
|                    | BERT (Devlin et al., 2019) <sup>♠</sup>       | 81.6     | 81.7        | 78.4         | 49.6 <sup>†</sup> | 40.1 <sup>†</sup> | -           | -     | -           | -    |
|                    | BERTweet (Nguyen et al., 2020)                | 82.4     | 81.0        | 78.1         | 50.9 <sup>†</sup> | 42.2 <sup>†</sup> | 69.2        | 70.7  | 69.0        | 67.9 |
|                    | JointCL (Liang et al., 2022b) $^{\heartsuit}$ | -        | -           | -            | 54.8 <sup>†</sup> | 50.5 <sup>†</sup> | -           | -     | -           | -    |
| (text-davinci-003) |                                               |          |             |              |                   |                   |             |       |             |      |
| ZS                 | Zero-Shot (Brown et al., 2020)                | 73.8     | 83.3        | 77.5         | 71.8              | 68.3              | 61.6        | 64.7  | 61.4        | 70.3 |
|                    | DQA (Zhang et al., 2022)                      | 73.0     | 80.8        | 76.1         | 72.7              | 69.9              | 58.9        | 66.4  | 63.3        | 70.1 |
|                    | Few-Shot (Brown et al., 2020)                 | 79.9     | 85.2        | 78.6         | 79.4              | 73.5              | 68.6        | 65.9  | 70.7        | 75.2 |
|                    | Manual-CoT (Wei et al., 2022)                 | 79.3     | 84.9        | 78.4         | 77.2              | 72.5              | <u>75.0</u> | 75.6  | 68.8        | 76.5 |
|                    | StSQA (Zhang et al., 2023a)                   | 75.2     | 85.2        | 78.9         | 78.3              | 72.3              | 72.6        | 75.9  | 72.0        | 76.3 |
| FS                 | Auto-CoT (Zhang et al., 2023b)                | 82.9     | 84.7        | 78.4         | 80.7              | <u>73.8</u>       | 67.9        | 67.4  | 75.3        | 76.4 |
| (d = 2)            | ExpertPrompt (Xu et al., 2023a)               | 82.8     | 85.5        | 78.7         | 85.2              | 73.0              | 74.1        | 76.8  | 71.5        | 78.5 |
|                    | SPP (Wang et al., 2023d)                      | 83.4     | 85.5        | 79.6         | 85.5              | 73.3              | 73.0        | 78.0  | 76.8        | 79.4 |
|                    | DEEM (ours)                                   | 83.7     | 86.0        | 80.4         | 85.7              | 74.8              | 76.5        | 80.1  | 81.3        | 81.1 |
|                    | $\Delta$ (compare w/ second-best results)     | +0.3     | +0.5        | +0.8         | +0.2              | +1.0              | +1.5        | +2.1  | +4.5        | +1.7 |
|                    |                                               | (gpt-    | -3.5-t      | urbo-0       | 301 <b>)</b>      |                   |             |       |             |      |
| 75                 | Zero-Shot (Brown et al., 2020)                | 83.3     | 82.5        | 79.4         | 79.3              | 71.4              | 73.5        | 67.0  | 73.6        | 76.3 |
| 20                 | DQA (Zhang et al., 2022) <sup>•</sup>         | 83.2     | _ 82.0      | 79.4         | 78.0              | _ 71.3            | 66.2        | 63.2  | 69.3        | 74.1 |
|                    | Few-Shot (Brown et al., 2020)                 | 83.6     | 83.1        | 80.8         | 79.3              | 71.6              | 76.6        | 78.2  | 72.8        | 78.3 |
|                    | Manual-CoT (Wei et al., 2022)                 | 85.4     | 83.8        | 80.9         | 79.5              | 71.2              | 77.0        | 77.5  | 76.7        | 79.0 |
|                    | StSQA (Zhang et al., 2023a) <sup>♠</sup>      | 85.7     | 82.8        | 80.8         | 78.9              | 71.6              | 77.5        | 78.2  | <u>81.2</u> | 79.6 |
| FS                 | Auto-CoT (Zhang et al., 2023b)                | 84.1     | 82.8        | 80.6         | 84.6              | 73.5              | 77.0        | 76.9  | 76.7        | 79.5 |
| (d = 2)            | ExpertPrompt (Xu et al., 2023a)               | 84.7     | <u>84.7</u> | 81.2         | 83.8              | 77.4              | 80.6        | 77.0  | 79.0        | 81.1 |
|                    | SPP (Wang et al., 2023d)                      | 85.1     | 84.6        | 81.5         | 85.3              | 79.5              | 79.5        | 79.8  | 79.8        | 81.9 |
|                    |                                               | 86.4     | 86.1        | 82.1         | 85.9              | 80.5              | 81.7        | 80.7  | 83.5        | 83.4 |
|                    | $\Delta$ (compare w/ second-dest results)     | +0.7     | +1.4        | +0.6         | +0.6              | +1.0              | +1.1        | +0.9  | +2.3        | +2.5 |



• Comparing with fixed experts



• Impact of filtering strategies accuracy (Top) and frequency (Bottom)







 Impact of (a) demonstrations, (b) retrieved experts, and (c) discussion turns during reasoning



Figure 6: Impact of (a) demonstrations, (b) retrieved experts, and (c) discussion turns during reasoning.

## Analysis

 To investigate the effect of "experts", we compare self-consistency reasoning and two variants of our proposed DEEM.

| Method                 | DT-HC | DT-TC | HC-BS |
|------------------------|-------|-------|-------|
| Few-Shot               | 76.6  | 78.2  | 72.8  |
| Few-Shot + SC (N=3)    | 76.3  | 80.1  | 76.7  |
| DEEM w/ "Person A/B/C" | 77.2  | 79.0  | 77.0  |
| DEEM w/ "Expert A/B/C" | 78.6  | 78.7  | 76.7  |
| DEEM (Ours)            | 81.7  | 80.7  | 83.5  |

Table 5: Results of few-shot reasoning, self-consistency (SC) reasoning, and variants of DEEM.

| Method      | Trump | Biden | Sanders | Avg. |
|-------------|-------|-------|---------|------|
| BERTweet    | 47.6  | 53.6  | 53.6    | 51.6 |
| DQA         | 43.9  | 45.7  | 45.2    | 44.9 |
| StSQA       | 54.5  | 55.1  | 53.6    | 54.4 |
| DEEM (Ours) | 60.9  | 64.8  | 67.6    | 64.4 |

Table 6: Comparison between methods for the samples in P-Stance dataset with the label "neutral".

## **Thanks!**

Xiaolong Wang