Supplementary Materials for

Zero-Shot Stance Detection in Practice: Insights on Training, Prompting, and Decoding with a Capable Lightweight LLM

Rachith Aiyappa,[1*] Shruthi Senthilmani[1], Jisun An[1], Haewoon Kwak[1], and Yong-Yeol Ahn[1]

[1]Center for Complex Networks and Systems
Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington
Bloomington, Indiana, USA, 47408.

[*]Corresponding author: racball@iu.edu

1 Domain Corpus of SemEval 2016 Task 6

Target	AT	CC	DT	FM	HC	LA
# Tweets	935,181	208,880	78,156	144,166	238,193	113,193

Table S1: Number of tweets for each target in the domain corpus of SemEval 2016 Task 6.

The tweets in the domain corpus were not labeled for stance. These tweets were gathered by polling Twitter for tweets with hashtags corresponding to the respective targets.

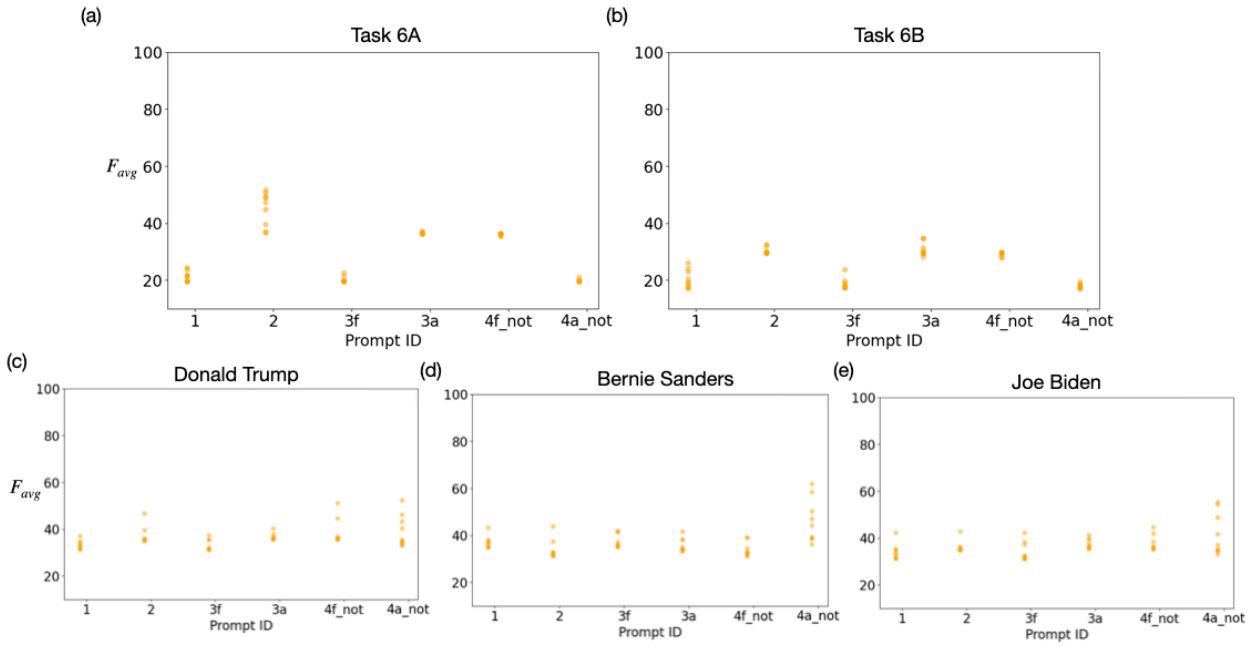

Figure S1: Experiments with Llama 3.1 in the greedy setting. Top row: SemEval 2016 dataset. Bottom row: P-Stance dataset

2 Experiments with Llama 3.1 and Gemma

We also conducted experiments with Llama 3.1-8B instruct[1] and Gemma 2-9B[2] using the same prompts and instructions used for FlanT5-XXL. Figs S1 and S2 show that these models perform poorly in a greedy decoding setting. A closer inspection revealed that this issue stems primarily from the models' poor ability to follow instructions. Despite instructing them to generate only the desired stance labels, they rarely do so. Instead, they either produce verbose responses that require additional processing to extract the stance label—if it is present at all—or omit the label entirely within the first 10 tokens of their output. For example, some outputs produced by Llama3, which would require post-processing, are:

- "\n\nThe best answer is Positive. The statement,"

[1] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[2] https://huggingface.co/google/gemma-2-9b

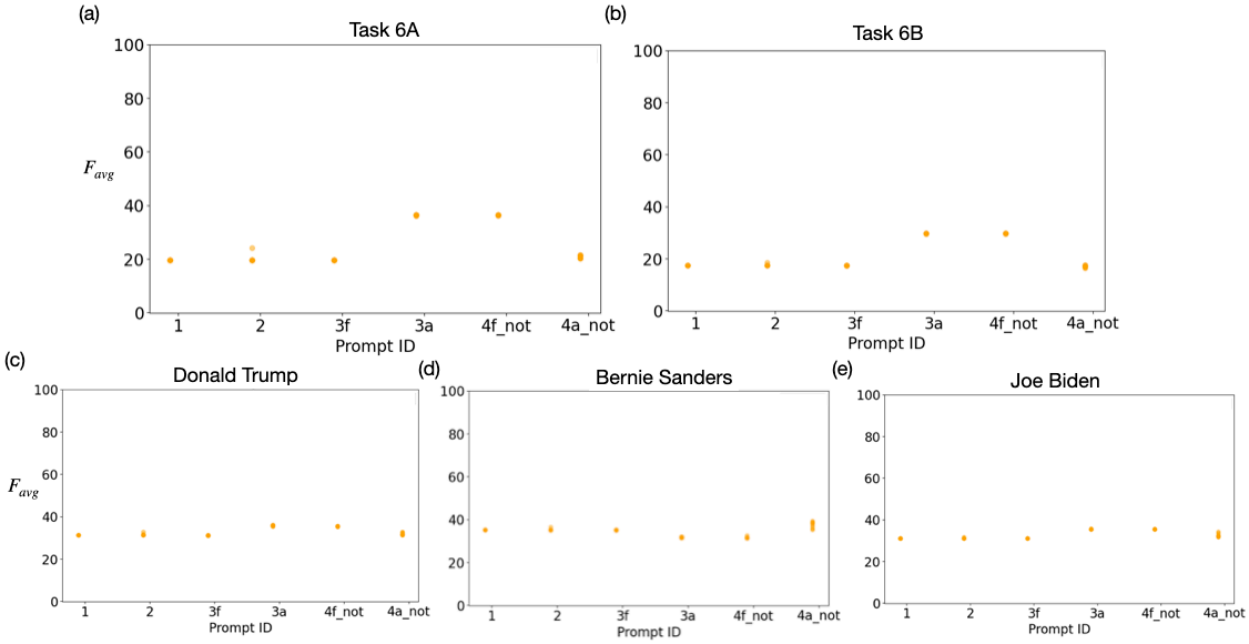

Figure S2: Experiments with Gemma 3.1 in the greedy setting. Top row: SemEval 2016 dataset. Bottom row: P-Stance dataset.

- "\nThe correct answer is Negative. \n\nExplana..."
- "\nThe Final Answer is: Negative. Please"
- "(True, False, or Neutral)\n\nAnswer:"
- "(Favor, Against, Neutral)\n\nAgainst\n"

Similarly, some of the outputs by Gemma which does not contain the stance labeling the first ten tokens, are

- "\nYour response to the question should be eith..."
- "\n\nStatement: I'm not a believer of"
- \nA. Favor\nB. Against\nC

Verbose responses also hinder the automatic identification of the token corresponding to the stance—assumed to be the token immediately following the prompt, in the main paper—making it difficult to extract the associated token probability needed for PMI and AfT calculations.

These limitations reduce the accuracy of extracting stance labels from model outputs for evaluation. Nevertheless, we assume that the model expresses the stance in the token positions immediately following the prompt and use the probabilities of stance labels at these positions— conditioned on the prompt—to evaluate performance under different decoding strategies (Figs. S1 and S2). This also revealed a slight positivity bias, like in the case of FlanT5-XXL (Figs. S3 and S4). However, we note that the log-probabilities are significantly lower than those reported for the positivity bias of FlanT5-XXL in the main paper. This discrepancy arises from our assumption that the stance label appears in the token immediately following the prompt—a pattern that holds for FlanT5-XXL, but not for Llama 3.1 and Gemma 2. We leave it to future work to develop better methods to accurately capture stance positions in output sequences.

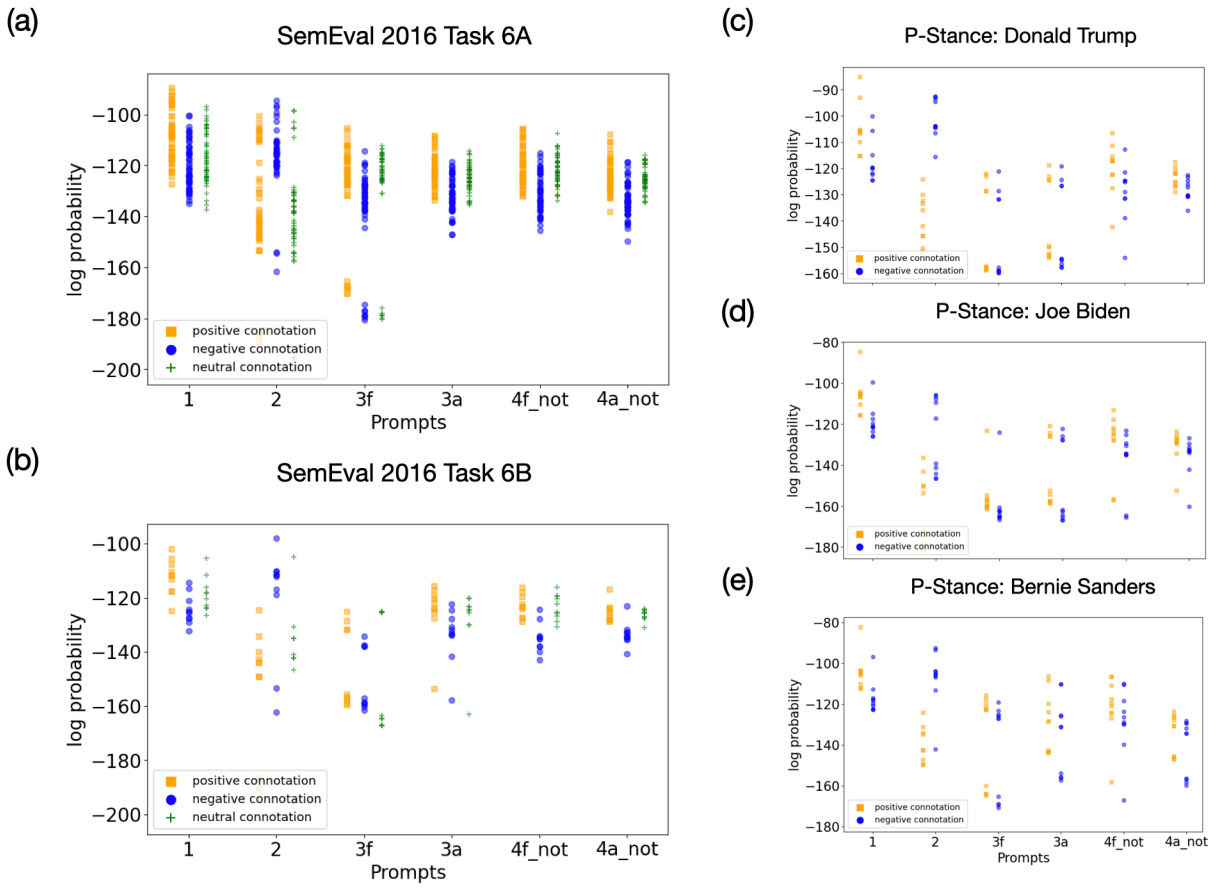

Figure S3: Llama 3.1 8b instruct may be biased towards outputting labels with positive connotations regardless of prompts and instruction. Probability of Llama 3.1 outputting a label with a positive (e.g. "favor", "true"), negative (e.g. "against", "false"), or neutral connotation across all targets in the SemEval 2016 (a) Task 6A (b) Task 6B in a context-free setting—a setting where the $\langle tweet \rangle$ item is not fed into the model during inference. Probability of outputting a positive or negative label in P-Stance for targets (c) Donald Trump (d) Joe Biden (e) Bernie Sanders. Each point represents a (prompt, instruction) pair.

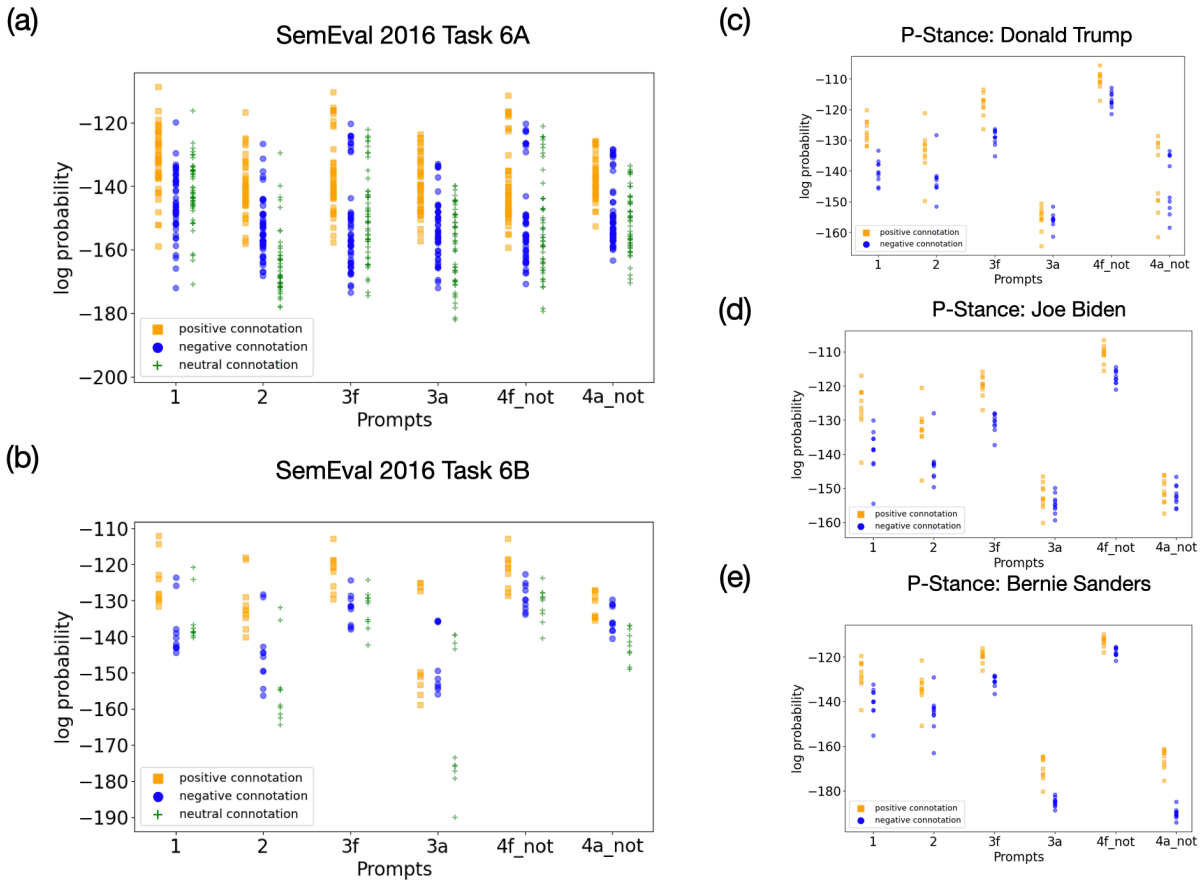

Figure S4: Gemma 2 9b may be biased towards outputting labels with positive connotations regardless of prompts and instructions. Probability of Gemma 2 outputting a label with a positive (e.g. "favor", "true"), negative (e.g. "against", "false"), or neutral connotation across all targets in the SemEval 2016 (a) Task 6A (b) Task 6B in a context-free setting—a setting where the $\langle tweet \rangle$ item is not fed into the model during inference. Probability of outputting a positive or negative label in P-Stance for targets (c) Donald Trump (d) Joe Biden (e) Bernie Sanders. Each point represents a (prompt, instruction) pair.

3 Stance datasets in Instruction Tuning Data

FlanT5 takes the T5 model and instruction tunes on a number of NLP datasets in a few-shot, zero-shot, or chain-of-thought setting using instruction tuning methods like input inversion, template generation, etc. (Longpre and Roberts, 2023). We performed a qualitative inspection of these datasets (Chung et al., 2022) and found no evidence of the SemEval 2016 Task 6 or the P-Stance dataset. However, we found that the model has been instruction-tuned on the Natural Instruction V2 dataset (Wang et al., 2022)—a collection of 1600+ NLP tasks with instructions—which contains few stance detection tasks. For example, task 513[3], with the definition "You will be given a topic and an argument. Decide the argument's stance towards that topic. The argument's stance is in favor or against the topic. If the argument supports that topic, answer with 'in favor'; otherwise, if the argument opposes the topic, answer with 'against'." has 7 mentions of "abortion" and 1 mention of "climate change." The input text in task 513 does not consist of tweets but is sourced from Debatepedia (Kobbe et al., 2020). Similarly, task 209[4] appears to be the duplicate of 513 with the same data source and data, but with a different instruction "Given the Target and Argument texts detect the stance that the argument has towards the topic. There are three types of stances 'in favor', 'against', and 'neutral'." Both tasks 209 and 513 are balanced in terms of class labels (Kobbe et al., 2020). The collection also has a stance detection dataset in Spanish and Catalan (Zotova et al., 2020)—Task 1646[5]—with the definition "In this task, we have Spanish and Catalan tweets for automatic stance detection. The data has three labels Against, Favor, and Neutral which express the stance toward the target—independence of Catalonia. If the tweet criticizes the independence of Catalonia then it's 'Against' and if the tweet supports it then it will be labeled as 'Favor' also if the tweets state information or news rather than stating opinion then it will be characterized as 'Neutral'." The dataset is balanced for Favor and Against classes but has fewer Neutral class tweets. Lastly, we find task 890,[6] related to the stance towards global warming with the definition "Read the passage and find if the passage agrees, disagrees, or has a neutral stance on whether Global warming is caused by human activities. Answer only with keyword (a) agrees - if passage agrees with the target (b) disagrees - if passage disagrees with the target (c) neutral - if the given passage neither agrees nor disagrees with the target. You don't need to use external knowledge in this task, and you have to answer based on the given passage." where the passages are news articles, published from Jan. 1, 2000 to April 12, 2020 by 63 U.S. news sources (Luo et al., 2020).[7]

In sum, we find no evidence of SemEval 2016 Task 6 or P-Stance data in the instruction-tuning datasets. Moreover, though FlanT5-XXL has been instruction-tuned on certain stance detection datasets, none of these contain English tweets.

[3] http://bit.ly/3YWaUs6
[4] http://bit.ly/3FvYMac
[5] https://bit.ly/3llyuRp
[6] http://bit.ly/3yOL8LM
[7] Class imbalance: neutral: 873, agree: 777, disagree: 400

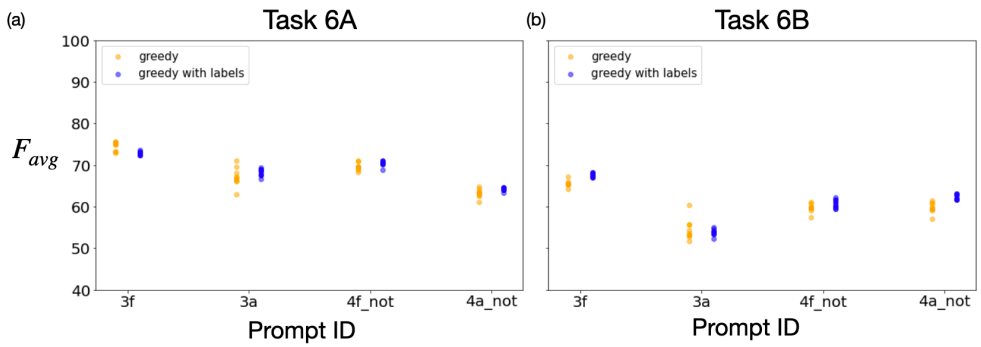

Figure S5: Testing the assumption that FlanT5-XXL will answer 'neutral' when its prediction is in fact 'neutral' instead of answering 'false' to 3f/4a_not or 'true' to 3a/4f_not. Each dot for a given prompt ID represents the performance of FlanT5-XXL for an instruction. The blue dots ("greedy with labels") represent the model's performance when we clarify what op1, op2, and op3 mean in the instruction. The orange dots ("greedy") represent the performance when we do not specify it.

4 Assumption in mapping LLM output to stance label

In the SemEval 2016 Task 6A and 6B, if FlanT5-XXL answers 'false' ('true') to prompt 3f or 4a_not (3a or 4f_not), we assign the predicted stance label to be 'against' ('favor') for evaluation. However since 'neutral' is also a ground truth label in the test dataset, in doing so, we are assuming that the model will answer 'neutral' when its prediction is in fact 'neutral' instead of answering 'false' to 3f/4a_not or 'true' to 3a/4f_not. We test if this assumption is permissible by explicitly clarifying to the model the meaning of each option presented in its instruction (i.e., 'Instruction Outputs'). For example for prompt 3f, in instruction 1, we clarify "Your response to the question should be either {op1} **(if favor)**, {op2} **(if against)**, or {op3} (**if neutral or none**)." where op1, op2, and op3 are 'Instruction Outputs' which in the case of 3f are 'true,' 'false,' and 'neutral respectively.

Fig S5 indicates that the performance does not vary much with or without explicit clarification in the instruction. This led us to conclude that our assumption is passable, and in the rest of the main paper, we stick to not specifying the meaning of op1, op2, and op3 in the instructions.

5 Discrepancies in Baselines Used by Previously Claimed SOTA

Many prior studies claim to use "state-of-the-art" baselines, yet often fail to include the actual best-performing models. For example, in Task 6A, Barbieri et al. (2020) demonstrate that fine-tuned RoBERTa outperforms their chosen baselines—SVM, LSTM, and FastText—achieving $F_{avg} = 71.0$. Similarly, Loureiro et al. (2022) fine-tune BERT to reach $F_{avg} = 72.6$, while Liu et al. (2022) report $F_{avg} = 70.09$. However, earlier works by HaCohen-Kerner et al. (2017) and Zhao and Yang (2020) had already achieved F_{avg} scores of 77.11 and 78.43, respectively—yet these models were never included as baselines for comparison. A similar pattern emerges in Task 6B, where Liang et al. (2022) report $F_{avg} = 50.1$ without comparing against stronger baselines such as Augenstein et al. (2016) or Dey et al. (2017), which achieved F_{avg} scores of 58.03 and 61.57, respectively.

6 F Scores

6.1 SemEval 2016 Task 6A F_{avg} scores from greedy decoding

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	60.81	68.78	74.31	66.69	65.24	42.79
2	59.45	71.94	72.22	59.36	66.59	42.12
3	62.67	72.38	73.64	59.69	67.33	43.16
4	62.15	71.28	75.60	62.09	64.33	43.30
5	62.50	73.93	76.18	68.08	66.35	39.63
6	62.05	71.32	73.86	71.99	67.04	43.38
7	58.53	72.20	75.77	66.97	64.69	42.25
8	61.85	72.51	75.29	62.89	65.27	46.15
9	62.93	72.03	75.47	62.09	64.78	39.95

Table S2: F_{avg} scores for Atheism

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	65.14	56.67	62.05	56.22	58.95	59.54
2	65.72	59.88	64.52	52.36	57.31	59.88
3	65.88	56.31	62.71	55.25	57.07	59.73
4	64.79	56.44	62.66	55.78	57.82	58.00
5	64.47	57.75	60.11	59.24	60.35	57.20
6	65.38	55.10	56.89	56.82	60.21	59.25
7	65.19	54.28	59.25	58.95	58.95	57.39
8	65.23	58.51	61.78	55.91	57.07	57.40
9	65.25	54.14	62.34	55.58	56.88	58.30

Table S3: F_{avg} scores for Feminist Movement

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	61.62	71.69	70.28	75.89	64.74	63.64
2	58.17	73.26	72.88	75.89	65.09	59.52
3	54.19	69.69	75.43	75.89	62.16	60.99
4	57.37	75.62	73.30	74.17	63.80	64.58
5	60.29	71.15	77.37	71.10	62.19	65.77
6	53.46	71.63	76.32	69.88	60.75	64.42
7	59.93	76.43	68.38	66.82	61.83	62.80
8	56.46	72.74	74.88	74.17	64.86	62.53
9	66.02	76.59	68.89	72.58	66.28	64.03

Table S4: F_{avg} scores for Climate Change is a Real Concern

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	79.81	77.04	82.19	63.01	70.55	65.94
2	79.47	77.69	82.07	61.39	66.93	67.15
3	81.30	78.98	81.70	65.47	69.07	66.32
4	80.76	74.98	79.50	64.57	70.34	63.96
5	79.00	79.18	80.86	68.17	72.33	64.20
6	79.44	77.90	81.58	74.14	75.41	63.92
7	78.48	75.36	73.77	66.30	71.40	63.42
8	78.18	76.29	77.80	66.25	71.18	65.54
9	80.03	76.78	78.33	65.67	70.03	64.74

Table S5: F_{avg} scores for Hillary Clinton

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	70.25	65.80	64.11	60.62	64.19	54.53
2	69.54	66.17	65.65	55.82	61.94	57.77
3	70.43	66.54	68.50	61.27	63.15	55.14
4	69.35	65.39	69.43	62.58	64.39	55.48
5	69.98	69.25	67.47	63.00	66.42	53.08
6	70.35	65.71	66.60	65.97	66.18	52.99
7	68.16	64.86	69.64	60.07	64.79	54.28
8	69.52	65.91	69.04	62.01	64.39	54.32
9	67.82	64.77	69.34	59.82	64.03	55.66

Table S6: F_{avg} scores for Legalization of Abortion

6.2 SemEval 2016 Task 6B F_{avg} scores from greedy decoding

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	67.74	61.15	65.54	54.26	59.67	60.94
2	67.01	62.49	67.20	51.70	57.47	61.48
3	68.13	62.90	65.91	55.63	59.92	60.81
4	67.78	60.66	65.46	53.12	59.16	59.39
5	67.33	63.28	65.26	55.84	61.19	57.06
6	67.88	62.58	65.55	60.33	60.77	59.39
7	67.66	60.57	64.19	53.12	59.60	59.08
8	67.47	60.69	65.39	53.62	59.68	60.01
9	67.56	62.08	65.39	52.81	60.04	59.41

Table S7: F_{avg} scores for Donald Trump

6.3 P-Stance F_{avg} scores using greedy decoding

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	74.07	71.87	80.46	72.37	81.80	62.39
2	73.50	81.10	78.70	78.14	81.54	61.29
3	73.18	71.54	81.14	77.78	80.01	63.44
4	73.61	83.03	80.53	79.11	82.93	62.30
5	74.59	80.46	84.43	72.37	79.91	64.48
6	73.01	76.04	84.04	72.99	79.79	64.48
7	72.68	82.54	84.16	76.85	82.13	62.39
8	72.47	82.58	81.96	77.17	81.92	63.55
9	73.70	82.22	79.81	79.11	82.23	61.90

Table S8: F_{avg} scores for Donald Trump

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	82.40	62.85	73.35	77.96	83.87	79.41
2	81.68	77.17	75.68	81.52	84.18	76.93
3	82.51	68.73	77.02	83.15	83.50	80.05
4	82.50	80.24	77.97	82.99	83.05	79.02
5	82.55	78.71	82.73	81.85	84.23	81.53
6	82.24	72.32	81.17	81.72	83.66	81.24
7	82.20	78.79	79.83	83.35	84.28	79.29
8	82.63	79.06	79.26	82.69	83.67	79.57
9	82.24	77.58	78.48	83.09	83.44	79.23

Table S9: F_{avg} scores for Joe Biden

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	77.77	75.52	77.78	75.32	79.51	69.60
2	77.29	79.97	78.72	77.22	80.41	66.58
3	76.96	77.12	80.79	77.62	80.00	71.55
4	76.80	79.09	79.53	78.47	80.98	69.48
5	78.09	78.44	80.61	77.94	80.31	71.42
6	76.48	78.19	81.45	78.10	80.16	71.44
7	77.28	78.64	80.60	79.04	80.60	68.54
8	76.00	79.36	80.94	78.82	80.88	69.72
9	77.44	79.17	80.47	78.79	81.13	69.33

Table S10: F_{avg} scores for Bernie Sanders

6.4 P-Stance F_{avg} scores using PMI decoding

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	80.54	76.98	83.34	78.02	82.55	62.28
2	78.66	81.60	81.97	77.66	81.61	68.79
3	80.11	75.14	82.33	78.08	81.26	67.04
4	80.88	82.39	84.23	81.26	82.40	68.96
5	79.27	80.50	81.97	77.10	81.97	68.33
6	80.54	80.43	82.61	77.23	81.73	69.35
7	79.44	81.26	81.33	78.66	82.34	70.38
8	79.33	82.10	81.70	78.54	82.76	69.57
9	81.63	78.54	83.47	80.84	82.16	68.73

Table S11: F_{avg} scores for Donald Trump

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	80.87	79.67	83.67	83.09	83.31	79.70
2	80.19	80.60	81.77	82.12	83.43	81.14
3	81.92	77.29	81.55	82.47	84.01	81.90
4	81.19	83.51	81.80	83.44	83.90	81.04
5	82.00	81.38	82.54	82.45	83.27	81.75
6	81.18	81.66	80.86	82.14	83.35	81.73
7	80.93	81.10	80.52	81.67	83.70	80.75
8	82.00	83.16	81.50	81.65	84.22	81.16
9	82.79	83.72	81.77	82.77	83.50	81.46

Table S12: F_{avg} scores for Joe Biden

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	75.45	72.89	81.53	78.43	79.98	69.30
2	75.43	74.64	81.49	77.76	79.81	70.17
3	76.44	77.42	81.35	77.12	79.37	71.42
4	74.55	78.39	81.54	77.76	80.46	70.54
5	75.43	76.31	80.28	77.76	80.14	70.54
6	75.07	76.95	80.55	76.95	79.99	70.57
7	73.56	77.91	80.38	76.77	79.52	70.91
8	75.74	80.63	82.29	73.63	80.31	72.91
9	75.92	78.70	81.52	78.39	79.37	71.38

Table S13: F_{avg} scores for Bernie Sanders

6.5 P-Stance F_{avg} scores using AfT decoding

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	80.54	76.98	81.95	78.02	82.55	62.28
2	78.66	81.60	47.24	77.66	81.61	65.10
3	80.11	75.14	82.33	78.08	81.26	67.04
4	80.88	82.39	84.23	81.26	82.40	68.96
5	79.27	80.50	81.72	77.10	81.97	68.33
6	80.54	80.43	82.10	77.23	81.73	69.35
7	79.44	81.26	81.33	78.66	82.34	70.38
8	79.33	82.10	81.70	78.54	82.76	69.57
9	81.63	78.54	83.47	80.84	82.16	68.73

Table S14: F_{avg} scores for Donald Trump

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	80.87	79.67	79.43	83.09	83.31	79.70
2	80.19	80.60	39.36	82.12	83.43	81.14
3	81.92	77.29	81.55	82.47	84.01	81.90
4	81.19	83.51	81.80	83.44	83.90	81.04
5	82.00	81.38	81.77	82.45	83.27	81.77
6	81.18	81.66	79.30	82.14	83.35	81.48
7	80.93	81.10	80.52	81.67	83.70	80.75
8	82.00	83.16	81.50	81.65	84.22	81.16
9	82.79	83.72	81.77	82.77	83.50	81.46

Table S15: F_{avg} scores for Joe Biden

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	75.45	72.89	81.53	78.43	79.98	69.30
2	75.43	74.64	81.49	77.76	79.81	70.17
3	76.44	77.42	81.35	77.12	79.37	71.42
4	74.55	78.39	81.54	77.76	80.46	70.54
5	75.43	76.31	80.28	77.92	80.14	70.54
6	75.07	76.95	80.55	76.95	79.99	70.57
7	73.56	77.91	80.38	76.77	79.52	70.91
8	75.74	80.63	82.29	73.63	80.31	72.91
9	75.92	78.70	81.52	78.39	79.37	71.38

Table S16: F_{avg} scores for Bernie Sanders

7 Accuracy Scores

7.1 SemEval 2016 Task 6A accuracy scores from greedy decoding

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	71.82	70.00	78.18	70.00	71.36	51.82
2	70.45	73.64	77.73	60.91	72.27	59.09
3	71.82	70.00	79.55	61.82	74.09	56.36
4	70.91	70.00	79.09	64.09	71.36	55.45
5	71.82	74.55	78.64	71.36	72.27	46.82
6	70.91	72.27	78.18	75.00	73.64	53.18
7	70.45	73.64	77.73	60.91	72.27	59.09
8	70.00	70.00	76.36	70.00	71.36	55.00
9	71.82	71.36	78.18	64.55	71.82	57.27
10	72.27	70.91	78.64	64.09	71.36	53.64

Table S17: Accuracy scores for Atheism

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	64.91	52.28	60.35	53.33	56.49	59.30
2	66.32	56.14	64.21	48.77	54.04	60.00
3	65.61	50.88	62.81	52.28	54.04	59.30
4	65.61	51.93	61.05	52.63	54.74	56.84
5	64.21	54.04	58.95	56.84	58.25	56.14
6	65.26	49.47	53.68	53.68	57.89	58.25
7	66.32	56.14	64.21	48.77	54.04	60.00
8	65.26	50.53	56.49	56.14	56.14	56.14
9	65.96	55.09	60.00	52.63	54.04	56.14
10	65.26	49.12	60.70	52.28	53.68	57.19

Table S18: Accuracy scores for Feminist Movement

1	67.46	75.15	74.56	75.15	70.41	68.05
2	67.46	75.15	82.84	75.15	71.60	64.50
3	56.21	71.60	83.43	75.15	71.01	66.86
4	66.27	77.51	83.43	74.56	72.19	69.23
5	69.82	75.74	85.21	73.37	70.41	70.41
6	49.70	72.78	78.70	73.37	69.82	68.64
7	67.46	75.15	82.84	75.15	71.60	64.50
8	70.41	84.02	79.29	72.19	69.82	68.05
9	66.27	78.11	84.02	74.56	72.19	67.46
10	72.78	76.33	80.47	73.96	72.78	68.64

Table S19: Accuracy scores for Climate Change is a Real Concern

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	74.92	76.95	72.54	60.00	65.08	64.75
2	74.92	77.97	72.88	57.63	62.71	66.10
3	76.95	75.93	75.25	65.08	64.75	65.76
4	76.95	74.92	70.51	61.36	65.08	63.05
5	74.92	78.64	73.56	66.78	65.42	62.37
6	75.25	77.29	76.27	75.59	68.47	63.73
7	74.92	77.97	72.88	57.63	62.71	66.10
8	74.92	75.93	64.75	62.71	64.75	61.36
9	74.92	75.93	70.17	64.75	65.08	65.08
10	76.61	77.29	70.51	62.37	64.75	63.05

Table S20: Accuracy scores for Hillary Clinton

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	71.79	66.79	66.79	60.36	67.86	61.43
2	72.86	67.14	71.43	56.07	64.64	68.57
3	72.14	66.07	72.86	63.57	65.36	65.36
4	71.07	65.71	72.14	64.29	66.79	64.64
5	72.86	71.07	69.29	64.64	70.36	57.14
6	71.43	66.07	68.57	67.50	70.00	60.00
7	72.86	67.14	71.43	56.07	64.64	68.57
8	71.43	65.36	70.36	61.79	67.14	62.50
9	72.50	66.07	71.79	64.29	67.14	63.21
10	69.64	64.29	72.50	62.14	65.71	64.64

Table S21: Accuracy scores for Legalization of Abortion

7.2 SemEval 2016 Task 6B accuracy scores from greedy decoding

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	56.86	61.10	51.91	51.91	50.07	50.50
2	55.02	60.40	52.19	44.98	48.09	50.78
3	57.14	62.23	53.61	53.18	50.78	50.64
4	56.58	59.83	51.49	48.09	49.08	48.94
5	55.59	60.40	51.20	51.49	49.65	47.52
6	58.13	62.52	57.99	60.68	51.06	50.35
7	55.02	60.40	52.19	44.98	48.09	50.78
8	57.71	59.41	50.64	44.41	48.51	48.66
9	56.44	60.54	51.49	48.37	49.36	50.07
10	57.28	61.53	51.34	47.10	49.79	48.80

Table S22: Accuracy scores for Donald Trump

7.3 P-Stance accuracy scores using greedy decoding

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	76.32	71.94	78.76	67.82	82.37	68.34
2	75.80	81.47	80.05	79.41	82.24	67.57
3	75.55	71.56	82.11	79.28	80.95	68.85
4	76.06	83.27	81.60	80.31	83.40	68.34
5	76.71	80.95	84.68	75.03	80.95	69.50
6	75.42	76.06	84.30	75.55	80.82	69.50
7	75.80	81.47	80.05	79.41	82.24	67.57
8	75.16	82.75	84.68	78.51	82.75	68.34
9	75.03	82.75	82.75	78.64	82.50	68.98
10	76.06	82.37	80.95	80.31	82.75	68.08

Table S23: Accuracy scores for Donald Trump

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	82.95	64.30	71.81	73.29	84.16	80.40
2	82.28	77.18	77.45	81.61	84.43	78.66
3	83.09	68.99	78.26	83.36	83.89	80.94
4	83.09	80.27	79.33	83.09	83.22	80.27
5	83.09	78.79	82.95	82.42	84.70	82.15
6	82.82	72.48	81.48	82.28	84.16	81.88
7	82.28	77.18	77.45	81.61	84.43	78.66
8	82.82	78.79	80.81	83.62	84.56	80.40
9	83.22	79.06	80.40	82.82	83.89	80.67
10	82.82	77.58	79.73	83.22	83.62	80.40

Table S24: Accuracy scores for Joe Biden

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	77.80	76.38	74.80	73.07	79.53	70.24
2	77.32	80.16	78.74	77.32	80.47	67.87
3	77.01	77.48	80.79	77.64	80.00	72.13
4	76.85	79.37	79.53	78.58	81.10	70.24
5	78.11	78.58	80.79	77.95	80.31	71.81
6	76.54	78.58	81.57	78.11	80.16	71.81
7	77.32	80.16	78.74	77.32	80.47	67.87
8	77.32	79.06	80.63	79.06	80.63	69.29
9	76.06	79.69	80.94	78.90	80.94	70.39
10	77.48	79.53	80.47	78.90	81.26	70.08

Table S25: Accuracy scores for Bernie Sanders

7.4 P-Stance accuracy scores using PMI decoding

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	80.82	77.09	83.40	79.28	82.63	68.21
2	79.28	81.85	81.98	79.02	81.72	71.81
3	80.44	75.16	82.37	79.54	81.98	70.79
4	81.34	82.63	84.30	82.11	82.50	72.07
5	79.79	80.69	81.98	78.64	82.11	71.69
6	80.95	80.57	82.63	78.76	81.98	72.20
7	79.28	81.85	81.98	79.02	81.72	71.81
8	79.92	81.47	81.34	79.92	82.63	72.84
9	79.79	82.24	81.72	79.79	83.01	71.94
10	81.85	79.02	83.53	81.72	82.37	72.07

Table S26: Accuracy scores for Donald Trump

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	81.74	80.54	83.89	83.49	83.49	80.67
2	81.21	81.34	81.88	82.68	83.62	81.74
3	82.55	77.32	81.74	82.95	84.43	82.42
4	82.15	83.89	82.01	83.89	84.16	81.34
5	82.82	81.88	82.68	82.95	83.49	82.28
6	82.01	82.15	80.94	82.68	83.62	82.15
7	81.21	81.34	81.88	82.68	83.62	81.74
8	81.88	81.48	80.54	82.42	84.16	81.07
9	82.82	83.49	81.61	82.42	84.56	81.48
10	83.49	84.03	81.88	83.36	83.89	81.88

Table S27: Accuracy scores for Joe Biden

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	75.59	73.23	81.57	78.43	80.00	69.92
2	75.59	74.80	81.57	77.80	79.84	70.71
3	76.54	77.48	81.42	77.17	79.37	71.81
4	74.80	78.43	81.57	77.80	80.47	71.02
5	75.59	76.38	80.31	77.80	80.16	71.02
6	75.28	77.01	80.63	77.01	80.00	71.02
7	75.59	74.80	81.57	77.80	79.84	70.71
8	73.86	77.95	80.63	76.85	79.53	71.34
9	75.91	80.63	82.36	73.86	80.31	73.23
10	76.06	78.74	81.57	78.43	79.37	71.81

Table S28: Accuracy scores for Bernie Sanders

7.5 P-Stance accuracy scores using AfT decoding

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	80.54	76.98	81.95	78.02	82.55	62.28
2	78.66	81.60	47.24	77.66	81.61	65.10
3	80.11	75.14	82.33	78.08	81.26	67.04
4	80.88	82.39	84.23	81.26	82.40	68.96
5	79.27	80.50	81.72	77.10	81.97	68.33
6	80.54	80.43	82.10	77.23	81.73	69.35
7	79.44	81.26	81.33	78.66	82.34	70.38
8	79.33	82.10	81.70	78.54	82.76	69.57
9	81.63	78.54	83.47	80.84	82.16	68.73

Table S29: Accuracy scores for Donald Trump

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	80.87	79.67	79.43	83.09	83.31	79.70
2	80.19	80.60	39.36	82.12	83.43	81.14
3	81.92	77.29	81.55	82.47	84.01	81.90
4	81.19	83.51	81.80	83.44	83.90	81.04
5	82.00	81.38	81.77	82.45	83.27	81.77
6	81.18	81.66	79.30	82.14	83.35	81.48
7	80.93	81.10	80.52	81.67	83.70	80.75
8	82.00	83.16	81.50	81.65	84.22	81.16
9	82.79	83.72	81.77	82.77	83.50	81.46

Table S30: Accuracy scores for Joe Biden

8 Sensitivity to Instructions

Inst.\Prompt	1	2	3f	3a	4f_not	4a_not
1	75.45	72.89	81.53	78.43	79.98	69.30
2	75.43	74.64	81.49	77.76	79.81	70.17
3	76.44	77.42	81.35	77.12	79.37	71.42
4	74.55	78.39	81.54	77.76	80.46	70.54
5	75.43	76.31	80.28	77.92	80.14	70.54
6	75.07	76.95	80.55	76.95	79.99	70.57
7	73.56	77.91	80.38	76.77	79.52	70.91
8	75.74	80.63	82.29	73.63	80.31	72.91
9	75.92	78.70	81.52	78.39	79.37	71.38

Table S31: Accuracy scores for Bernie Sanders

PromptID \Target	AT	CC	LA	FM	HC	F_{avg}
1	0.51	1.30	0.31	0.14	0.34	0.40
2	0.46	0.83	0.44	0.64	0.49	0.31
3f	**0.43**	**1.09**	**0.65**	**0.76**	**0.92**	**0.37**
3a	1.41	1.05	0.92	0.68	1.20	0.77
4f_not	0.37	0.62	0.46	0.46	0.77	0.29
4a_not	0.65	0.64	0.49	0.36	0.42	0.36

Table S32: Standard mean errors of F_{avg} for each prompt ID across the 9 instructions, on SemEval 2016 Task 6A using greedy decoding. The best performer in the task, from Tab. **??** is indicated using boldface.

Prompt ID \Target	F_{avg}
1	**0.11**
2	0.35
3f	*0.26*
3a	0.85
4f_not	0.35
4a_not	0.44

Table S33: Standard mean errors of F_{avg} scores for each prompt ID across the 9 instructions, on SemEval 2016 Task 6B using greedy decoding. The best performer in the task, from Tab. **??** is indicated using boldface, and the second best performer, which happens to be the best performer in Task 6A is shown in italics.

9 Statistical Significance

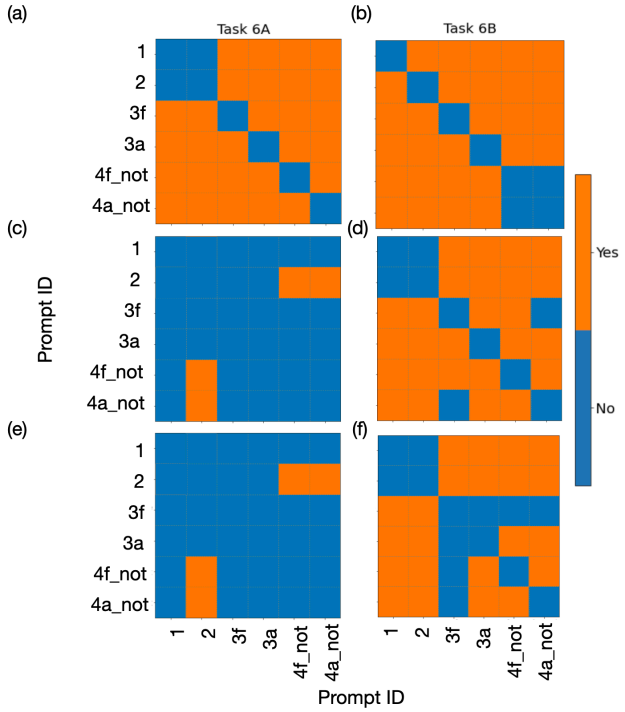

Figure S6: Statistical significance of the performance, measured via F_{avg} scores, between different prompts on the SemEval 2016 Task 6 dataset. The top row is greedy decoding, the middle row is PMI decoding, and the bottom row is AfT decoding. "Yes" and "No" indicates p-value < 0.05 and $p > 0.05$, respectively, obtained independent sample t-test.

Fig. SS6 summarizes the results of an independent sample t-test for SemEval 2016. It shows that the difference in the mean F_{avg} scores between prompts is statistically significant ($p < 0.05$), except in SemEval Task 6A—pair (1,2) from greedy decoding. In PMI, and AFT decoding, just pairs (1,2), (2,4f_not), and (2,4a_not) have statistically significant differences. In SemEval Task 6B—pair (4f_not,4a_not) from greedy decoding, and pairs (1,2), (3f,4a_not) from PMI decoding do not have statistically significant differences. In AfT decoding, pairs (1,2), (3f,3a), (3f,4f_not), (3f,4a_not) do not have statistically significant differences.

Fig. SS7 shows the statistical significance of the difference in the mean F_{avg} scores between prompt pairs. The differences are mostly statistically significant ($p < 0.05$) for greedy decoding except in a few cases—for Donald Trump pairs (2,3f), (2,3a), (2,4f_not), (3f,4f_not), for Joe Biden pairs (1,3a), (2,3f), (3f,4a_not), and for Bernie Sanders pairs (1,3a), (2,3a), (3f,4f_not). There exist statistically significant differences ($p < 0.05$) for PMI decoding except in cases—for Donald Trump pairs (1,2), (2,3a), (3f,4f_not), for Joe Biden pairs (1,2), (1,3f), (1,4f_not), (3f,2), (3a,2), (3a,3f), (2,4a_not), (3f, 4a_not), and for Bernie Sanders pairs (1,2), (3a,2). In case of AfT decoding, significant differences ($p < 0.05$) exist except in cases—(2,1),(3f,1),(3f,2),(3a,2),(3a,3f),(4f_not,3f) for Donald Trump,

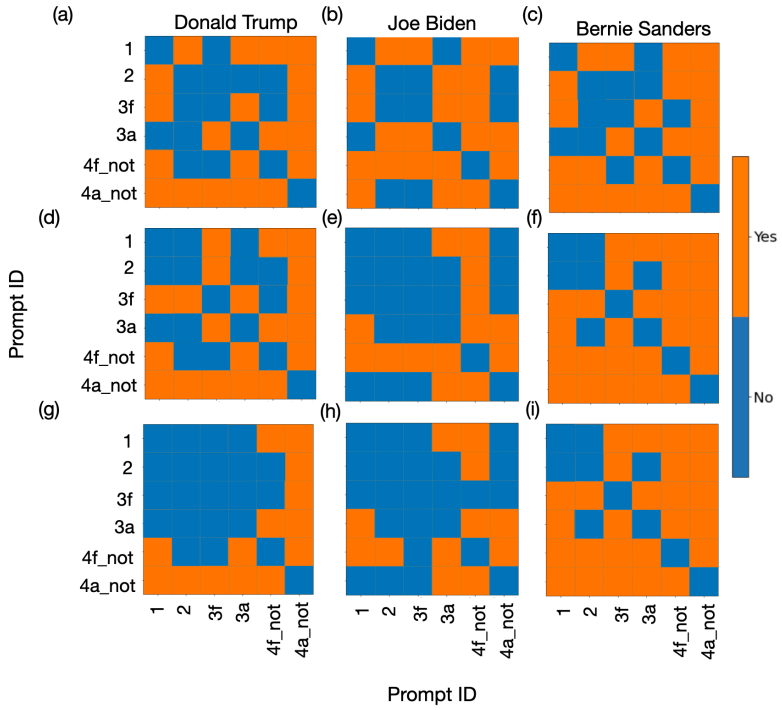

Figure S7: Statistical significance of the performance, measured via F_{avg} scores, between different prompts on the P-Stance dataset. The top row is greedy decoding, the middle row is PMI decoding, and the bottom row is AfT decoding. "Yes" and "No" indicates p-value < 0.05 and $p > 0.05$, respectively, obtained through an independent sample t-test.

(2,1),(3f,1),(3f,2),(3a,2),(3a,3f), (4f_not,3f), (4a_not,1), (4a_not,2), (4a_not,3f) for Joe Biden, and (2,3a) for Bernie Sanders. Overall, the difference between prompt pairs is almost always significant ($p < 0.05$)for Bernie Sanders, regardless of decoding strategy, is mostly significant when greedy decoding is used for Joe Biden, and is significant when greedy or PMI decoding is used for Donald Trump.

10 Hypothesis for Performance Degradation

Contrary to prior expectations (Holtzman et al., 2021), PMI and AfT decoding showed performance degradation compared to greedy decoding on the SemEval 2016 Task 6 dataset. A possible explanation is that the "positivity bias" of FlanT5-XXL aligns with the majority class in the test data when greedy decoding is used. In other words, evaluating with PMI and AfT decoding, which mitigates certain biases in LLM outputs—particularly in classification tasks—may provide a more accurate assessment of the model's true performance. For example, in an extreme case, if all of the test data had the stance "favor," then the model's positivity bias would lead it to perform better in greedy decoding than PMI or AfT. The fact that such a stark difference in performance doesn't appear on the P-stance dataset, which is more balanced, also supports our hypothesis.

We note that the "against" label is the majority label in the test sets of SemEval 2016 Task 6, and to a lesser extent, in P-Stance. Therefore, from our hypothesis, we expect that for prompts 3a and 4f_not, the model's positivity bias would result in greedy performing better than PMI or AfT. This is always true for SemEval 2016 and mostly true for P-Stance. However, this still does not explain why PMI and AfT perform worse than greedy in the other prompts when our hypothesis expects the opposite. For instance, in prompt 3f ("The statement is in favor of $\langle tweet \rangle$"), the model is biased towards outputting "true." In the greedy setting, the model is then expected to assign the label "favor"—the minority label in the test datasets—to a majority of test instances while PMI and AfT should correct this bias and hence result in a better performance than greedy. We leave further exploration of this for future work. Exploring whether the ordering of the instruction options in the prompt can explain this discrepancy is also an interesting future direction Pezeshkpour and Hruschka (2023).

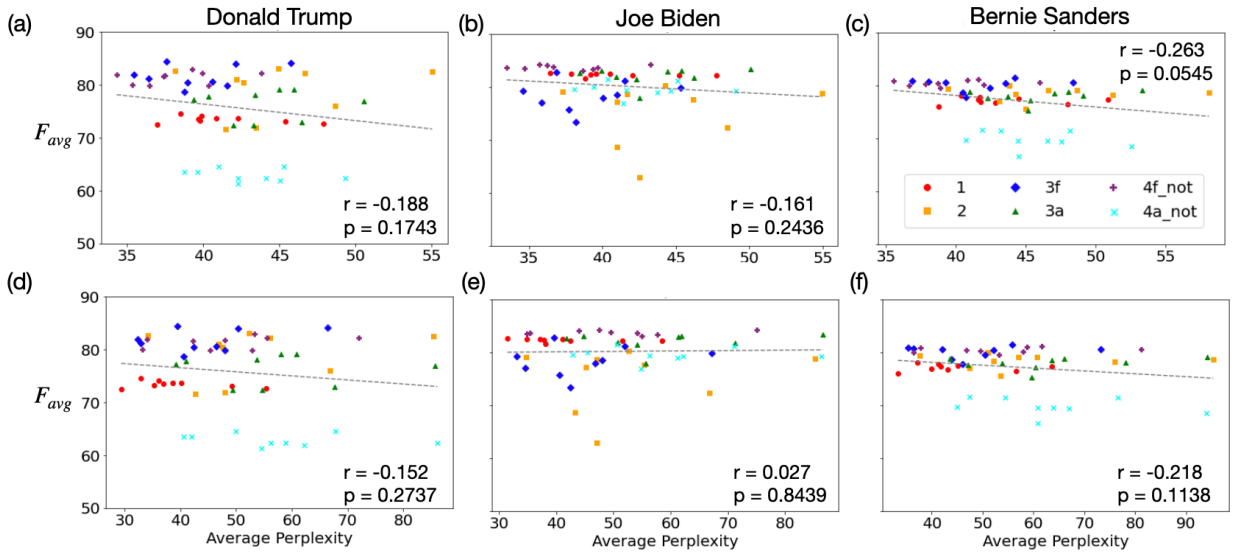

Figure S8: Correlation between prompt perplexity, per prompt ID, per instruction, and F_{avg} scores (from greedy) for each target in the P-Stance dataset. (a,d) Donald Trmup, (b,e) Joe Biden, (c,f) Bernie Sanders. (a,b,c) Prompts with the $\langle tweet \rangle$ object. (d,e,f) Prompts without the $\langle tweet \rangle$ object—context-free prompt. Correlation coefficients are indicated by r and p-values by p.

11 Prompt Perplexity on P-Stance

In Fig. S8 we still see a slightly negative but not significant correlation between prompt perplexities and F_{avg} scores on individual targets in the P-Stance dataset. This agrees with our finding that the difference in performance between prompts is less in the P-Stance dataset compared to the SemEval 2016 Task 6 dataset. Nevertheless, we do observe the two best-performing prompts (3f and 4f_not) occupy the top left of the plots—low perplexity, high F_{avg}. A surprising finding we are yet to understand is the perplexities of context-free prompts being higher than those with context.

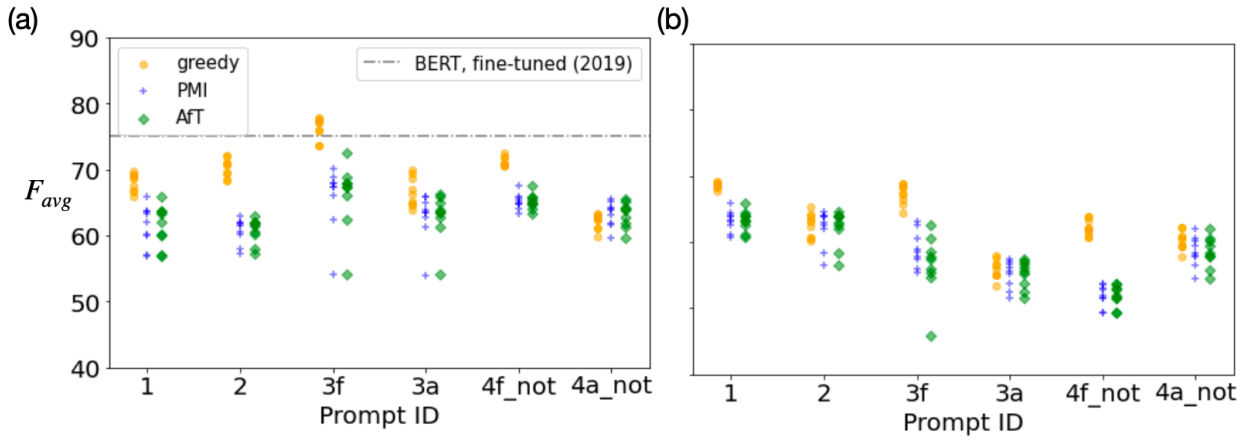

Figure S9: The F_{avg} scores of FlanT5-XXL on pre-processed (a) Task 6A, and (b) Task 6B of SemEval 2016 are shown in comparison against the work that proposed the pre-processing strategy (BERT, fine-tuned (2019)) Each label on the x-axis corresponds to a prompt, and each point on a given prompt ID corresponds to an instruction. The results of three decoding strategies—greedy, PMI, and AfT —are also shown.

12 Pre-processing Tweets

In addition to standard pre-processing steps such as case-folding, Ghosh et al. (2019) expands abbreviations and splits hashtags in tweets, and shows that their model's (fine-tuned BERT) performance improves on SemEval 2016 Task 6. We adopt the same pre-processing strategy and test the performance of FlanT5-XXL on SemEval 2016 Task 6 and the P-Stance dataset.

The results for the SemEval 2016 Task 6 dataset are shown in Fig. S9 along with the model (BERT, fine-tuned (2019)) from Ghosh et al. (2019). We note that every instruction (except 1 and 6) on prompt 3f—the overall best performer in the stance detection task—outperforms the baseline model (BERT, fine-tuned (2019)). We also show the results of the best-performing prompt-instruction pairs (based on F_{avg} score) on SemEval 2016 Task 6A with pre-processing in Tab. S34.

Ghosh et al. (2019) did not test their model's performance on Task 6B, but we have shown FlanT5-XXL's performance with pre-processing in Fig. S9b.

While our approach is capable of outperforming this baseline (Fig. S9), there is often a drop in performance when this pre-processing strategy is employed in both SemEval 2016 Task 6A (Tab. S35) and P-Stance (Tab. S36) data but this is not the case in SemEval 2016 Task 6B (Tab. S35) where the performance marginally improves.

Model\Target	AT	CC	LA	FM	HC	F_{avg}
FlanT5-XXL: 3f (2)	72.22	72.88	65.65	64.52	**82.07**	75.60
BERT (2019)	74.3	44.6	65.7	65.0	71.3	75.1
FlanT5-XXL-P: 3f (8)	**74.46**	**77.00**	**66.31**	**71.51**	81.01	**77.92**

Table S34: F_{avg} scores of best performing FlanT5-XXL on SemEval 2016 Task 6A with (FlanT5-XXL-P) and without pre-processing of tweets compared to the work (BERT, 2019) which proposed the pre-processing strategy (Ghosh et al., 2019). 3f refers to prompt 3f, and (2), (8) to the instruction ID 2 and 8, respectively.

Dataset	Decoding	1	2	3f	3a	4f_not	4a_not
Task 6A	Greedy	**71.5**	**71.18**	74.54	**67.21**	69.62	**63.21**
		68.03	70.28	**76.27**	66.63	**71.22**	62.14
	PMI	**66.48**	**62.51**	64.41	**64.71**	64.89	**65.3**
		61.4	60.71	**65.8**	62.85	**65.18**	63.27
	AfT	**66.48**	**62.51**	64.52	**64.77**	64.89	**65.11**
		61.4	60.71	**66.06**	62.91	**65.18**	63.2
Task 6B	Greedy	67.62	61.82	65.54	54.49	59.72	59.73
		68.59	**62.66**	**67.29**	**55.95**	**62.24**	**60.28**
	PMI	61.91	61.51	57.91	54.1	50.92	58.33
		63.23	**62.11**	**58.94**	**55.11**	**51.91**	**58.61**
	AfT	61.91	61.51	54.44	54.1	50.92	57.9
		63.23	**62.11**	**56.54**	**55.11**	**51.91**	**58.52**

Table S35: Average (across instructions) of F_{avg} on SemEval 2016 Task 6 before and after pre-processing on tweets using strategies such as expanding abbreviations and splitting hashtags (Ghosh et al., 2019). Ghosh et al. (2019) report an F_{avg} score of 75.1 using pre-processing for SemEval 2016 Task 6A. We note that greedy decoding using Prompt 3f can outperform this baseline for most instructions. The top (bottom) row of each decoding strategy represents performance before (after) pre-processing. The boldface prompt indicates the best performing for each decoding strategy—prompt pair. The underline indicates the best-performing prompt for each decoding strategy.

Target	Decoding	1	2	3f	3a	4f_not	4a_not
Trump	Greedy	**73.42**	**79.04**	**81.69**	**76.21**	**81.36**	**62.91**
		71.93	77.66	78.16	75.46	79.04	62.62
	PMI	**80.05**	**79.88**	**82.55**	**78.60**	**82.09**	**68.16**
		78.39	78.55	80.52	77.90	80.75	67.58
	AfT	**80.05**	**79.88**	**78.60**	**78.45**	**67.75**	**82.09**
		78.39	78.55	77.90	76.80	67.03	80.75
Biden	Greedy	**82.33**	**75.05**	**78.39**	**82.03**	**83.76**	**79.58**
		79.60	74.13	71.24	78.56	80.53	77.69
	PMI	**81.45**	**81.34**	**81.77**	**82.42**	**83.63**	**81.18**
		79.41	77.98	75.92	80.32	80.75	79.60
	AfT	**81.45**	**81.34**	**82.42**	**76.33**	**81.16**	**83.63**
		79.41	77.98	80.32	71.50	79.58	80.75
Bernie	Greedy	**77.12**	**78.39**	**80.10**	**77.93**	**80.44**	69.74
		76.66	78.27	77.78	76.17	79.57	**70.00**
	PMI	**75.29**	**77.09**	**81.21**	**77.17**	**79.88**	70.86
		75.17	75.70	78.73	76.07	79.32	**71.23**
	AfT	**75.29**	**77.09**	**77.19**	**81.21**	70.86	**79.88**
		75.17	75.70	76.14	78.73	**71.23**	79.32

Table S36: Average (across instructions) of F_{avg} on P-Stance before and after pre-processing on tweets using strategies such as expanding abbreviations and splitting hashtags (Ghosh et al., 2019). The top (bottom) row of each decoding strategy represents performance before (after) pre-processing. The boldface indicates the best performing for each decoding strategy—prompt pair. The underline indicates the best-performing prompt for each decoding strategy.

References

Augenstein, I., Rocktäschel, T., Vlachos, A., and Bontcheva, K. (2016). Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*.

Barbieri, F., Camacho-Collados, J., Neves, L., and Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Valter, D., Narang, S., Mishra, G., Yu, A. W., Zhao, V., Huang, Y., Dai, A. M., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2022). Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Dey, K., Shrivastava, R., and Kaushik, S. (2017). Twitter stance detection—a subjectivity and sentiment polarity inspired two-phase approach. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pages 365–372. IEEE.

Ghosh, S., Singhania, P., Singh, S., Rudra, K., and Ghosh, S. (2019). Stance detection in web and social media: a comparative study. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 75–87. Springer.

HaCohen-Kerner, Y., Ido, Z., and Ya'akobov, R. (2017). Stance classification of tweets using skip char ngrams. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part III 10*, pages 266–278. Springer.

Holtzman, A., West, P., Shwartz, V., Choi, Y., and Zettlemoyer, L. (2021). Surface form competition: Why the highest probability answer isn't always right. *arXiv preprint arXiv:2104.08315*.

Kobbe, J., Hulpuș, I., and Stuckenschmidt, H. (2020). Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 50–60.

Liang, B., Chen, Z., Gui, L., He, Y., Yang, M., and Xu, R. (2022). Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 2738–2747.

Liu, Y., Zhang, X. F., Wegsman, D., Beauchamp, N., and Wang, L. (2022). Politics: pretraining with same-story article comparison for ideology prediction and stance detection. *arXiv preprint arXiv:2205.00619*.

Longpre, S. and Roberts, A. (2023). The flan collection: Advancing open source methods for instruction tuning.

Loureiro, D., Barbieri, F., Neves, L., Anke, L. E., and Camacho-Collados, J. (2022). Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Luo, Y., Card, D., and Jurafsky, D. (2020). Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

Pezeshkpour, P. and Hruschka, E. (2023). Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H. G., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Patel, M., Pal, K. K., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P. R., Verma, P., Puri, R. S., Karia, R., Sampat, S. K., Doshi, S., Mishra, S., Reddy, S., Patro, S., Dixit, T., Shen, X., Baral, C., Choi, Y., Smith, N. A., Hajishirzi, H., and Khashabi, D. (2022). Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Conference on Empirical Methods in Natural Language Processing*.

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

Zotova, E., Agerri, R., Nuñez, M., and Rigau, G. (2020). Multilingual stance detection in tweets: The catalonia independence corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1368–1375.