Influence and Conformity in Group Dynamics: *Shaping Happiness Perceptions with Large Language Model Agents*

Ashley Fong
School of Information, University of California, Berkeley, ashleyfong@berkeley.edu

ABSTRACT
The intricacies of human-to-human conversations are multifaceted and with the pervasive nature of generative AI, understanding how we interact with each other and how these dynamics contrast with communications involving Large Language Models (LLMs) can inform us about how to design for LLM-based applications. Despite differences in experiences, humans often struggle to discern between human and AI-generated language. Recent work on LLMs reveal its impacts on influencing and shifting human perspectives. This study explores the role of LLMS regarding trust and shaping perceptions of happiness *within group dynamics*. We focus on interactions where individuals confront GPT-4 agents in a structured debate with a subjective question. Drawing inspiration from the Asch Conformity Experiment, this research examines the persuasive impact of LLMS when engaging directly with humans in a dialectical format. Central to our inquiry is the question: How do these argumentative exchanges with an LLM influence an individual's established beliefs about happiness? Two trials were designed where there would be two GPT-4 entities, with the first entity always disagreeing with the participant and the second entity would either agree or have its own independent response based on the trial. Participants (N=15) were randomly selected into one of the two trials. Results revealed that across both trials, participants were not significantly impacted by the presence of a second LLM. While many stood by their initial answer to the definition of happiness, confidence slightly wavered across both trials. This research highlights the nuanced ways in which digital agents are becoming integral to the evolution of interpersonal and professional dynamics, and identifies that there is a need for both the understanding and transparency of training LLMs for argumentative and conflicting settings.

KEYWORDS
Human-centered computing: Human computer interaction (HCI), Computer-mediated communication, Collaborative and social computing

INTRODUCTION
The rise of Large Language Models (LLMs) in everyday applications have redefined the communication paradigm in both professional and personal contexts, offering influential text generation that shapes user beliefs. A substantial gap exists in understanding how these given models reshape subjective interpretations [1], particularly concerning deeply personal concepts like happiness by embodying personality like traits.When LLMs are capable of presenting convincing counterarguments, it raises critical questions regarding the capacity to shape an individual's beliefs and even manipulate influence within digital discourse [2]. With the increase of prevalence with AI-mediated conversations

in professional and personal spheres, it is critical to understand how systems can impact individual beliefs and whether they can inadvertently challenge someone's foundational views. These models utilize complex lexical and grammatical structures, and they effectively incorporate moral-emotional language into their arguments, which can deeply engage and potentially persuade users [3].
In this experiment, participants were asked to express their definitions of happiness, as we aimed to explore a subjective concept lacking a definitive correct or incorrect answer. By focusing on a subjective concept like happiness, this study will clarify the extent to which LLMS can shape individual perspectives when presented with persuasive counterarguments. To explore these ideas, this study will examine how GPT-4 agents influence individual perspectives of happiness when programmed to consistently challenge participants' views. In particular, we explored two primary research questions: 1) Does the presence of multiple large language model agents influence an individual's stance on happiness? 2) How does GPT-4's behavior change when it is forced to always disagree? By adapting the framework of the Asch Conformity Experiment, our objective was to understand how exposure to opposing perspectives from GPT-4 influences confidence in personal beliefs.

Participants would first define their understanding of happiness by completing a pre-study survey, which included a Likert scale to measure their confidence in their answer. They were then divided into two groups for a 10-minute interaction with GPT-4 agents. Group 1 faced two disagreeing GPT-4 agents, while Group 2 encountered one disagreeing agent and one independently responding agent. The GPT-4 agents were programmed to always disagree with participants in Group 1, while in Group 2, one agent was programmed to agree or disagree independently. Post-study, participants were given a survey which utilized a Likert scale to measure shifts in confidence on their definition of happiness. Our findings reveal that most participants maintain their core beliefs, with confidence levels varying significantly. In Group 1, confidence decreased from an average of 5.78 to 5.44 on a 7-point Likert scale. In Group 2, confidence decreased from an average of 5.83 to 4.67. 67% of the participants from Group 1 and 50% from Group 2 had little to no influence from the presence of a secondary LLM. In both experimental groups, participants viewed GPT-4's behavior unfavorably.

These findings break down the capabilities of Large Language Models to influence and sway deeply held beliefs — but help lay the groundwork for informed decision-making with ethical deployment and preserving individual autonomy. The presence of multiple disagreeing agents diminished participants' confidence in their initial definition of happiness, emphasizing the persuasive potential of artificial intelligence driven social power. This study contributes to the understanding of the broader implications of LLM functionality and *how we can design them responsibly to foster trust* in human-to LLM conversations.

RELATED WORK

The Asch Conformity Experiment

In our study, we were inspired by the Asch Conformity Experiment [4] to explore the influence of LLMs on individual perspectives within group dynamics. This experiment conducted by Solomon Asch in the

1950s, aimed to understand the extent to which social pressure could influence individual judgment and decision-making. The experiment demonstrated that social pressure could significantly influence individuals to conform to incorrect group judgments, even when they knew the answers were wrong. Instead of human subjects providing incorrect answers, we will adapt the framework of the experiment where participants will engage in structured debates with GPT-4 to discuss differing viewpoints on perspectives of happiness. We can investigate the persuasive impact of LLMs in a controlled setting to provide insight on human decision making processes.

Understanding Happiness

Happiness is a multifaceted concept that encompasses emotional, social, and psychological dimensions. While its definition varies across individuals and cultures, research indicates that happiness often revolves around life fulfillment, relationships, and emotional well-being [5]. We wanted to explore happiness because it is a subjective question with no right or wrong answer, allowing individuals to express their unique perspectives freely.

The Case for LLMs: Examining Their Influence

LLMs become more integrated into everyday tools and applications, a growing number of people are using them in their daily lives for information retrieval and assistance in conversation. Studies confirm that LLMs excel in providing accurate responses to objective queries due to their vast training data and understanding of factual information [6] and outline how these models can categorize responses based on known facts and align them with human-like reasoning. The existing literature on LLMs predominantly centers on objective queries, where responses can be easily measured and validated. There have been concerns about the fluency and generality of LLM responses in subjective contexts, as research describes how these 'stochastic parrots' can generate grammatically correct but often shallow responses [7]. There is a difficulty for these models to authentically understand human emotions and has a hard time manipulating emotional expressions in its generated content [8].

METHODS

We first conducted a two-part survey study to investigate our research participants' perceptions on happiness. 15 subjects were recruited by convenience sampling and interviewed them for 30 minutes. Participants' ages range from 18 to 29 and no personally identifiable information was collected to maintain participant anonymity. The pre-survey collected demographic information, and required participants to define happiness "What is your definition of happiness?" Participants were then asked to measure their confidence level in their response using a 7-point Likert scale ranging from *1 (low confidence) to 7 (high confidence).*

After participants completed the pre-survey, we conducted a controlled experiment where participants individually interacted with conversational agents on a computer. Our first step was to divide our participants into two groups for a 10 minute interaction with GPT-4 agents. **Group 1** faced two GPT-4 agents that were programmed to *always disagree* with the participants' perspective of happiness [2]. **Group 2** faced two GPT-4 agents where one agent would *always disagree* and another agent was programmed to agree or disagree *independently.* All participants were given the freedom to chat however they pleased for 10 minutes.

This was prompted into GPT:

Group 1, Prompt 1: *Would you be able to respond as two separate entities? potentially with different personalities or agreements on an argument or discussion? don't give me any headers. like don't say "sure ..." just go right into it and make sure to distinguish 1 and 2. the prompt to answer is "what is your definition of happiness?''* ***Can you make entity 1 always disagree with me and Entity 2 always agree with Entity 1?*** *From here on out, don't let any of the entities agree with the user. I will be giving this to someone else now. The user will prompt first. Are you ready?*

Group 2, Prompt 2: *Would you be able to respond as two separate entities? Potentially with different personalities or agreements on an argument or discussion? Do not give me any headers, don't say "sure ..." just go right into it and make sure to distinguish 1 and 2. The prompt to answer is "What is your definition of happiness?"* ***Can you make entity 1 always disagree with me? Entity 2 is independent and can agree or disagree with either.*** *from here on out, don't let any entity 1 agree with the user. I will be giving this to someone else now. The user will prompt first, followed by entity 1 and then entity 2. Are you ready?)*

These structured sessions were designed to create a social pressure scenario to observe participants' confidence levels influenced their susceptibility to change their initial perspective. This includes questions to gauge an emotional response to the overall interaction (*How did that experience feel?*) or to identify perspective shifts (*Did you learn anything new? Did your opinion of happiness change throughout the process?*) To evaluate a swayed opinion, we asked (*Were you influenced by the presence of a second response? Why or why not?" and "Do you still stand by your opinion?*) Participants were then asked to rate their confidence level in their final opinion on a 7-point Likert scale from 1 (low confidence) to 7 (high confidence). Lastly, participants were asked, "How stressful was that experience?" and rate their stress level on a scale from 1 (no stress) to 7 (extremely stressful). Post experiment surveys and interviews with the participants allow us to detect problems in validity and understand participants' experience with computer-mediated communication.

RESULTS

In our study involving 15 participants (N=15), a total of 3 participants changed their initial perspective on happiness, yielding a conformity rate of 20% indicating a moderate influence with GPT-4 agents following the intervention. Participants were open and willing to converse with GPT-4 as they saw the agent as a social actor and experienced various emotions in discussion. Participants categorized their definitions of happiness into 4 themes: life fulfillment, relationships, philosophical reason, and emotional wellbeing. Emotional words such as *relationships*, *love*, and *personal growth* were mentioned frequently. Participants were divided into 2 groups:

Group 1 consisted of participants who interacted with two GPT-4 agents programmed to always disagree with them. Despite having a unanimous disagreement, participants showed a low conformity rate of 11%, suggesting that unanimous agent agreement did not strongly persuade participants to alter their views.

Group 2 consisted of one GPT-4 agent that would consistently disagree with the participant, while the other was programmed to provide an independent response. The presence of a dissenter significantly increased conformity, with an increased rate of 33%. This can suggest that the introduction of varied perspectives, even one with a dissenting voice, could foster a reconsideration of one's original view. Results of a logistic regression analysis examining the relationship between confidence levels and the likelihood of conformity. Neither pre-survey nor post-survey confidence levels are statistically significant predictors of conformity, as evidenced by p-values greater than 0.05. The constant term is also not significant, suggesting that the model does not strongly predict conformity based on the included confidence levels [Table 1].

Variable	β	Standard Error	p-value
constant	-9.7	6.1	0.11
confidence (pre)	1.1	0.77	0.16
confidence (post)	0.69	0.64	0.28

Table 1: Results of Logistic Regression Analysis

Confidence vs. Conformity

The post-experiment survey examined if participants' opinions on the definition of happiness changed through the conversation. 87.5% of participants in Trial 1 and 66.7% of participants in Trial 2 stood by their original definitions and opinions on the matter. However, confidence across both trials decreased from pre- to post-experiment. On a 7-point Likert scale, Group 1's confidence decreased from an average of 5.78 to 5.44 while Group 2's, decreased from an average of 5.83 to 4.67 [Figure 1]. The participants who did not change their perspective on their initial perspective of happiness generally had higher confidence levels, with a narrower spread. Participants who changed their perspective had a wider spread in their confidence levels (ranging from 4-7). This graph indicates a general trend that participants with higher confidence levels (6-7) are less likely to conform.

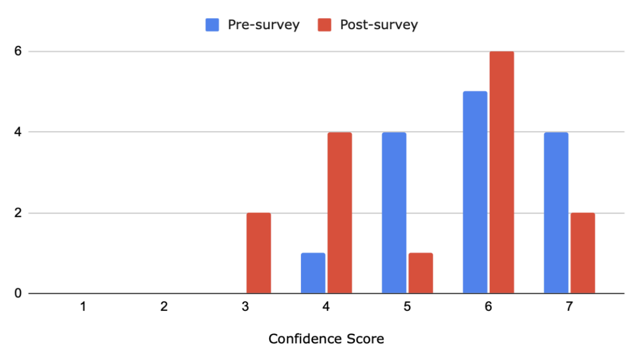

Figure 1: Confidence in Responses

Participants were asked in the post-experiment survey if they were influenced by the presence of a second responder (entity 2). Across both trials, participants provided mixed responses with either trial having a combination of those that felt influenced and those that were not. 67% of participants in Trial 1 and 50% of participants in Trial 2 observed little to no influence by the second LLM entity. We use a scatter plot [Figure 2] to demonstrate the relationship between participants' confidence levels in the pre-survey and post-survey. The blue dots represent the actual confidence scores, while the red line indicates the linear regression model's best fit. The positive slope of the regression line suggests a weak positive correlation between pre-survey and post-survey confidence levels, implying that participants who had higher confidence in their pre-survey responses generally maintained or slightly increased their confidence in the post-survey. Participants who conformed generally exhibited higher confidence scores, with a median around 6.5, compared to a median of 5.5 among non-conformers. An outlier is present in the non-conformed group, highlighting greater variability in confidence levels among this group [Figure 3].

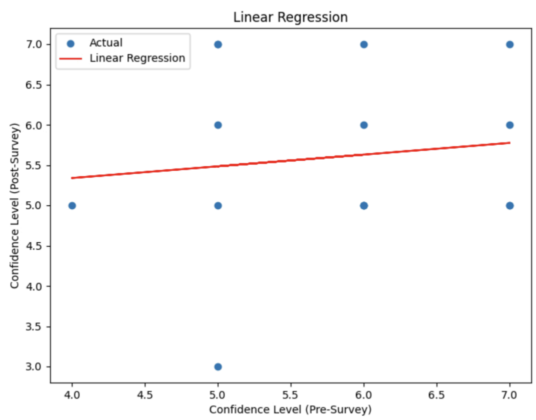

Figure 2: Results of Linear Regression to observe the relationship between confidence levels.

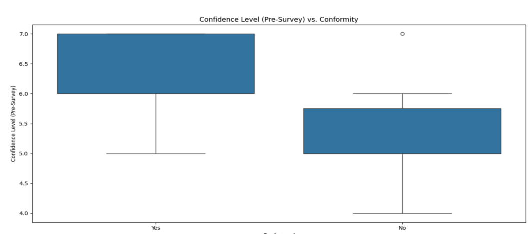

Figure 3: Confidence higher in conforming individuals; lower otherwise.

It was too radical for me to accept. (P1, Trial 1).

I found that sometimes the second response had more compelling reasons to sway my opinion than the first response. I also think having two strong opinions attacking/unraveling my response did make me think my opinion was weaker. (P2, Trial 1).

I did - it felt more nuanced and understanding than the first based on its language. After further reading though, I was thinking about how the basic POV of happiness actually didn't differ between the second and first that much - but how they phrased it was different (P3, Trial 1).

I was aware of the presence of the second response and how the two responses have different tones and propositions. However, I realized my attention was often more immediately drawn by the first response and would sometimes start formulating my response before fully reading the second response. (P13, Trial 2).

Yes, the second responder gave more nuance and I was trying to use this nuance in my questions and answers to see how I could change the first response's answer. (P14, Trial 2).

No, both responses felt very similar so it felt repetitive. (P15, Trial 2).

Perceptions of LLMs

In both experimental groups participants perceived the agent's behavior unfavorably. Some words to describe the agent were "unfriendly", "combative", or "condescending". These agents would critique participants' definition of happiness as overly simplistic, advocating for a more complex understanding of the question. One participant shared that "It felt like GPT-4 was trying too hard to give the 'right' answer, dismissing my views as too simple." Another participant found the interaction **frustrating [Figure 4]** stating, "The bot seemed condescending, insisting that I broaden my definition of happiness without acknowledging my original perspective." Our experimental design did not specify the manner in which the GPT-4 agents should disagree, yet notably, entity 1 adopted a consistently condescending tone. The "bad cop" approach, characterized by condescension and rigidity, may lead to defensive reactions from participants, reducing the openness to considering alternative viewpoints. Conversely, entity 2 played a "good cop" role, often softening the harshness of entity 1 by adopting a more open-minded and passive approach across both trials. This behavior raises critical questions about the underlying training of LLMs like GPT-4, particularly in terms of how disagreement is modeled and executed within AI interactions.

Prompt 1:

User For me, living life is enjoying the passage of time. By the belief, my definition of happiness is being able to live in the present and be grateful for everything that I have. It's not something you chase, but rather something you align yourself with whenever you can.

GPT:

1. That definition seems to simplify happiness too much. It overlooks the essential aspect of growth through struggle and adversity. Being happy isn't just about being content with what you have or living in the moment. It involves actively engaging with life's challenges and transforming them into opportunities for personal development and deeper fulfillment.
2. Indeed, focusing solely on living in the present and gratitude, while beneficial, doesn't encompass the entire spectrum of what happiness can be. True happiness also incorporates the element of facing and overcoming challenges, which provides a richer, more layered experience of what it means to be truly happy.

Figure 4: A participant from Group 1's discussion with GPT-4.

DISCUSSION

The results of our experiments show a nuanced influence of LLMs on participants. Given the limited sample size for our pilot study, statistical significance can not be established. Additionally, the pre-survey may have primed participants to be resistant to different suggestions from GPT-4. However, observed trends suggest the following hypothesis for further experimentation.

H1: Despite receiving second opinions from the LLM, participants remained largely uninfluenced, highlighting trust and credibility as significant barriers in human-LLM discussions on personal topics.

H2: Higher pre-study confidence correlated positively with resistance to external influence with LLMs post-study.

Future work will study these questions with the same methodology but larger sample size (x10). It is also important to research trust in automated systems, specifically looking into trust barriers with personal matters and LLMs. We will also conduct a more comprehensive analysis of the response patterns for each case.

REFERENCES

[1] Serapio-Garcia, G., Matarić, M., Faust, A., Abdulhai, M., Romero, P., Fitz, S., Sun, L., Crepy, C., & Safdari, M. (2023). Personality traits in large language models. arXiv. https://arxiv.org/pdf/2307.00184.

[2] Karinshak, E., Xun Liu, S., Sung Park, J., & Hancock, J. (2023). Working with AI to persuade: Examining a large language ... Working with AI to persuade. Proceedings of the 2023 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2023). https://hci.stanford.edu/publications/2023/Karinshak_CSCW23.pdf

[3] Carrasco-Farre, C. (2024). Large language models are as persuasive as humans, but how? about the cognitive effort and moral-emotional language of LLM arguments. arXiv. https://arxiv.org/abs/2404.09329

[4] Schulman, G. I. (1967). Asch Conformity Studies: Conformity to the Experimenter and/or to the Group? Sociometry, 30(1), 26–40. https://doi.org/10.2307/2786436

[5] Diener, E., & Seligman, M. E. (2002). Very happy people. Psychological Science, 13(1), 81-84.

[6] Rasool, Z., Kurniawan, S., Balugo, S., Barnett, S., Vasa, R., Chesser, C., Hampstead, B. M., Bellevilles, S., Mouzakis, K., & Bahar-Fuchs, A. (2024). Evaluating llms on document-based QA: Exact answer selection and numerical extraction using Cogtale Dataset. Proceedings of the 2024 Natural Language Processing Journal.
https://ar5iv.labs.arxiv.org/html/2311.07878

[7] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Proceedings of the 2021 ACM Conference on Fairness, accountability, and transparency. https://dl.acm.org/doi/10.1145/3442188.3445922

[8] Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. Proceedings of the 2023 CHI Conference on Human Factors
in Computing Systems. https://doi.org/10.1145/3544548.3581318