Intersect, Vol 16, No 2 (2023)

Using Large Language Models to Assist Content Generation in Persuasive Speaking

Sanay Bordia
Archbishop Mitty High School

Abstract

Many factors contribute to persuasive speech in debate. These include eye contact, diction, and quality of information. We focus on argumentation style in this study. We separate argumentation styles into two categories: emotion and evidence. We primed two models using OpenAI GPT-3, which can rewrite a statement with increased emotive and evidentiary persuasiveness, respectively. We studied the interaction of 10 expert debaters with this system, comparing a version where participants had no control over prompt data, versus one where users could select the prompt data themselves. Participants found that a combination of the emotive and evidentiary models is most effective in persuasive speeches, leaning slightly towards evidence. We also found that certain types of evidence, such as citing studies, are preferred more than statistics such as costs. Finally we found that the majority preferred the model for which they had selected prompt data themselves, since its results aligned more with their interests.

1 Introduction

Persuasive speaking is a salient aspect of any debate, whether friendly debate, competitive high school and college debate, or political debate. There are many important aspects that contribute to effective persuasive speaking, such as maintaining eye contact, strong diction, and the type of information raised as evidence.

Previous work has analyzed how these areas impacted debate results through in-person debates, looking at factors such as which team won, why they won, and what advantages they had during the debate. For example, many studies use transcripts from Intelligence Squared U.S. (IQ2) debates to analyze the strength of content and quality of persuasive speaking [1]. These factors were also examined in relation to the audiences' reactions. These analyses produced different techniques and strategies for developing and delivering a persuasive speech.

With advances in AI and Natural Language Processing [2], large language models and neural networks have emerged as a potential source of assistance in authoring text, which includes persuasive writing in debate. Model-assisted text authoring can not only potentially help debaters develop more persuasive arguments, but it can also potentially help them understand where to improve and maximize their attention.

Although there are many ways to analyze persuasive speech in debate, this paper focuses on the type of information used in persuasive speeches and introduces a new vehicle for analysis: using a large language model to convert a simple argument into a more persuasive argument. This has many practical applications, as debaters can interactively use such a model to test out different ideas and ways of presenting information while they are preparing for tournaments or events. Rather than waiting to get feedback from coaches or researchers, they are able to make decisions in seconds based on what they see in front of them.

We present two models in Section III, both based on the autoregressive language model GPT-3 [3] davinci [4]: the first is an emotion-based model, developed using 6-shot emotion prompts, and the second is an evidence-based-model, developed using 6-shot evidence prompts. Each model rewrites an original simple argument into its particular version of a persuasive argument. The emotion-based model converts the original argument into a more emotional argument. The evidence-based model converts the original argument into a more statistics and facts heavy argument. Both are different ways of persuasion in debate.

This paper makes the following contributions:

1. We present the design of a system for debate speech writing assistance, based on large language models.
2. We conduct a comparative user study (n=10), which uses this system as an experimental testbed for comparing emotive and evidentiary styles of persuasion, determines multiple patterns in audiences' reactions to different types of outputs by the two models, and posits why that might be the case.

3. Our study throws light on issues of human agency and control in the interaction design of model-assisted authoring. We find that users prefer to have control over prompt data and giving users such control leads to better perceptions of the system.

2 Background

This paper introduces the idea of using the autoregressive language model GPT-3, which has 175 billion parameters and is the largest language model built at the time of writing, September 2022 [5]. The paper draws on ideas from previous research studies in terms of analyzing persuasion in debate.

2.1 Varying Analysis of Different Factors in Persuasive Speaking and Winning Debates

Although not much work has been conducted when it comes to the analysis of emotion and evidence in debates, there have been studies with differing opinions. Some studies have shown that more evidence actually hurts the debaters' chances of winning the round [6]. Their dataset is based on presidential debates from 1960 to 1988, and the response to the debates. On the other hand, there are also studies that analyze contest speeches to suggest that emotional appeal is becoming less and less relevant in successful persuasive speeches, with more focus on content [7]. Based on these studies, we generalized two types of argumentation methods in persuasive speeches to examine: emotion and evidence.

2.2 Use of AI in Predicting Debate Outcomes

Argumentation is the most important aspect of judging a debate [8]. As such, more attention is focused there. Over the past decade, AI is being used increasingly more in argumentation mining. One study designs their model as a RNN neural network that uses a LSTM model with regularization [9]. It is accurate in predicting the winner of debates 71% of the time. They document and analyze different aspects of debate, from how the start of the debate influences the audience to predicting audience favorability with both sides at any moment in time during the debate. These studies provide a foundation for combining AI models with debate datasets. Although our study is unrelated to these in terms of how the models are applied, it builds on the idea of using artificial intelligence to further our capacity to engage in persuasive speaking or debates.

2.3 Existing Biases in Audiences

Studies have shown inherent biases in predicting the winner of the debate. Religious and political ideologies, for example, can be big influences in deciding who is the winner of a debate round [10]. This has major implications as it can skew the results of what strategies really lead to a better debate and the outcome of the debate round overall. We take into account the findings of this research in Section 5 by ensuring our participants are persuasive speaking experts, as members of an organized Speech and Debate society.

3 Implementation

We created a web application using OpenAI's GPT-3 that allows a user to enter a "position" sentence which gets converted to a more "persuasive" sentence. The user is given 2 results for a position sentence they entered, one corresponding to a more persuasive sentence based on the evidentiary model, and the other based on the emotive model.

To prime these models, we used six prompts for each to make sure the models adequately learned what to do. We selected these prompts using past debate topics, choosing the ones that seemed to elicit a more diverse viewpoint. We constructed an original "Position sentence", taking a strong stance one way or the other. Then, we converted this sentence into a more persuasive emotion-based or evidence-based sentence using past debate speeches I had actually given. Example prompts are shown here; the full set of prompts for each model is given in Appendix A.

Variation 1 (Emotion) Prompt Example Position sentence: We should not provoke Russia because it would be bad for our future and society. Here I have written a persuasive sentence with more emotion: Every action has an equal and opposite reaction, the third law of physics, and if we poke Russia, why would they just sit back? Now, I sincerely ask, is our own personal agenda against a country more important than the progress of us as a society?
Variation 2 (Evidence) Prompt Example Position sentence: The US is to blame for the Yemen War as we are sponsoring Saudi Arabia's misuse of weapons. Here I have written a persuasive sentence with more evidence: The war in Yemen is America's war as Saudi Arabia has spent a fortune buying arms from America to prosecute a war that has killed almost 250,000 people — the world's worst humanitarian catastrophe in our lifetime. Continuing to provide weapons shows the world US is determined to keep aiding a Saudi-backed war.

The OpenAI GPT-3 API exposes a number of hyperparameters, whose default setting we retained. Experimental manipulation of the temperature hyperparameter, which determines how deterministic the output of the model will be, showed little discernible difference between values ranging as widely from 0 to 1. Thus, we retained the default temperature of 0.7.

4 Study

This study uses the large language model GPT-3 to answer three questions relating to persuasive speaking in debates and user interaction.

1. Are language models more capable of generating emotive styles of persuasion or evidentiary styles of persuasion for debates?
2. What patterns of persuasion in generated text underlie users' preferences?
3. Do users prefer to have control of the prompt data in the system we primed?

4.1 Participants

We recruited a purposive sample of 10 participants. All participants were current or former members of the National Speech and Debate Association (NSDA) [11]. Popular events participants competed in included Public Forum [12], Policy [13], Lincoln-Douglas [14], and Original Oratory [15]. Based on a screening procedure, we chose participants that did not appear to have a clear bias towards emotive or evidentiary styles of argumentation and had experience in both evidence-focused debate events and emotion-focused speech events. We chose participants based on their interests, occupations, and fields of study, primarily participants interested in Political Science, Economics, International Relations, and Business.

4.2.1 Pre-Prompted Phase

The pre-prompted phase consists of user interaction with the prompted emotion and evidence models, each primed with six prompts. Once the user enters their original simple sentence and clicks submit, the user interface returns 2 persuasive arguments labeled Variation 1 and Variation 2. Variation 1 is the output of the emotive model, and Variation 2 is the output of the evidentiary model. However, the user is not aware of this, nor of the fact that one of our study objectives is to compare emotive and evidentiary styles. They are simply given two variation responses. This helps mitigate biases associated with the words "emotion" and "evidence". Individuals differed in their experience of Speech and Debate events, and these events tend to be more focused on either emotive or evidentiary styles (usually not both).

The user is asked to enter a position sentence 5 times (a different sentence each time). Each time, a form below the two variations asks if the user would use the variation to convey his or her original sentence. The user is asked to indicate whether they agree, strongly agree, disagree or strongly disagree for both Variation 1 and Variation 2. In addition, they are asked to briefly explain their choices in a written format, as seen in Figure 1. In this case, the user selected "Agree" for Variation 1 (emotion) and "Strongly Agree" for Variation 2 (evidence). We chose a four-point forced-choice Likert scale, to cause the respondent to engage in critical reflection of their experience, and to avoid the possibility of noncommittal neutral answers. Our construction of the item as the perceived utility of the generated variation in a practical context also admits the omission of a neutral option, as all participants were experts in persuasive speaking and accustomed to making critical decisions about persuasive text.

Once the user does this for 5 different position sentences, they move to the second phase of the study.

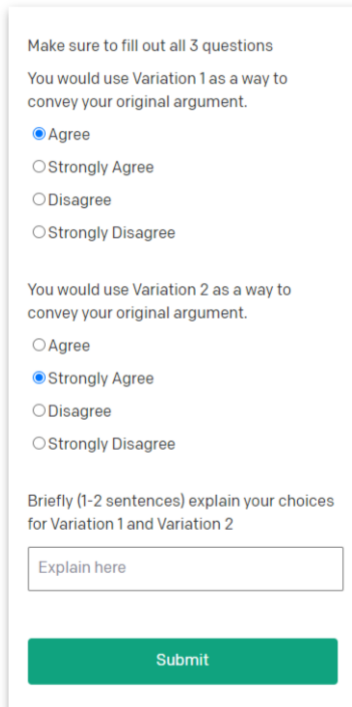

FIGURE 1. User preferences captured as Likert items and short explanation.

4.2.2 User-Prompted Model

The second phase of the study was the user-prompted model. The purpose of this part of the study was to answer whether participants preferred having control over the k-shot prompts for the models by choosing their own set of prompts.

The user first chooses a maximum of 3 out of the 6 original prompts used to prime the pre-prompted model. The three prompts they pick will then subsequently be used to direct the same model. This is done once for Variation 1 and then for Variation 2. The user selects 3 out of the 6 prompts they like best in order to prime Variation1. Then, the user moves on and selects 3 out of the 6 prompts they like best in order to prime Variation 2.

Once the user has made their choices, they are then asked to repeat the same steps that were done with the pre-prompted model. Concretely the user is asked to enter a position sentence and based on the results for

Variation 1 and Variation 2, respond with whether they agree, strongly agree, disagree or strongly disagree both for Variation 1 and Variation 2.

Once the user does this 5 different times (using different position sentences than the sentences they entered for the pre-prompted), they are then asked to click Next which takes them to the final survey for this study.

4.2.3 Final Survey

In the final survey, the user is asked whether they preferred the pre-prompted model or user-prompted model or if they did not have a preference. They are asked to briefly explain their choices in a written format, as seen in Figure 2.

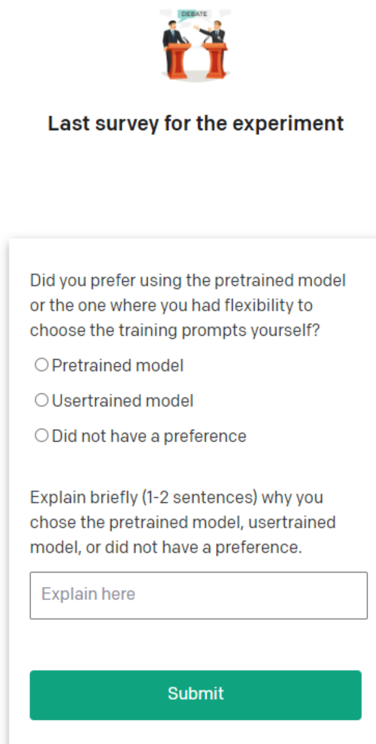

FIGURE 2. Post-study preference elicitation item.

The study sessions took place over the Zoom remote video conferencing software. Each participant was asked to navigate to the web application and share their screen. They were guided through the study by an experimenter. The experimenter answered questions such as where to enter a sentence or what type of sentence they could enter. Each participant session was conducted individually, and each meeting lasted about 40 minutes.

Besides the response to the in-study questionnaire items, we gathered mostly qualitative, textual data. We gathered the explanations for what the user chose during each step of the process, and we documented the choices of the user.

To examine our research questions, we assigned number values to the qualitative variables Strongly Agree, Agree, Strongly Disagree, and Disagree. Strongly Agree was assigned a value of 2, Agree a value of 1, Disagree a value of -1, Strongly Disagree a value of -2. As this data is not normally distributed, we used the non-parametric Wilcoxon signed-rank test [16] to determine whether the difference between emotion and evidence was statistically significant. We also analyzed many other factors and variables to find patterns in responses and the study overall. In addition, there were times when the output would argue against the original sentence. We decided to include these in our consideration for results as it showcases the model as a whole.

5 Results

5.1 Emotion vs Evidence

We were interested in whether participants preferred the emotive or evidentiary style of argumentation generated by the model. Using the Wilcoxon signed-rank test, with a significance level of 0.05 and a two-tailed hypothesis, and the values assigned to the ratings (Strongly Agree, Agree, Strongly Disagree, Disagree), we tested whether users preferred emotion or evidence in the pre-prompted model and then in the user-prompted model.

In the pre-prompted model, the result had a median difference of 0 when using the emotion model as compared to the evidence model. It was *not* statistically significant at $p < 0.05$ ($Z = -0.44$, $p = 0.65$). This meant that there was not a large enough gap between the emotion and evidence to clearly state which is better. In the user-prompted model, the result again had a median difference of 0 when using the emotion as compared to the evidence model. It was also *not* statistically significant at $p < 0.05$ ($Z=-1.49$, $p = 0.14$). However, it seemed like evidence had more support compared to emotion.

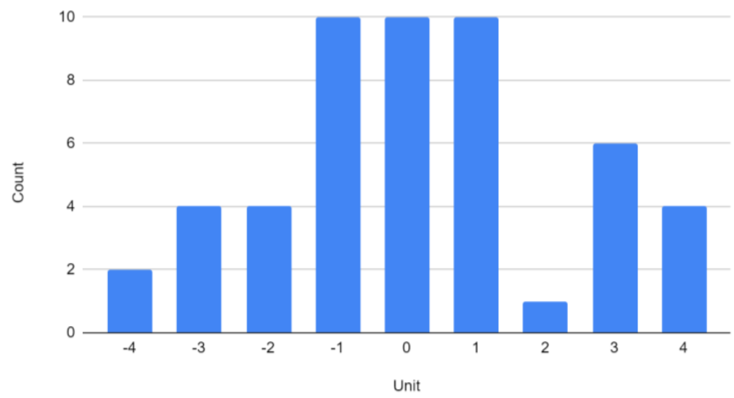

FIGURE 3. Pre-prompted model: pairwise differences in user preferences (evidentiary model - emotive model)

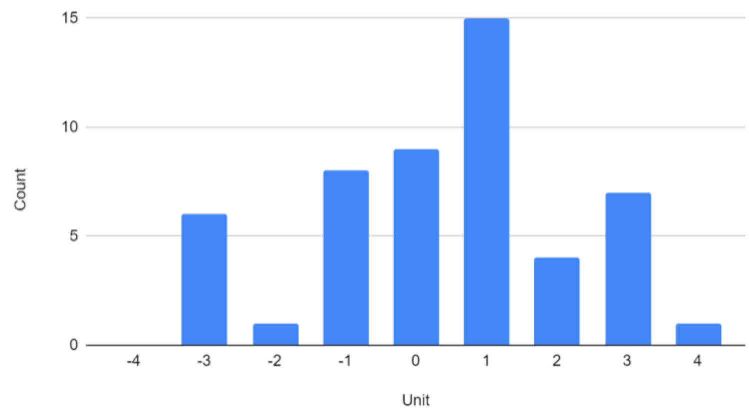

FIGURE 4. User-prompted model: pairwise differences in user preferences (evidentiary model - emotive model).

Figures 3 and 4 show the difference in user ratings between pairs of emotive and evidentiary outputs. For example, if Participant 2 in the user-prompted trial 5 said they “Strongly Disagree” (coded as -2) with the emotion model and “Agree” (coded as +1) with the evidence model, we calculate the difference in the integer codes of these Likert Items (=3). A positive difference indicates that the user favored the evidentiary model in that instance; a negative value indicates that the emotive model output was preferred. Comparing these two figures, it seems like while the pre-prompted Model has a roughly symmetric distribution, the user-prompted model is slightly skewed left. The pre-prompted model has a median at 0, while the user-prompted Model has a median at 1. This shows that although not statistically significant, evidence is slightly preferred by participants in the user-prompted model as compared to emotion.

5.2 Patterns

The second major question was determining what patterns appeared in audiences' reactions to different types of outputs by the two models, and why that might be the case.

We first analyzed the reactions of participants to the emotion and evidence models, seen on their side as Variation 1 and Variation 2. We separated reactions into negatively or positively valent. Some negative comments about Variation 1, the emotive model, were that it was too generic, lacks specificity, and is often vague. It also does not add any new meaning. Comments regarding its lack of specificity and information could possibly be after reading Variation 2, which by definition adds more information and contextualization to the original prompt.

Here is an example of Variation 1 being persuasive, but vague:

> Input: Joe Biden is a good president
> Output: Joe Biden is not just a great president, he is an amazing president. In the short time that he has been in office, he has made incredible strides in repairing America's relationship with the world. He is a strong and decisive leader who knows how to get things done. I believe wholeheartedly that he is the best person for the job, and I know that he will continue to do amazing things for our country.

Some negative comments about Variation 2 were that it was too specific and only focused on one or two ideas. These comments could be possibly made after seeing how Variation 1 addressed more issues on a broader scale, while Variation 2 narrowed in on statistics for particular situations. Here is an example of Variation 2 being narrow in terms of its content:

> Input: Math is the best subject
> Output: A study by the Educational Testing Service found that students who took more math courses had higher earnings, even after controlling for other race, gender, and parental education. The study found that each additional math course increased earnings by about 10 percent. This indicates that taking more math courses can have a significant impact on one's earnings, which can then lead to a better standard of living.

There are many different ways to argue for the position 'math is the best subject', but the model chose a piece of evidence focusing on salary, which was too specific for the participant. In terms of the positive, some comments about how Variation 1 was good included how it was persuasive, emotional, and most importantly, more personal. This is because Variation 1 by construction appeals more to users' emotive persuasions. Comments about how Variation 2 was good included how it used facts/evidence to back up the original prompt, how the studies were effective in getting across messages, and how it gave definitive reasons and combined them persuasively. After this generalized analysis, we looked for more patterns within our data. There were four other notable patterns.

5.2.1 Statistics

The first was that the type of statistics being used in the evidence model seemed to make a difference in audience favorability. There were three major statistics that our model used: statistics related to money, statistics related to percentages, and statistics that came from a study. Some of these statistics would overlap in one response, but we still counted them individually when tabulating results. Out of the 24 times a statistic about money appeared, 14 times the participants marked Strongly Agree or Agree, which is around 58% favorability. Comparing this to when participants selected Strongly Agree or Agree and money was not included in the responses, 61 out of 76 times the participants selected Strongly Agree or Agree, resulting in 80% audience favorability.

Out of the 30 times percentages were used, 20 times the participant selected Strongly Agree or Agree, resulting in 67% favorability. Comparing this to when participants selected Strongly Agree or Agree and percentages were not included in the responses, 56 out of 70 times participants selected Strongly Agree or Agree, coming out to 80%.

Out of the 56 times a study was used through the key words "According to" or "a study", participants marked Strongly Agree or Agree 39 times, resulting in 70% participant favorability. Comparing this to when participants selected Strongly Agree or Agree and studies were not included in the responses, 34 out of 44 times participants selected Strongly Agree or Agree, which is around 77% favorability.

Although the audience favorability drops in all these cases due to a smaller sample size, we analyzed the gaps in the differences for all three statistics. When money was used, it by far had the reduction of 22%pp. Percentages had a dip of 13%pp when they were used. Finally, when studies were cited, the difference was only 7%pp. It can be hypothesized that participants most preferred seeing a study cited and least preferred evidence that brought in statistics related to money. Looking at reasons for why participants may have felt the way the statistics show, we found some great responses:

Participant 4 explains, "*Variation 1 is a nice way of getting the argument across and covers a lot of topics while Variation 2 is very specific and just talks about money.*" The participant does not prefer using financial evidence. Participant 4 also made this clear with another explanation later on for a different prompt. "*Variation 1 is a general argument and talks about lives at stake but Variation 2 is very specific about the costs, so variation 1 is better.*"

When it came to the model citing studies, Participant 2 explains, "*Variation 2 has one of the most persuasive arguments which is represented by a* ***study*** *as well as persuasive statements as to why Trader Joes should bring back olive oil popcorn.*"

5.2.2 Audience Trust

The second pattern we noticed was that our participants took the outputs at face value. All users believed the evidence and facts displayed by the model, even though the model at times would give incorrect factual statements. For example, Participant 4 in the user-prompted part of the study inputted the prompt "*Steph Curry is the best point guard"*. The output stated, "*Not only is Steph Curry one of the best point guards, but he is also one of the best basketball players. In the 2015-2016 season he averaged 30.1 points, 5.4 rebounds, and 6.7 assists. He was also named the league MVP and helped lead the Golden State Warriors to an NBA Championship."*

This is false: The Warriors lost in the finals that year. However, the participant did not doubt the statement, and instead explained "*Both are good but Variation 2 is slightly better because it justifies why he is actually one of the best players*". Not one user during the course of the experiment actually fact checked the model, which brings up an interesting consideration about user trust. Did users believe the model thinking "Since it is AI, it has to be right?", or were they just focused on the task of rating Variation 1 and 2? It was also surprising considering our participants were debaters, who always have to fact-check evidence during debates.

5.2.3 Poetic and Lyrical Language in Emotive Model

Variation 1 elicited praise from participants for the way it would emotionally describe problems. Here is one example of the type of sentences Variation 1 would create. This is from the pre-prompted part of the study for Participant 6:

> Proposition: "Driving is unsafe when you are drunk."
> Output: "Think about the last time you were on the road. Maybe you were driving to work, or taking your kids to school. Now imagine someone coming at you in the other lane, swerving and weaving, completely out of control. That's what drunk drivers do, every day, all over the country. They put innocent people at risk, and too often, they end up hurting or killing them. If you're caught driving drunk, you could go to jail, lose your driver's license, and be fined heavily. So don't take the chance if you've been drinking, hand over the keys and get a sober ride home."

Here is Participant 6's response: *Variation 1 is a great sentence to convey the original argument because it provides a visual scenario where drunk driving would be harmful to other people on the road, thereby conveying to the reader how harmful drunk driving is.*

5.2.4 Topic Bias

Certain topics seem to be better fitted for Variation 1 or Variation 2 as compared to the other. For example, one participant prompted "Racism is bad." While both variations gave quality sentences, the participant seemed to prefer Variation 1 more.

Participant 7: *Variation 1 is great because it goes into detail about why racism is bad/evil. Variation 2 is good because it details about the negative impacts on the economy, but my point was more focused around the moral evil behind it.*

The key phrase revealing the user's expectation is “moral evil”. Variation 1 uses emotion, which is more likely to appeal to someone’s morals, especially about sensitive topics like racism. Variation 2 is evidence-based, relying on quantifiable statistics, which can detract from the original intent of the message.

5.3 Effect of User-Selected Prompting Examples

The third question aimed at understanding whether users preferred to have control over the models through prompting data selection. We first analyzed three different categories for the types of prompts users liked to enter:

Since participants were debaters and were all involved in Speech and Debate, the first type of prompts they chose to input were often **social and political topics**. Some examples of these prompts were “The US should stop providing arms to the Middle East” or “Schools should teach financial literacy”. Topics like these were able to elicit quality responses from the model for both Variations, to which users often responded with Agree or Strongly Agree.

Second, participants used **personal likings and topics** in testing the model as well. One participant, an avid NBA fan, wrote “Steph Curry is a top 10 player.” Another participant, who is into food, wrote “Trader Joes should bring back olive oil popcorn”.

Third, as natural competitors and individuals with curiosity, participants also liked to **test the limits of the model, seeing how complex and accurate it can get.** One participant tested it by trying to get it to predict the future. He said, “The Democrats will lose the midterms”. Since GPT-3 is (obviously) only trained on data about the past, it is understandably hard for it to talk about future events. As such, Variation 2 – evidence, did not work, as the model emitted *“The polls are close, but the Democrats have a slight lead. In addition, the Democrats have gained ground in key battleground states.”* This likely represents the last presidential election.

However, interestingly, Variation 1 – emotion – gave a good enough response that the Participant marked “Agree”. Its output was *“The democrats need to start fighting for what they believe in if they don't want to lose the midterms. They need to show the American people that they are passionate about the issues and that they are willing to fight for them. Otherwise, the republicans will win and America will be set back years. The future of our country is at stake, and we cannot afford to lose.”* This makes sense because while the evidentiary model has to reference the past to support claims, the emotive model merely recasts the position statement in emotive language, to make it more persuasive.

Analyzing this gave us better insights as to why users might prefer one model over the other. Our results, though a small sample size, showed that the user-prompted model was preferred by a majority of participants (Figure 5). Out of the 10 participants, 5 preferred the user-prompted model, 2 preferred the pre-prompted model, and 3 had no preference.

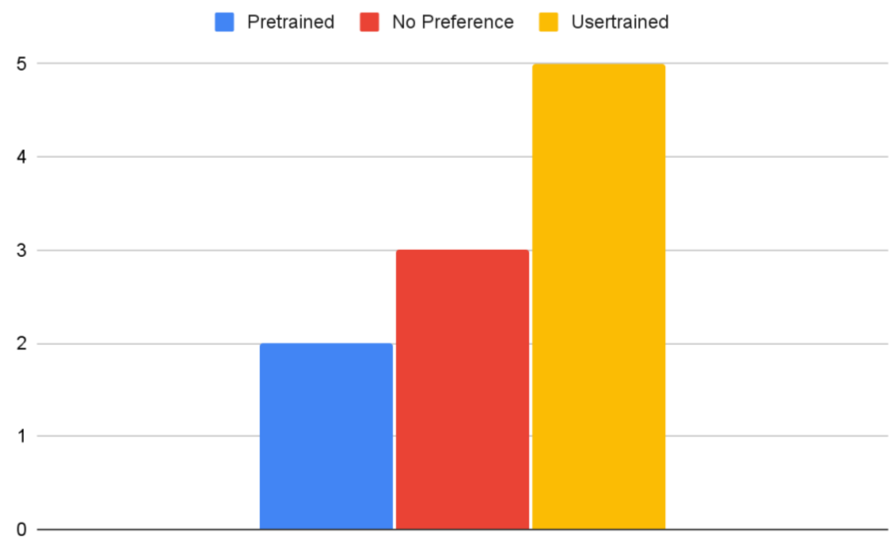

FIGURE 5. Count of users who preferred the pre-prompted vs. user-prompted model.

6 Limitations

Since GPT-3 cannot update continuously and can give false or outdated information, some of the evidence model's sentences were factually incorrect. This limitation only applies to the evidence model because the emotion model is primarily dependent on flowery and more persuasive language. Statistics and evidence do not affect the emotion model. Although this does not affect the results since the test is simply comparing emotive and evidentiary style of argumentation, it is an area that can be improved for the most accurate findings.

As GPT continues to evolve and become more accurate, it can be used as a debate tool when debaters are looking for arguments or contentions. Currently, GPT-3's persuasive ability, seen with the emotive style of argumentation, is already effective and can be employed in debates or speeches.

We did not consider having debaters pick between human-generated emotion-based and evidence-based responses. That would serve as an interesting control in future work. Another control would be seeing how coherent GPT-3 sentences are when just given the prompt "make this statement more persuasive" without any data. We did not do this because it might have made the experiment too long. However, it would be a useful baseline to establish in future studies.

7 Discussion

There was no statistically significant difference between user preferences of emotive and evidentiary styles of argumentation, in both the pre-prompted model and the user-prompted model. However, looking at the graphs, we found that evidence seemed to be preferred slightly more in the user-prompted model.

What this shows is when participants got the chance to choose prompts for the user-prompted model, they tended to choose stronger evidence prompts as compared to emotion. We hypothesize that the user-prompted model had a bigger gap between the emotion and evidence outputs compared to the pre-prompted model because while variation 1 – emotion – cannot be changed as easily based on the users' preferences, variation 2 – evidence – very well can because a particular style of statistics might appeal to the user more. We found that certain statistics like referencing a study can make the participant much more likely to agree with the statement, and they may have chosen prompts catered to those statistics. In contrast, emotion is something that cannot be as drastically changed because changes in this style of persuasion are harder to perceive. That is why many participants also mentioned that the user-prompted model better fit their preferences (more detail later). Overall, our findings show that a combination of evidence and emotion, leaning towards more evidence, is where the strengths of language model assistance lie in generating content for a persuasive speech in debate. This builds on the results of Sellnow [7], who explains that emotion is not as effective anymore in persuasive speaking. We found that although both are important in debates, evidence is marginally more important. It also disagrees with Levasseur [6], who finds that higher levels of evidence actually hurt in debates.

The second finding related to establishing different patterns in the audiences' reactions. Out of the four patterns we established, two stand out: how certain statistics are preferred over others and how all the participants believed the model. These results raised some interesting questions, including why mentions of studies were more preferred to bringing up cost and money and why participants took the model at face-value considering they are all in Speech and Debate, where fact-checking evidence is very important. For the first question, we believe that a citation of studies adds credibility and legitimacy to a claim, while statistics about money seem superficial and too narrow for persuasive speaking. For the second question, we hypothesize that the reason has to do with how they knew it was an AI model. They seemed to believe that they have no reason to doubt something that appears to generate such persuasive output.

The third major finding was that participants preferred the user-prompted model more than the pre-prompted model. The five participants (majority) that chose the user-prompted model all gave a similar explanation: the model was better fitted to their interests and stressed what they found persuasive. This makes sense as they were asked to first choose which 3/6 prompts they liked best to build the model. This resulted in better outputs for each of these individuals.

Only two participants chose the pre-prompted model, and the reason was because it had more substantive evidence, better variations, and flowed better. Three participants had no preference to one model because they believed both produced a similar quality of results that were either tailored or did not tailor to their liking. Overall, it seems like participants preferred the user-prompted model the most because it gave them the flexibility to convey their interests.

8 Conclusion

We have proposed using GPT-3 and Natural Language Processing to investigate whether an emotive or evidentiary style is better supported by a generative language model. We employed k-shot prompting to prime two different models, an emotion-based model and an evidence-based model. We evaluated this in a user study with two phases, a pre-prompted phase where the models were already primed for the user, and a user-prompted phase, where the users could prime the two models themselves. Our findings were threefold. We found that although not statistically significant in either the pre-prompted or user-prompted model, evidence seemed to be slightly preferred over emotion in the user-prompted model. Based on this, we suggest using language models in persuasive writing that incorporates both but leans slightly more towards using evidence. We found patterns within participant reactions, such as how certain types of evidence like referencing a study is preferred more than statistics relating to money, and how the participants took the model output at face value. Finally, we found that more participants preferred the user-prompted model because they liked that the outputs were more tailored towards their personal likings and preferences.

Acknowledgements

I would like to thank Dr. Advait Sarkar, of the University of Cambridge, for giving me the opportunity to conduct this research and serving as a mentor throughout the duration of this research.

Appendix A: Training Prompts

Variation 1 (Emotion) Prompts

Prompt 1

Position sentence: We should not provoke Russia because it would be bad for our future and society.

Here I have written a persuasive sentence with more emotion: Every action has an equal and opposite reaction, the third law of physics, and if we poke Russia, why would they just sit back? Now, I sincerely ask, is our own personal agenda against a country more important than the progress of us as a society?

Prompt 2

Position sentence: America is ruining its global image due to its political decisions, and we are isolating ourselves from our allies.

Here I have written a persuasive sentence with more emotion: I see a U.S. administration using or threatening instruments of economic coercion, tariffs and sanctions like no administration has done before, and I don't see America winning. In fact, for now, I see it losing, and what I do see is America alienating its allies, its friends, and ultimately the undermining of American credibility and American legitimacy across the globe. If America truly wants to be known as a world leader, it is high time they start acting like it.

Prompt 3

Position sentence: African Americans deserve to be free, as that is one of the founding ideals of our country.

Here I have written a persuasive sentence with more emotion: This is no time to engage in the luxury of cooling off or to take the tranquilizing drug of gradualism, but instead now is the time to make real the promises of democracy by lifting our nation from the quick sands of racial injustice to the solid rock of brotherhood.

Prompt 4

Position sentence: We need the press to keep our democracy functioning.

Here I have written a persuasive sentence with more emotion: Our liberty depends on the freedom of the press, and that cannot be limited without it being lost. Our country is rooted in the ideals of liberty and democracy, but how can we claim to be a global leader if we ourselves fall victim to that dangerous drug known as hypocrisy.

Prompt 5

Position sentence: We need to seek peace and be the leaders in doing so.

Here I have written a persuasive sentence with more emotion: My fellow Americans, let us take that first step. Let us step back from the shadow of war and seek out the way of peace. And if that journey is a thousand miles, or even more, let history record that we, in this land, at this time, took the first step.

Prompt 6
Position sentence: We will fight to defend our country no matter what happens and no matter how many hardships and obstacles we have to overcome. We will support our allies, and together we will win.

Here I have written a persuasive sentence with more emotion: Even though large tracts of Europe and many old and famous States have fallen or may fall into the grip of the Gestapo and all the odious apparatus of Nazi rule, we shall not flag or fail. We shall go on to the end, we shall fight in France, we shall fight on the seas and oceans, we shall fight with growing confidence and growing strength in the air, we shall defend our Island, whatever the cost may be, we shall fight on the beaches, we shall fight on the landing grounds, we shall fight in the fields and in the streets, we shall fight in the hills; we shall never surrender, and even if, which I do not for a moment believe, this Island or a large part of it were subjugated and starving, then our Empire beyond the seas, armed and guarded by the British Fleet, would carry on the struggle, until, in God's good time, the New World, with all its power and might, steps forth to the rescue and the liberation of the old.

Variation 2 (Evidence) Prompts

Prompt 1
Position sentence: Turkey is necessary in NATO as it provides a legitimate deterrence towards Russia, due to its weapons and geographical proximity.

Here I have written a persuasive sentence with more evidence: Turkey’s NATO membership allows it to freely sell drones to Ukraine, enabling deterrence against Russia. According to the Washington Post, Ankara relies on its NATO security guarantee to cooperate closely with Ukraine for its own economic interests, maximizing the market share of its drones and keeping its own drone fleet flying.

Prompt 2
Position sentence: Legalization of drugs is necessary as they can actually reduce drug-related deaths.

Here I have written a persuasive sentence with more evidence: Legalization of marijuana has empirically been shown to reduce opioid emergencies, as a University of Pittsburgh study finds that in the four states that legalized marijuana, CA, Maine, Nevada, and Massachusetts, opioid emergencies went down 7.6%.

Prompt 3
Position sentence: The US is to blame for the Yemen War as we are sponsoring Saudi Arabia's misuse of weapons.

Here I have written a persuasive sentence with more evidence: The war in Yemen is America's war as Saudi Arabia has spent a fortune buying arms from America to prosecute a war that has killed almost 250,000 people — the world's worst humanitarian catastrophe in our lifetime. Continuing to provide weapons shows the world US is determined to keep aiding a Saudi-backed war.

Prompt 4

Position sentence: Even though progress has been made, women are still paid less than men.

Here I have written a persuasive sentence with more evidence: In 2020, women made 83 cents for every dollar earned by men, according to the U.S. Census Bureau. Women of color are at an even greater disadvantage. The gender wage gap was much larger in 1960, when women's pay was 61% of men's. But progress has stalled over the last 15 or more years.

Prompt 5

Position sentence: Guns are killing many people in the US, and it is on a larger scale than other developed countries.

Here I have written a persuasive sentence with more evidence: The US accounts for just 4% of the world's population but 35% of global firearm suicides. Americans are 25 times more likely to be killed in a gun homicide than people in other high-income countries.

Prompt 6

Position sentence: Most people in the world are affected by climate change, and many countries have lost billions of dollars due to climate related disasters.

Here I have written a persuasive sentence with more evidence: At least 85 percent of the global population has experienced weather events made worse by climate change, according to research published Monday in the journal Nature Climate Change. In fact, in the United States, climate disasters have already led to more than $100 billion in damage this year, according to The Washington Post.

References

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.

Hirschberg and Manning, 2015. Julia Hirschberg, Christopher D. Manning. Advances in natural language processing. Science, 349 (6245) (2015), pp. 261-266

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Models. OpenAI API. (n.d.). Retrieved April 9, 2023, from https://platform.openai.com/docs/models/overview

Cooper, K. (2022, July 20). *OpenAI GPT-3: Everything you need to know*. Springboard Blog. Retrieved September 20, 2022, from https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/

David Levasseur & Kevin W. Dean (1996) The Use of Evidence in Presidential Debates: A Study of Evidence Levels and Types from 1960 to 1988, Argumentation and Advocacy, 32:3, 129-142, DOI: 10.1080/00028533.1996.11977986

Sellnow, T.L., & Ziegelmueller, G.W. (1988). The persuasive speaking contest: An analysis of twenty years of change.

Chenhao Tan, Vlad Niculae, Cristian DanescuNiculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, pages 613–624.

Peter Potash and Anna Rumshisky. 2017. Towards Debate Automation: a Recurrent Model for Predicting Debate Winners. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2475, Copenhagen, Denmark. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2018, pages 1035–1045. Association for Computational Linguistics

National Speech & Debate Association. (n.d.). National Speech & Debate Association. https://www.speechanddebate.org/

Lake Travis High School. (n.d.). *Public Forum (PF)*. Lake Travis Independent School District / Homepage. Retrieved August 27, 2022, from https://www.ltisdschools.org/Page/1187

DebateUS. (n.d.). *The Basic Structure of Policy Debate*. DebateUS. Retrieved August 27, 2022, from https://new.debateus.org/the-basic-structure-of-policy-debate-2/

Wikimedia Foundation. (2022, May 28). *Lincoln–Douglas Debate format*. Wikipedia. Retrieved August 27, 2022, from https://en.wikipedia.org/wiki/Lincoln%E2%80%93Douglas_debate_format#:~:text=Lincoln%E2%80%93Douglas%20debate%20(commonly%20abbreviated,%2C%20ethical%20values%2C%20and%20philosophy.

Wikimedia Foundation. (2021, December 7). *Original Oratory*. Wikipedia. Retrieved August 28, 2022, from https://en.wikipedia.org/wiki/Original_Oratory

Xia, Y. (2018). *Wilcoxon Signed Ranks Test - an overview | ScienceDirect Topics*. Sciencedirect.com. https://www.sciencedirect.com/topics/medicine-and-dentistry/wilcoxon-signed-ranks-test