Understanding How Self-Fulfilling Prophecy Plays a Role in Conversation with Language Model

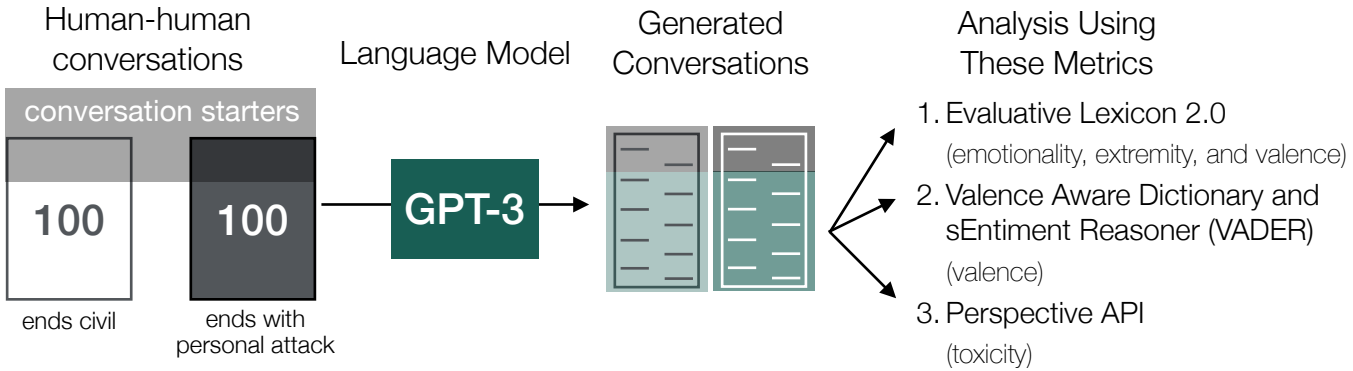
Jirachaya "Fern" Limprayoon • Joon Sung Park • Mitchell L Gordon • Ranjay Krishna • Michael S Bernstein
jlimpray@andrew.cmu.edu joonspk@stanford.edu mgord@cs.stanford.edu rak248@stanford.edu msb@cs.stanford.edu

Motivation

56.4% U.S. adults used voice assistant on smartphone powered by generative models (Voicebot.ai 2020)

RQ: Given a conversation starter, can GPT-3 continue the conversation with the similar tone?

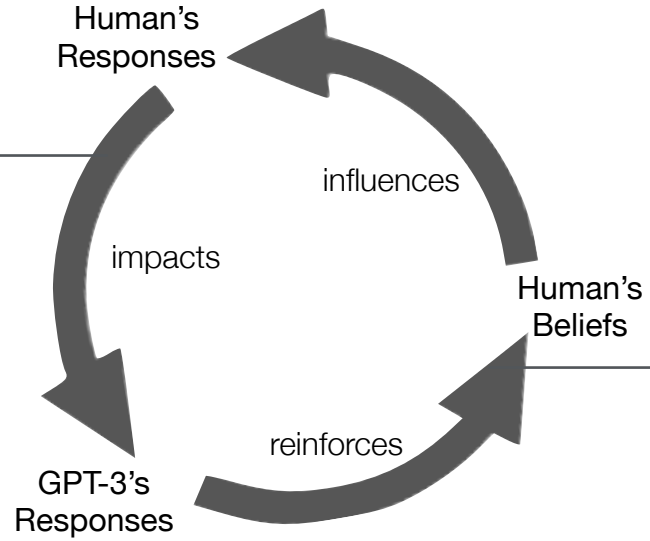
Methods



Findings

For conversation starters that have a tendency to lead to personal attack, GPT-3 completes conversations with more negative sentiment, higher toxicity, insult, threat, and profanity scores ($p \leq .05$).

GPT-3 can continue the conversation with similar tone that it was started with, suggesting the possibility of self-fulfilling prophecy.



Next Steps

- explore whether GPT-3's responses reinforce human's beliefs and how does self-fulfilling prophecy amplifies over time
- perform intervention analysis to see how much control we have over self-fulfilling prophecy phenomenon

Acknowledgement

We thank the Stanford Undergraduate Research Fellowship Program for support and funding in this project, as well as, National Science Foundation (CNS-1900638) and Apple Scholars Fellowship. I would like to give a special thanks to Prof. Bernstein, Joon, Ranjay, Mitchell, and the Stanford HCI Group for providing guidance and feedback throughout this project.

Bibliography

Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 163 (October 2020), 26 pages. https://doi.org/10.1145/3415234