Priming for Tonal Consistency in AI Text Generation

Aimee Haas

2023

Abstract

This study explores the effect of conditioning on text generation, with a particular focus on tonal consistency in artificial intelligence (AI) responses. By comparing responses from AI and a human respondent, I investigate how priming influences tonal consistency. My findings provide insights with potential applications in sectors such as customer service and education, where maintaining a consistent tone is crucial.

1 Introduction

The integration of AI into our daily lives presents new opportunities and challenges. One such challenge lies in conditioning text generation, with the aim of developing an AI model capable of maintaining a consistent tone throughout a conversation. This is a significant hurdle for AI systems.

The concept of "shot learning" in AI refers to the number of examples given to the model during learning. Understanding these learning methods is vital for grasping AI's potential in text generation.

My research focused on whether AI can match human tonal consistency, a question with implications in various sectors and ethical considerations around academic integrity. Using the advanced GPT-3.5-turbo AI model and varying priming levels, I compared AI and human responses to assess the model's efficacy.

This research goes beyond evaluating AI capabilities. It seeks to understand potential sector impacts and ethical dilemmas, aiming to contribute both to the technical understanding of AI and discussions about its integration into society.

2 Related Work

A pivotal work in the field of large language models was the demonstration of GPT-3's capabilities by Brown et al. (2020)[1]. This study established that large language models could excel in a variety of NLP tasks by using a few-shot learning approach, where the model generalizes from a small number of examples provided in the prompt. This approach fundamentally serves as a form of priming, guiding the model's generation behavior. It suggested that, given sufficient contextual cues or priming instances, these models could perform tasks effectively without explicit fine-tuning.

This idea of few-shot demonstrations acted as a rudimentary form of priming, providing initial insights into how language models could be conditioned to generate more contextually aware outputs. However, the conditioning depended on the examples provided, which had to be carefully constructed.

Brown et al.'s work serves as a significant point of reference for my research. It underpins the idea of priming as a potentially powerful tool for influencing the behavior of large language models. Yet, the challenge of conditioning a model to maintain a consistent tone across multiple text generations, akin to a human conversation, still exists. My work extends the insights provided by Brown et al., exploring the effectiveness of priming in promoting tonal consistency and considering its implications for practical applications.

3 Task

The primary task of my research is to examine whether conditioning text generation on examples primed with human-authored answers leads to more consistent tone than examples primed with AI-generated responses. I use the method of shot learning, pairing human-authored questions—emulating a typical professor-student discussion board—with both human and AI-generated answers to form the priming pairs. The same sets of questions are used in both cases to ensure consistency in context and complexity.

4 Model

For this study, I use the GPT-3.5-turbo model from OpenAI, an advanced language model. To quantify tonal consistency, I use the SentenceTransformer model with 'distilbert-base-nli-stsb-mean-tokens'. This model generates sentence embeddings that I use to measure the semantic similarity between pairs of responses, serving as my metric for tonal consistency.

5 Experiment

To investigate the potential of AI in emulating human tone in text generation, I designed and conducted a series of experiments. The data for this study were derived from real-world discussion prompts from an introductory level geography course that I took.

In order to maintain ethical integrity, I chose to use only my own responses to these prompts instead of involving other students' work. These discussion prompts, along with the respective responses, formed the human-authored example pairs used for priming the AI model. The AI-authored pairs, conversely, were generated by providing the same discussion prompts to the GPT-3.5-turbo model and saving its responses.

The experiment involved running the GPT-3.5-turbo model with varying levels of priming, ranging from no priming (0) to seven (7) priming examples, using both the human-authored and AI-authored pairs. For each priming level, the model was given the corresponding number of example pairs before being asked to generate a response to a new discussion prompt.

The responses generated by the AI model for each priming level were then saved. These responses were compared within each priming group to determine the level of tonal consistency, using BERT-based semantic similarity calculations.

The process was repeated for 20 iterations to ensure a robust sample size for each priming level. The responses from these iterations, and the corresponding similarity scores, form the crux of the results and observations detailed in the following sections.

Through this study, I sought not only to assess the capabilities of the AI model in mimicking human tone in text responses but also to explore the impacts of priming on the tone consistency in AI-generated text. The implications of these findings are discussed in the subsequent sections.

6 Results

The study examined the effect of different priming levels on both real and AI-generated student responses, considering key metrics including mean, median, range, 25th percentile, 75th percentile, and standard deviation.The analysis reveals a number of important trends. Real students performed best at priming levels 2, 4, and 5, with mean similarity scores of 0.862496, 0.874104, and 0.841059, respectively (as seen in Table 1 and Figure 1). AI students, on the other hand, achieved the highest mean similarity scores at priming levels 3 and 7, which were 0.879742 and 0.830002 respectively, as per Table 3 and Figure 1.

Notably, the variability within each priming level, as indicated by the standard deviation and range (Tables 1 and 3, Figures 2 and 5), was substantial. This underscores the complexity of the impact of priming on student responses. Specifically, priming level 2 for real students and level 1 for AI students showed the greatest variability.

Insight into how the lowest-performing quartile of students is affected by different priming levels can be gained by examining the 25th percentile (Tables 2 and 4, Figure 3). For the highest-performing quartile of students, the 75th percentile offers similar insights (Tables 2 and 4, Figure 4).

The statistical significance values show that, except for the priming level 1 in real students, all other levels have a significant effect on the outcome (Tables 2 and 4).

Table 1: Real Student Answer Priming Results

	Mean	Standard Deviation	Median	Range
Priming Level				
0	0.761385	0.132329	0.769439	0.614629
1	0.663537	0.273452	0.801488	0.780228
2	0.862496	0.080955	0.883616	0.400025
3	0.817859	0.086863	0.839072	0.382006
4	0.874104	0.073522	0.891763	0.310368
5	0.841059	0.137673	0.922892	0.420877
6	0.823131	0.120182	0.86674	0.423894
7	0.668184	0.202088	0.683566	0.740583

Table 2: Real Student Answer Priming Results Continued

	25th Percentile	75th Percentile	Stat. Significance
Priming Level			
0	0.699885	0.858907	
1	0.353112	0.892312	0.4270
2	0.838924	0.915826	0.0000
3	0.750789	0.883273	0.0001
4	0.809037	0.942788	0.0000
5	0.693598	0.947415	0.0000
6	0.702175	0.933303	0.0000
7	0.493445	0.862063	0.0001

Table 3: AI Student Answer Priming Results

	Mean	Standard Deviation	Median	Range
Priming Level				
0	0.692499	0.150759	0.690355	0.548257
1	0.475718	0.214089	0.505027	0.844209
2	0.741282	0.25741	0.836771	0.865159
3	0.879742	0.050141	0.879064	0.229335
4	0.829996	0.117703	0.856047	0.489258
5	0.813032	0.096754	0.790829	0.340434
6	0.822828	0.096625	0.814009	0.3234
7	0.830002	0.093903	0.825931	0.313015

Table 4: AI Student Answer Priming Results Continued

	25th Percentile	75th Percentile	Stat. Significance
Priming Level			
0	0.589499	0.804681	
1	0.271873	0.634754	0.0000
2	0.747823	0.903618	0.0000
3	0.848205	0.916973	0.0000
4	0.800878	0.908517	0.0000
5	0.737409	0.902244	0.0000
6	0.740318	0.92538	0.0000
7	0.740364	0.926502	0.0000

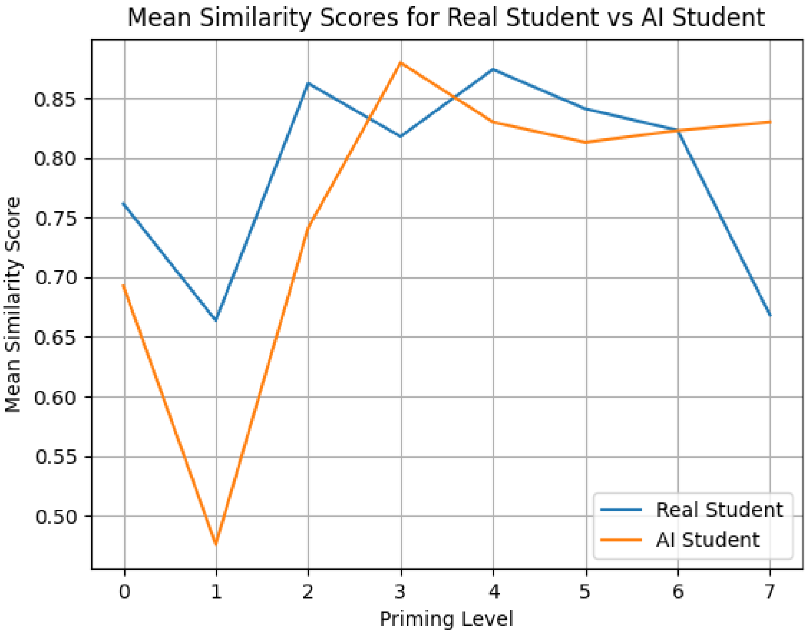

Figure 1: Mean Similarity

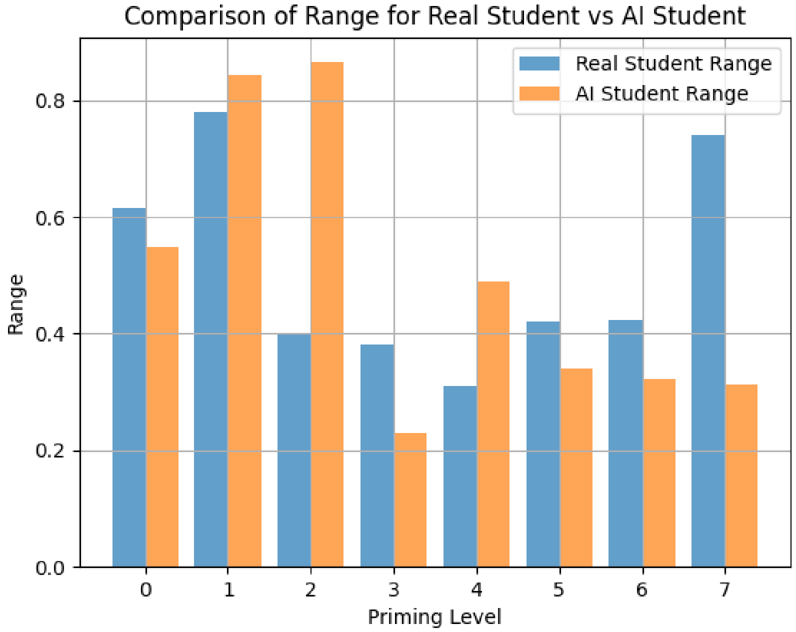

Figure 2: Range of Similarity

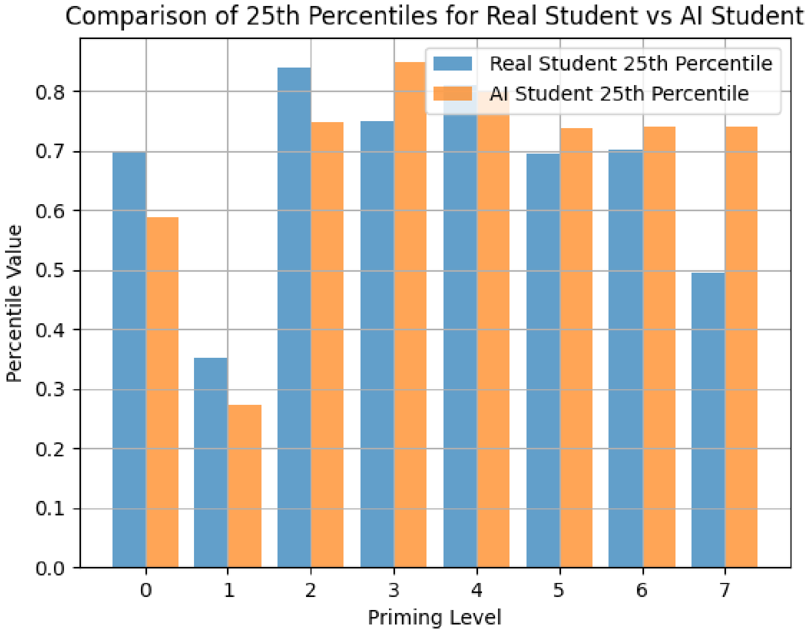

Figure 3: 25 Percentile

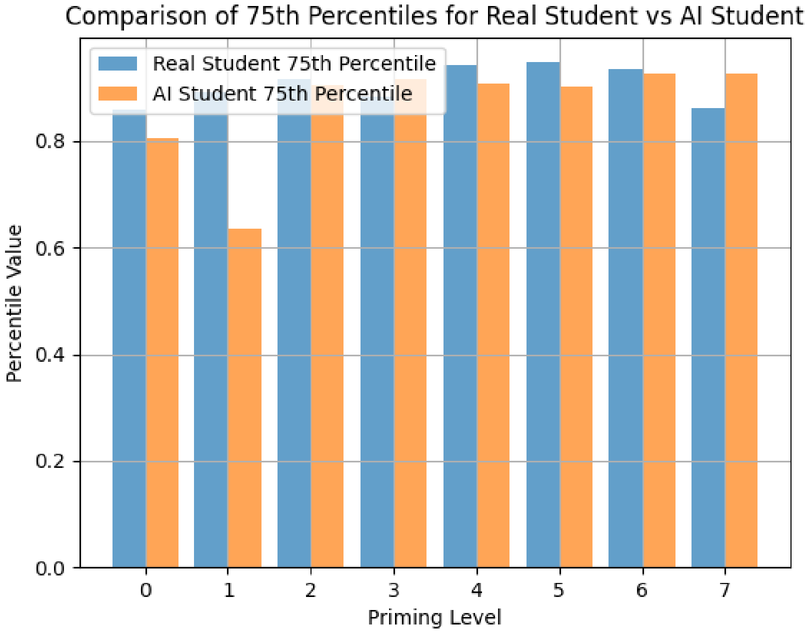

Figure 4: 75 Percentile

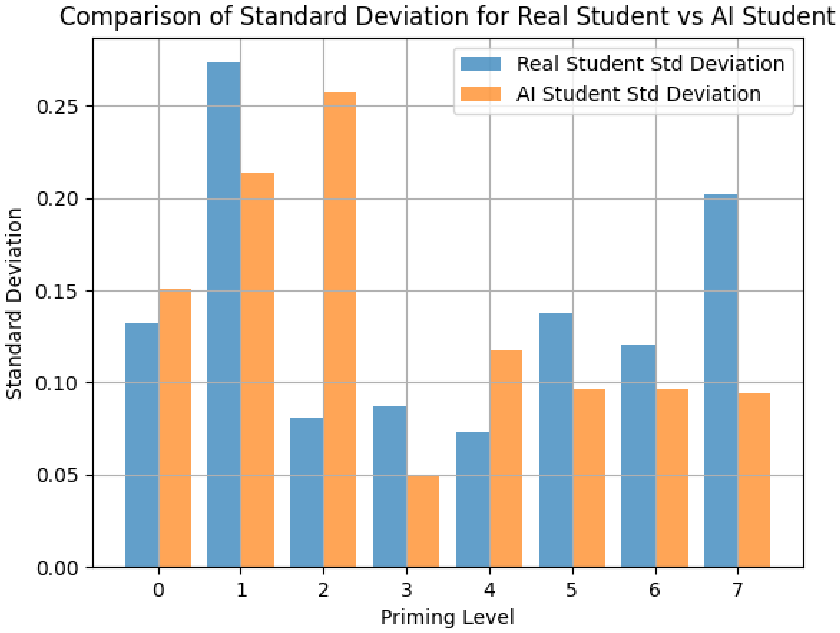

Figure 5: Standard Deviation

7 Conclusion

In conclusion, the results from this study illuminate intriguing implications about the influence of priming on student responses and AI algorithms. Notably, it was the AI-primed model, not the one primed with human-authored responses, that achieved the highest text similarity, signifying the most consistent tone.

This finding suggests that although a person may use an AI model like ChatGPT with the intention of ”training” it to emulate their unique tone for tasks like homework, the outcomes may not always align with their expectations. The efficacy of priming appears to be quite variable, underscoring the need for further research.

8 Future Work

Future investigations could consider larger text bodies and the effects of priming on different AI models. Another prospective exploration is the personalization of priming levels based on learner characteristics or AI specifics. These research directions could enhance our understanding of priming's impact and its optimization in educational settings.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.