Evaluating GPT-3.5 Turbo for Predicting Persuasive Arguments

Justin Hwang

Chicago Human + AI Lab Report

12 May 2025

Abstract

This project investigates whether large language models (LLMs), specifically GPT-3.5 Turbo, can predict which of two responses in an online discussion is more persuasive. Using the ChangeMyView (CMV) dataset, we prompt GPT to select the reply most likely to earn a “delta” from the original poster, which signals a persuasive impact. We compare two prompting strategies: predict-then-explain (PTE), where GPT chooses first and then justifies its answer, and explain-then-predict (ETP), where it reasons before committing to a choice. Across 500 randomized A/B reply pairs, the ETP strategy outperforms predict-then-explain, achieving 58.2% accuracy compared to 52.8% under deterministic settings (temperature = 0). However, we find that a higher temperature of 0.5 narrows the discrepancy by improving PTE performance, while ETP remains stable. Subsequent feature analysis reveals that GPT relies heavily on surface cues such as length and formatting.

1 Introduction

Persuasive communication is central to effective debate, education, and decision-making. As LLMs become more prominent in real-world applications, it is important to understand whether they can identify persuasive arguments, and what features influence their judgment.

This project builds on earlier work by Tan et al. (2016), which used logistic regression with handcrafted features to predict which replies in Reddit’s r/ChangeMyView (CMV) forum were successful in changing a poster’s opinion [1]. In CMV, users submit opinions and award a “delta” (Δ) to replies that successfully alter their view. This high-quality environment provides a suitable dataset for studying persuasion.

We ask whether a modern LLM, given two replies to the same post, can predict which one earned the delta. More importantly, we investigate how prompting structure affects this ability. Specifically, we compare two strategies:

- **Predict-then-explain (PTE)**
  - GPT selects a reply and then justifies its choice.

- **Explain-then-predict (ETP)**
    - GPT explains which reply is more persuasive, then states its decision.

This framing allows us to test whether reasoning before commitment improves judgment. Additionally, we vary the model's temperature to observe how randomness interacts with prompting style.

Our goal is to both evaluate zero-shot LLM performance, and reflect on how reasoning order, input features, and randomness affect persuasiveness prediction. We aim to assess the extent to which GPT-3.5 can replicate human judgments of argument quality.

2 Dataset

We use the **CMV held-out pair dataset**, which contains an original post (op_text) and two replies: one that earned a delta (positive), and one that did not (negative).

To control cost and ensure clarity in evaluation, we sample 500 post-reply pairs and randomize the order of the two replies for each instance. This helps mitigate position bias and simulates a true A/B comparison format. Each sample is labeled with the correct answer ("A" or "B") according to which reply received the delta.

3 Method

3.1. Prompting Strategy

We test two variants of prompt design, outlined in Figure 1 below. For the "predict-then-explain" (PTE) prompt, GPT is asked to choose the more persuasive reply first, then explain its reasoning. Conversely, for "explain-then-predict" (ETP), GPT is asked to explain which reply is more persuasive and then make a choice.

Each model call is made independently using OpenAI's GPT-3.5 Turbo API. The model receives the post and both replies, randomized as Reply A and Reply B, and is evaluated based on whether its choice matches the reply that earned a delta.

Predict-then-explain prompt instruction	Explain-then-predict prompt instruction

`A Reddit user posted:`
`"[POST]"`
 	
`Two users replied:`
 	
`Reply A:`
`"[Reply A]"`
 	
`Reply B:`
`"[Reply B]"`
	`A Reddit user posted:`
	`"[POST]"`
	
	`Two users replied:`
	
	`Reply A:`
	`"[Reply A]"`
	
	`Reply B:`
	`"[Reply B]"`

`Which reply is more persuasive and more likely to change the original poster's view? Answer with "A" or "B", then explain your reasoning.`	`Explain which reply is more persuasive and why. After your explanation, state your choice: "A" or "B".`

Figure 1. Prompt instruction for GPT according to predicting/explaining order.

3.2. Temperature Values

We selected two representative temperature settings to test how randomness affects GPT's evaluation of persuasive arguments. A temperature of 0 forces the model to be deterministic and conservative, producing the most likely output based on its training. This setting is commonly used when consistency is prioritized, especially in classification tasks.

In contrast, a temperature of 0.5 allows for more variability and creativity in generation, without being too random. Our goal in including it was to assess whether a looser sampling distribution might lead the model to explore different reasoning strategies that could affect judgment.

3.3. Evaluation

We extract GPT's choice ("A" or "B") using a parser that looks for the first valid letter in the response. For each prompt type, we compute the accuracy, calculated as the proportion of examples where the GPT's choice matches the correct label. Furthermore, we examine the confusion matrix for each prompt type. Finally, we examine textual features such as the number of words, characters, questions, hedges, links, and bullets to determine which factors correlate to GPT's choice.

4 Results

4.1. Accuracy

With the temperature value set to 0, ETP prompting yields a 5.40-point accuracy gain over the PTE approach, as shown in Figure 2. This suggests that generating an explanation before committing to a choice helps the model reason more effectively.

With the temperature value set to 0.5, GPT's ETP approach performs similarly. However, the PTE approach performs substantially better, trailing the ETP approach by only 1.4 points. This narrowed advantage suggests that some of the benefit of explanation-first prompting at temperature = 0 may come from mitigating model overconfidence or positional bias.

Prompt Strategy	Temperature = 0	Temperature = 0.5
Predict-Then-Explain (PTE)	52.80%	57.00%
Explain-Then-Predict (ETP)	**58.20%**	**58.40%**

Figure 2. Accuracy according to prompt strategy and temperature.

4.2. Confusion Matrix and Statistical Comparison

Figure 3 shows the confusion matrices for both prompting strategies at temperature = 0, where the model operates deterministically. The ETP strategy improves correct identification of persuasive B replies by 13 points (126 vs. 113), while also boosting A-corrects by 14 (165 vs. 151). This indicates that reasoning before decision-making reduces impulsive or positionally-biased choices.

To determine whether this observed performance difference is statistically meaningful, we conducted **McNemar's test** on paired predictions across all 500 examples. This test is appropriate for evaluating whether two classifiers differ significantly in accuracy when applied to the same instances. The test yielded a p-value of $p < 0.001$, which indicates that the accuracy gain from ETP is statistically significant. Together, the confusion matrix and inferential test show that explanation-first prompting yields more robust judgments of persuasiveness at temperature = 0.

Figure 4 presents the corresponding confusion matrices at temperature = 0.5, where the gap between prompting strategies narrows. Here, the relative improvement from ETP is smaller, and a statistical test shows no significant difference between the two approaches ($p = 0.48$), which indicates that increased randomness may reduce the need for explanation-first reasoning.

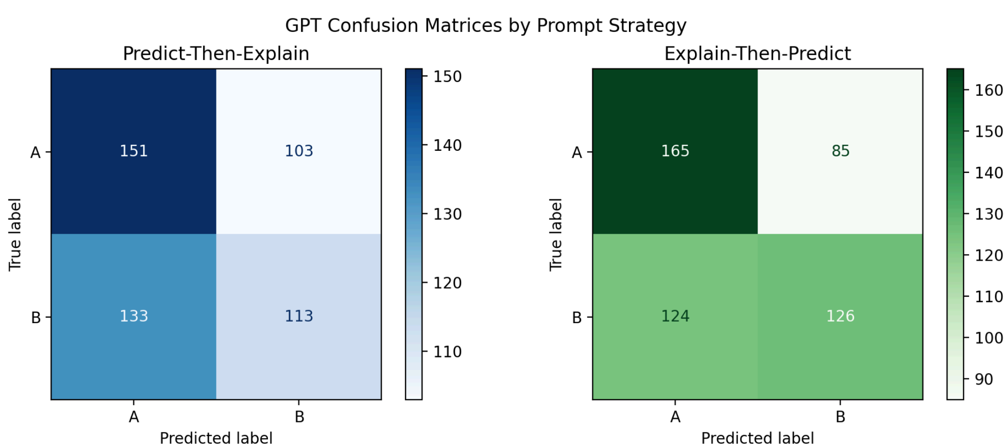

Figure 3. Confusion matrices comparing the two prompting strategies at temperature = 0.

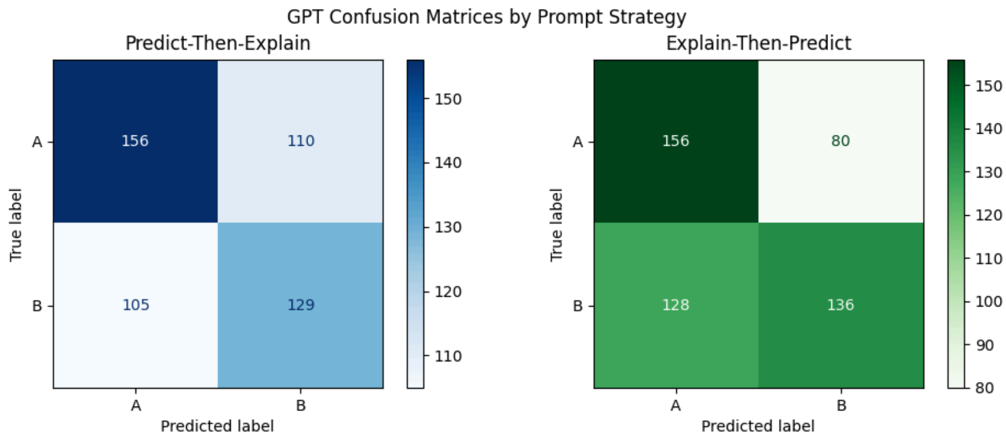

Figure 4. Confusion matrices comparing the two prompting strategies at temperature = 0.5.

4.3. Case Study: Batman

To illustrate the practical difference between prompting styles, we include an example from the CMV dataset using a temperature of 0. The original post argues that Batman's actions are ineffective and self-serving, suggesting that systemic investments would be more impactful than vigilantism.

In Figure 5, two replies are presented, with Reply A earning a delta (Δ) from the original poster. Then, GPT's response to each prompt is shown.

Notably, GPT spends more time summarizing the arguments in the chosen reply (bolded below) in the ETP condition, leading to the correct classification.

Original Post	
He's intelligent and wealthy and adaptable apparently. I think the majority of Gotham's problems could've solved with smart investments in social programs and industry and the police force which Bruce has money for. I swear the guy is anonymous being Batman, which means he can get away with killing, which would solve all his super villian issues immediately. How naive are superheros when same villains keep coming back. Considering his track record, you think he really can't his stance on the not killing rule. And even when it comes to I dunno fighting regular crime, I'm pretty sure a neighborhood watch group and 100 patrolmen would do a much better job than one guy flying around million dollar super planes. I think that just makes you an asshole. I'm done: You guys gave me more reasons to hate Batman more and superheros in general.	
Reply A (Δ)	**Reply B**
I think the majority of Gotham's problems could've solved with smart investments in social programs and industry and the police	He's intelligent and wealthy and adaptable apparently. I think the majority of Gotham's problems could've solved with smart

force which Bruce has money for. [+ 229 words]	investments in social programs and industry and the police force which Bruce has money for. [+ 333 words]
Predict-then-Explain Response (Incorrect)	
`I believe reply B is more persuasive and more likely to change the original poster's view. Reply B provides more specific examples and explanations to counter the original poster's arguments. It addresses each point made by the original poster and provides logical reasoning as to why Batman's actions and methods are justified. Reply B also brings up valid points about the` **`limitations of regular law enforcement and the unique abilities and resources that Batman possesses.`** `Overall, reply B presents a more thorough and convincing argument.`	
Explain-then-Predict Response (Correct)	

`Reply A is more persuasive because it provides more detailed explanations and examples to support its points. It addresses the idea that` **`Bruce Wayne already invests in social programs and the justice system is failing Gotham by not executing repeat offenders.`** `It also explains why` **`Batman's physical abilities and determination make him more effective than a neighborhood watch group or patrolmen.`** `Reply B also makes valid points, but Reply A goes into more depth and provides a more thorough analysis of the situation.`
 	
`Choice: A`	

Figure 5: Example of difference in GPT response between prompts.

4.4. Feature Analysis

To understand what drives GPT-3.5 Turbo’s decision-making and correctness, we analyzed surface-level features of replies under the PTE and ETP prompts with temperature = 0. We examined two types of correlations:

- Differences between Reply A and B, correlated with whether GPT chose Reply A (a_minus_b features)
- Differences between the chosen and unchosen replies, correlated with whether GPT was correct (delta features).

As shown in Figure 6, under the PTE prompt, GPT shows a mild preference for replies with more characters (+0.17) and more words (+0.15), as well as longer average word length (+0.16). These patterns suggest a bias toward longer or denser replies, even if such choices aren’t always correct. Other features, like hedging language or the number of questions, showed minimal influence on the model’s choices.

When GPT is correct, the chosen reply tends to differ more from the unchosen reply in three key ways: more bullet points (+0.24), more characters (+0.24), and more words (+0.23). These correlations are notably stronger than those that influence GPT's choice, especially the number of bullet points. This indicates that formatting and length are more predictive of persuasiveness than GPT accounts for. In contrast, hedging, links, and questions continue to show weak associations with correctness.

As shown in Figure 7, under the ETP prompt, GPT displays a strong preference for longer and more detailed replies. Particularly, character count (+0.41) and word count (+0.39) have the strongest positive correlations. GPT also favors replies with more bullet points (+0.28) and longer words (+0.23).

When GPT is correct, the features of the chosen reply are similar to that of PTE; the number of bullets (+0.21), characters (+0.20) and words (+0.19) are most predictive of correctness.

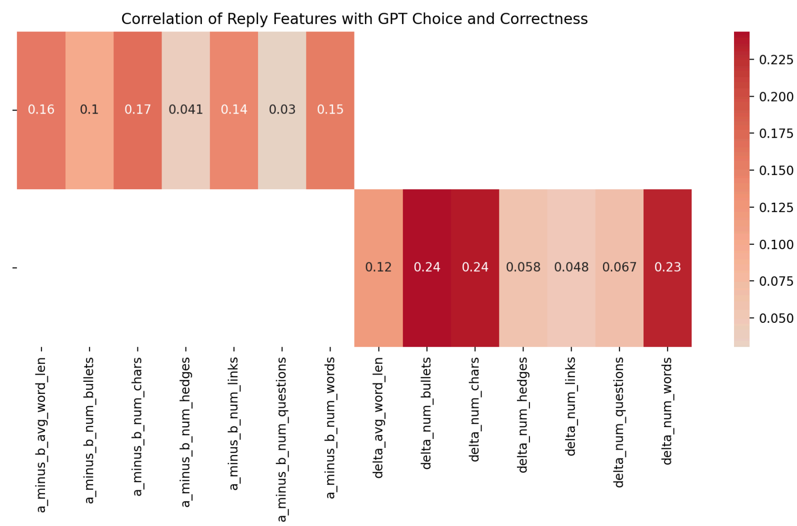

Figure 6. Correlation of reply features with GPT choice and correctness for PTE.

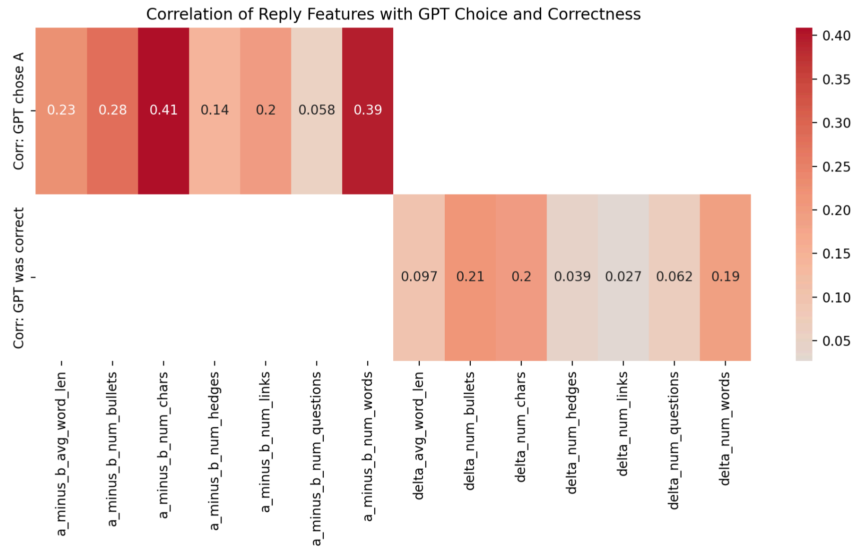

Figure 7. Correlation of reply features with GPT choice and correctness for ETP.

5 Discussion

Our results show that prompt structure significantly affects GPT-3.5 Turbo's ability to predict persuasive arguments. The ETP strategy outperforms the PTE strategy at lower temperature settings where the model operates more deterministically, while it is more comparable for a higher temperature setting.

Deeper analysis of reply features suggests that GPT's decisions are guided in part by surface-level heuristics, such as reply length and formatting. In both prompting styles, GPT shows a strong preference for longer replies with more words and characters, especially under ETP, where correlations with reply length exceed +0.40. Yet when correlating those same features with correctness, the associations are notably weaker. This shows that GPT's confidence often relies on surface features rather than substantive reasoning, and its preferred features only partially align with what actually makes a reply persuasive to humans.

The number of bullet points was one of the strongest predictors of correctness in both prompting conditions. However, GPT appears to underweight this feature in making its choices. Therefore, this raises the opportunity for prompt re-design or model finetuning that explicitly emphasizes structural cues.

Additionally, the narrowing of performance differences at higher temperatures suggests that randomness can mitigate overconfidence or positional biases.

5.1. Limitations and Future Directions

To ensure rigorous and consistent decisions, future experiments could use the explanation itself to generate a score (e.g., using LLMs or classifiers) and then make a decision based on the explanation's quality.

Because GPT-3.5 Turbo is now succeeded by newer models like **GPT-4.1**, these models are likely to have higher accuracy (a sample of 50 at temperature = 0 yielded 66% accuracy for PTE). However, due to the exponentially higher cost, we were unable to use a larger sample of 500.

Rather than prompting GPT to make a binary A/B choice, future setups could use pairwise preference prompts like "Why is A more persuasive than B?" to force comparative reasoning.

Even with randomized reply order, GPT consistently favors Reply A. Future designs could explicitly instruct GPT to ignore ordering.

A more granular setup could break down argument quality into dimensions such as relevance, clarity, structure, or emotional appeal, and assess which of these GPT picks up on.

Finally, training LLaMa on the Anthropic/persuasion dataset is a fruitful possibility that was omitted due to a lack of resources.

6 Conclusion

This study examined whether GPT-3.5 Turbo can predict which of two replies in a dialogue is more persuasive, using data from the ChangeMyView (CMV) dataset. We found that prompting structure significantly impacts performance: prompting GPT to explain before predicting (ETP) led to a higher accuracy than predicting first, especially in deterministic settings (temperature = 0).

Our feature analysis revealed that GPT relies heavily on surface-level heuristics, such as reply length and structure, when selecting persuasive arguments. While these cues sometimes align with human judgments, they are not always indicative of actual persuasiveness, which limits the model's generalization. This misalignment is especially notable in the PTE setting, where GPT often overcommits to shallow signals. Interestingly, introducing temperature (0.5) helps mitigate rigid biases, particularly in PTE, likely by increasing variability in response selection.

Although higher temperature settings helped mitigate overconfidence and improved performance in PTE, our findings highlight ongoing limitations in zero-shot LLM judgment. More advanced models such as GPT-4.1 may yield improved results, but were too costly to evaluate at scale.

Future work should explore alternative prompting formats, including pairwise comparisons and explanation-based scoring. More granular rhetorical analysis and training open models like LLaMa on persuasion-focused datasets are promising next steps for advancing LLM alignment with human evaluative reasoning.

References

[1] T, Chenhao, V, Niculae, C, Danescu-Niculescu-Mizil, L, Lee. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions, 2016, chenhaot.com/pubs/winning-arguments.pdf.