

PERSUASIVE ARGUMENT PREDICTION

ADITYA SINGH

ABSTRACT. This project replicates the pair task experiment from Winning Arguments ([1]) using the GPT-4o mini API.

CONTENTS

1. Introduction	1
2. Experiments	2
2.1. Experiment 1 - Basic Prompt	2
2.2. Experiment 2 - Explain-then-predict Prompt	2
2.3. Experiment 3 - Explain-then-predict-2 Prompt	3
3. Conclusion	3
References	4
4. Appendix	4

1. INTRODUCTION

In Winning Arguments ([1]), Tan et al analyze the dynamics of persuasion through a pair prediction problem (section 4.1 of [1]). Given an original post on the CMV (Change My View) Reddit, pairs of counter-responses that attempt to change the OP’s (original poster’s) view are collected. One and only one of the counter-responses is successful in doing so. Moreover, the pairs of counter-responses are selected so that they have high Jaccard similarity (content overlap – see section 4.1 of [1] for a precise definition), so that the emphasis is not on *what* is being said, but rather *how* it is being said.

In this experiment, we focus on the *root reply* version of the pair prediction problem, in which only the first reply to the OP’s post is included for each counter-argument, as opposed to the entire chain of replies. In the original paper, Tan et al train various logistic regression models using stylistic features of the arguments (see section 4.3 of [1]). In this replication experiment, we format the original post and the pair of counter responses into a prompt for GPT-4o mini, and use its generation as the classification prediction.

In particular, we experiment with three separate prompting schemes, which we call “basic”, “explain-then-predict”, and “explain-then-predict-2”. The goal of the experiment is not a high classification accuracy, but an exploration of *natural language explanation*. **How can we prompt LLMs to not only classify nontrivial**

tasks such as persuasive argument prediction, but also have a transparent thinking process/explanation? Due to time and cost constraints, all experiments are done on the same subset of 500 test examples from the heldout pair data.

Our main results are as follows: the pair task prediction is *extremely difficult* for GPT-4o mini. In the three experiments, we report classification accuracies of 58.81%, 57.80%, and 53.80% respectively. However, the most important highlight of this paper is the development of the “explain-then-predict-2” prompt, which can be viewed as rectifying a major shortcoming in the “explain-then-predict” prompt. As is discussed in [subsection 2.2](#), when GPT-4o mini cogenerates an explanation and prediction from a single prompt, it likely does not properly separate the two stages, but instead tailors its explanation so that its subsequent prediction will match well. To address this shortcoming, we make use of the “assistant message” functionality included in the GPT API, in which one can feed GPT’s response from a previous prompt back into the current prompt as context. Thus, explain-then-predict-2 separates the analysis and prediction into two separate parts, resulting in a more unbiased analysis as is discussed in [subsection 2.3](#).

Although the explain-then-predict-2 prompt actually has the worst classification accuracy, our analysis in [subsection 2.3](#) shows that from the perspective of *unbiased natural language explanation*, it is actually the most successful. The fact that GPT-4o mini does the worst with this prompting scheme may be because it is forced to make its decision based on its unbiased analysis. In the basic prompt, in which GPT can reach its classification decision in any way it wants since we do not ask for an explanation, it may use “hacky” solutions such as knowledge of spurious patterns it has gathered in pre-training that help it with the pair prediction task. On the other hand, when GPT-4o mini has to classify based on its unbiased analysis, there is often no obvious answer. If we want to use language models as tools to analyze concepts such as persuasion, then a two-prompt style approach as in explain-then-predict-2 seems like a promising option.

2. EXPERIMENTS

2.1. Experiment 1 - Basic Prompt. As a baseline experiment, we do not ask GPT-4o mini for an explanation or thinking process, and have it just predict which argument was successful. We use the “basic” prompt shown in [section 4](#). Out of the 500 examples, GPT-4o mini records an accuracy of 58.81%.

2.2. Experiment 2 - Explain-then-predict Prompt. In the second experiment, we use the “explain-then-predict” prompt shown in [section 4](#), and record an accuracy of 57.80% across the 500 examples. We notify GPT-4o mini that the pairs of responses have been curated for high content overlap, and ask it to prioritize stylistic analysis as in the original paper. We explicitly suggest four distinct areas of stylistic analysis to give structure to an intermediate thinking process for GPT-4o mini.

While GPT-4o mini is instructed to first explain, and then predict, we observe a phenomenon where the explanations are “prediction-aware,” exemplified in the “explain-then-predict” sample response shown in [section 4](#). Across all four stylistic categories we ask GPT-4o mini to look at (tone and emotional approach, rhetorical techniques, sentence structure and readability, and engagement with the original poster), GPT-4o mini praises the first response while criticizing the second response,

and finally predicts that the first response was successful in persuading the OP. We observe this phenomenon of GPT-4o mini preferring the label it is going to predict across all four categories in nearly all the examples, which the reader can further inspect at [2]. Thus GPT-4o mini is not honestly thinking about the analysis as an intermediate step before the classification decision, but tailoring its analysis so that its subsequent label will fit. We address this shortcoming in the next experiment.

2.3. Experiment 3 - Explain-then-predict-2 Prompt. In the third and final experiment, we use the “explain-then-predict-2” prompt shown in section 4, and record an accuracy of 53.80% across the 500 examples. We separate the prompting process into two separate prompts – first, we feed GPT the original post along with the pair of responses, and ask it to analyze the persuasion style of the responses without telling it to predict. Using the assistant message functionality of the GPT API, we then feed GPT’s “prediction-unaware” analysis back in as context, along with the direction to classify which argument succeeded *based on its prior analysis*.

To investigate the effects of this approach, we run GPT-4o mini twice on the same prompt, and compare its responses, shown in “explain-then-predict-2 prompt response 1” and “explain-then-predict-2 prompt response 2” in section 4. The first major observation is that since GPT-4o mini does not know it needs to predict based on the first prompt, it gives an unbiased analysis of the responses across the four categories. In both runs, as can be seen in the conclusion sections of the sample responses, GPT-4o mini decides that the comparison comes down to the neutral logical reasoning approach of the first argument versus the direct confrontational approach of the second argument. In the first run, GPT-4o mini decides that the more direct approach of the second argument is better and predicts the second argument as successful, while in the second run, GPT-4o mini decides that the calmer logical approach of the first argument is better and predicts the first argument as successful.

Based on GPT-4o mini’s unbiased reasoning after the first part of the prompt, there is no clear correct answer as to which argument was successful, and the comparison of these two runs reflects that uncertainty.

3. CONCLUSION

In summary, we observe a surprising trend where the more faithful GPT-4o mini must be to independent analysis of the persuasion style of the responses, the worse its classification accuracy becomes. One possible explanation for this is the inherent difficulty in the pair prediction problem, where it is relatively subjective as to which argument will persuade the OP. Thus, if GPT is free to predict without also producing a coherent explanation, it may be able to rely on hacky solutions such as spurious patterns it has identified in its pretraining when it comes to similar tasks.

The worst classification performance comes from the two-part prompting scheme as was discussed in subsection 2.3, and a reasonable explanation for this is that GPT-4o mini must base its decision on its prior unbiased analysis, from which there is often no clear preferable answer.

The goal of *natural language explanation* is to improve the usability and interpretability of LLMs by forcing them to generate coherent natural language explanations to accompany classification decisions. If we wish to treat the natural language

explanations as indicative of the LLM’s intermediate thinking process when coming to its decision, then a two-prompt style approach as in “explain-then-predict-2” which fully separates the explanation context and prediction context should be preferred to a joint prompt like “explain-then-predict”.

REFERENCES

- [1] Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions, Chenhao Tan and Vlad Niculae and Cristian Danescu-Niculescu-Mizil and Lillian Lee, 2016, Proceedings of WWW.
- [2] <https://github.com/adsingh-64/CMV>

4. APPENDIX

Basic Prompt

Task: Analyze two responses to an original post and predict which successfully changed the original poster’s view. Provide your prediction by stating either ‘\$\$Answer: First Response\$\$’ or ‘\$\$Answer: Second Response\$\$’.

Original Post Title: {entry[‘op_title’]}

Original Post: {entry[‘op_text’]}

First Response: {entry[‘argument_a’]}

Second Response: {entry[‘argument_b’]}

Explain-Then-Predict Prompt

Task: Analyze two responses to an original post and predict which successfully changed the original poster’s view.

1. Compare the responses, focusing on stylistic elements since they have high Jaccard similarity (content overlap). Consider:

- Tone and emotional approach
- Rhetorical techniques
- Sentence structure and readability
- Engagement with the original poster

2. Provide your prediction by stating either ‘\$\$Answer: First Response\$\$’ or ‘\$\$Answer: Second Response\$\$’

3. Briefly explain the key factor(s) that led to your prediction.

Original Post Title: {entry[‘op_title’]}

Original Post: {entry[‘op_text’]}

First Response: {entry[‘argument_a’]}

Second Response: {entry[‘argument_b’]}

Explain-Then-Predict-2 Prompt

Task: Analyze two responses to an original post that try to change the original poster's view.

Focus on comparing stylistic elements since they have high Jaccard similarity (content overlap). Consider:

- Tone and emotional approach
- Rhetorical techniques
- Sentence structure and readability
- Engagement with the original poster

Original Post Title: {entry['op_title']}

Original Post: {entry['op_text']}

First Response: {entry['argument_a']}

Second Response: {entry['argument_b']}

MODEL RESPONSE

Now, based on your analysis, predict which response successfully changed the original poster's view.

1. Provide your prediction by stating either '\$\$Answer: First Response\$\$' or '\$\$Answer: Second Response\$\$'
2. Briefly explain the key factor(s) that led to your prediction.

Explain-Then-Predict Prompt Sample Response

1. **Comparison of Responses**:

Tone and Emotional Approach:

- The first response maintains a factual and somewhat analytical tone. It seeks to clarify legal principles relevant to the case while expressing a level of sympathy towards the original poster's misunderstanding of liability.
- The second response is more confrontational, directly challenging the original poster's viewpoint and asserting the notion that their reasoning overlooks critical aspects of the case. It carries a slightly accusatory tone by stating that the original poster is arguing from a "wrong stand point," which may provoke defensiveness.

Rhetorical Techniques:

- The first response employs logical reasoning by referencing the principle of 'Comparative Negligence' and discussing McDonald's responsibility in a structured manner. It invites the reader to understand the nuances of legal responsibility without dismissing the original poster outright.
- The second response uses contrast more aggressively, highlighting the shortcomings in the original poster's argument regarding customer expectations and pointing out broader societal standards in food service.

This technique serves to emphasize the perceived ignorance in the original argument.

Sentence Structure and Readability:

- The first response is well-structured, with clear sentences that guide the reader through its logic step-by-step. Its readability is enhanced by straightforward language and organization.
- The second response, while also structured, uses longer and more complex sentences. This may make it harder for some readers to follow the argument effectively. It focuses heavily on exposition and criticism, which can detract from a smooth reading experience.

Engagement with the Original Poster:

- The first response tries to engage by addressing potential gaps in the original post's argument without directly undermining the original poster's position. This could encourage a more constructive dialogue.
- The second response actively dismantles the original poster's argument, which may provoke a defensive reaction rather than fostering a collaborative discussion. It does not pose questions or invite further exploration of the topic, potentially limiting engagement.

2. **Prediction**:

Answer: First Response

3. **Key Factors for Prediction**:

The first response effectively uses a logical, respectful approach that clarifies legal concepts and encourages understanding while remaining factual and non-confrontational. This style is more conducive to changing someone's view, as it invites consideration of new information rather than leading to defensiveness. The second response, however, is more critical and aggressive in its approach, which may alienate the original poster and reduce the likelihood of them re-evaluating their stance.

Explain-Then-Predict-2 Prompt Sample Response 1

Both responses to the original post share a high degree of content overlap while employing distinct rhetorical strategies and stylistic choices aimed at persuading the original poster (OP) to reconsider their views on McDonald's liability in the coffee spill incident. Let's analyze the key differences:

Tone and Emotional Approach:

- First Response: The tone is formal and informational. It employs legal terminology and references established principles like "Comparative Negligence," which gives the response an authoritative and scholarly feel. The language is neutral yet assertive, aiming to present facts rather than evoke emotional responses.
- Second Response: This response takes on a more conversational and somewhat provocative tone. The phrasing "you're arguing from the wrong

standpoint” can come across as dismissive or confrontational, which may alienate the OP. However, it aims to engage the reader by incorporating the audience directly into the discussion.

****Rhetorical Techniques**:**

- First Response: The argument is structured logically, beginning with an explanation of the legal framework surrounding the case, which helps to build a credible foundation. The points are carefully sequenced, concluding with the assertion of liability based on evidence. The response invokes authority in law but avoids strong emotional appeals.
- Second Response: This response employs a more confrontational rhetorical style. It uses phrases like “the loose thread in your reasoning” to highlight flaws in the OP’s argument, aiming to provoke critical thinking by juxtaposing the OP’s views against broader industry standards. This approach can appeal to an emotional response but may risk coming off as aggressive.

****Sentence Structure and Readability**:**

- First Response: The sentences are predominantly complex, containing multi-clause constructions that present detailed information. It demands focused reading but provides a wealth of information that underscores its points. The style maintains clarity while introducing legal concepts.
- Second Response: This response features shorter, more direct sentences that make it easier to digest at a glance. The informal structure encourages engagement, inviting the OP to reflect on their perspective without getting bogged down by legalese. However, this may also oversimplify a complex legal issue.

****Engagement with the Original Poster**:**

- First Response: The engagement is indirect, focusing heavily on the legal aspect rather than relating directly to the OP’s experience of the incident. Although it provides a compelling argument, it does not actively solicit the OP’s input or feelings, which might limit dialogue.
- Second Response: The response actively critiques the OP’s reasoning and invites further thought, which fosters a more interactive discussion. It seeks to challenge the OP’s views head-on, potentially opening up a more dynamic conversation. By doing so, it may resonate more deeply with the OP, despite its confrontational tone.

****Conclusion**:**

In summary, the first response appeals through authority and logical structure, positioning itself as a factual rebuttal without emotional engagement. In contrast, the second response attempts to engage more personally with the OP, albeit with a tone that could be perceived as aggressive. Both approaches present valid points but do so through markedly different stylistic choices, impacting their effectiveness in changing the OP’s viewpoint.

SECOND PART OF ANALYZE-THEN-PREDICT PROMPT

\$\$Answer: Second Response\$\$

The key factors leading to this prediction are the engagement style and emotional appeal of the second response. Unlike the first response, which primarily presents logical arguments and legal principles, the second response actively challenges the original poster's reasoning and encourages them to think critically about their stance. This confrontational yet engaging approach is more likely to resonate on a personal level, potentially prompting the OP to reconsider their views. The use of relatable analogies regarding industry standards also makes the argument more accessible, thus increasing the chance of influencing the OP's perspective. Additionally, the second response invites dialogue, opening a path for further discussion that could lead to a change in opinion.

Explain-Then-Predict-2 Prompt Sample Response 2

Comparing the two responses to the original post reveals notable differences in tone, emotional approach, rhetorical techniques, sentence structure, readability, and engagement with the original poster.

****Tone and Emotional Approach**:**

- The First Response maintains a formal and informative tone, focusing on the legal concepts of "Comparative Negligence" and the responsibility of McDonald's without expressing strong emotional reactions. It employs a cautious approach, aiming to present facts and clarify the legal framework of the case. The emphasis seems to be on correcting the original poster's misunderstanding of liability.
- The Second Response, on the other hand, carries a more confrontational and assertive tone. It directly challenges the original poster's viewpoint by stating, "you're arguing from the wrong standpoint." This phrasing establishes an immediate sense of disagreement and a less neutral engagement compared to the First Response.

****Rhetorical Techniques**:**

- The First Response uses a logical structure that relies on legal principles to make its case. It introduces key terms (like "Comparative Negligence") and explains them succinctly, aiming to educate the reader. This approach builds credibility and supports the argument with factual content rather than emotional appeal.
- The Second Response employs a rhetorical contrast by emphasizing the difference between customer responsibility and the establishment's duty to serve safe products. It presents a "middle ground" argument, making it clear that while customers should be careful, there are standards that businesses must also meet. This method appeals to the reader's sense of

fairness and responsibility..

****Sentence Structure and Readability**:**

- The First Response features relatively complex sentence structures but maintains clarity. It lays out ideas in a well-organized manner, making it easy for the reader to follow the argument. For example, it breaks down the implications of the jury's findings without overwhelming the reader.
- The Second Response uses shorter sentences and a more straightforward style, which can enhance readability. However, certain phrases, like "I see it," and "loose thread in your reasoning," may come off as condescending or dismissive, impacting how engaging it feels.

****Engagement with the Original Poster**:**

- The First Response interacts with the original post in a respectful manner, focusing on clarification and education. It does not directly attack the poster's intelligence or understanding but seeks to expand upon the notion of liability found within the legal context.
- The Second Response launches directly into counterarguments. While it engages with the original poster's claims, it does so in a way that undermines their stance. Phrases like "not mutually exclusive" and "very extensive middle ground" imply a critique of the original poster's logic, which could alienate the original poster rather than encourage constructive dialogue.

****Conclusion**:**

Overall, the First Response sets a more respectful and informative tone while focusing on legal principles, whereas the Second Response adopts a more assertive and confrontational approach, directly challenging the original poster's reasoning. The differing tones and engagement styles highlight various pathways for persuasion and dialogue within this discussion.

SECOND PART OF ANALYZE-THEN-PREDICT PROMPT

\$\$Answer: First Response\$\$

The key factors leading to this prediction include the respectful and informative tone of the First Response, which fosters a more constructive dialogue than the confrontational approach of the Second Response. The First Response employs logical argumentation grounded in legal principles, which could resonate more effectively with someone who is open to reconsidering their position. By clarifying the concept of "Comparative Negligence" and presenting factual information without attacking the original poster's viewpoint, the First Response makes a strong case that is likely to encourage reflection and possibly shift the original poster's perspective. In contrast, the Second Response's dismissive tone and confrontational rhetoric may alienate the original poster, reducing the likelihood of changing their viewpoint.