Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference

Thitaree Tanprasert
tt1996@cs.ubc.ca
University of British Columbia
Vancouver, British Columbia, Canada

Sidney Fels
ssfels@ece.ubc.ca
University of British Columbia
Vancouver, British Columbia, Canada

Luanne Sinnamon
luanne.sinnamon@ubc.ca
University of British Columbia
Vancouver, British Columbia, Canada

Dongwook Yoon
yoon@cs.ubc.ca
University of British Columbia
Vancouver, British Columbia, Canada

ABSTRACT

Exposure to diverse perspectives is helpful for bursting the filter bubble in online public video platforms. The recent advancement of Large Language Models (LLMs) illuminates the potential of creating a debate chatbot that prompts users to critically examine their stances on a topic formed by watching videos. However, whether the viewer is influenced by the chatbot may depend on its persona. In this paper, we investigated the effect of two relevant persona attributes - social identity and rhetorical styles - on critical thinking. In a mixed-methods study (n=36), we found that chatbots with outgroup (vs. ingroup) identity ($t(33)=-2.33$, $p=0.03$) and persuasive (vs. eristic) rhetoric ($t(44)=1.98$, $p=0.05$) induced critical thinking most effectively, making participants re-examine their arguments. However, participants' stances remain largely unaffected, likely due to the chatbot's lack of contextual knowledge and human touch. Our paper provides empirical groundwork for designing chatbot persona for remedying filter bubbles in online communities.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

filter bubble, critical thinking, conversational agents, agent personas, online public videos

ACM Reference Format:
Thitaree Tanprasert, Sidney Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3613904.3642513

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642513

1 INTRODUCTION

The phenomenon of 'filter bubbles' is an ongoing concern on social media platforms [81, 82], especially on online video platforms such as YouTube, which has more than 2.7 billion monthly active users and 3.5 billion daily searches as of 2023 [1]. While text-based platforms facilitate brief exchanges that quickly switch across multiple topics, YouTube's longer-format videos allow for a more in-depth presentation of ideas, making it a popular learning resource on content about culture, belief systems, politics, and societal issues [54]. However, as "prepackaging of intellectual positions and views is so ingenious that thinking seems unnecessary," [70] the video's audio-visual nature, combined with the influential status of YouTubers, can lead to more immersive and, hence, more passive consumption of content. This is particularly concerning as the platform's recommendation algorithms can limit exposure to diverse viewpoints, potentially prompting the bandwagon or false consensus effect that reinforces existing societal biases and stereotypes [19, 88]. Moreover, YouTube's susceptibility to misinformation, including fake news and unverified content [48], is amplified by the fact that video media literacy education is still not prevalent compared to that of text-based media [26]. This gap contributes to reinforcing filter bubbles and undermines YouTube's capacity for fostering cultural competence [9, 13, 97].

Given these challenges, critical thinking skills become especially necessary on video platforms like YouTube. Critical thinking involves the objective analysis and evaluation of information to make a judgment. It not only helps viewers navigate through algorithmic biases and passive content consumption but also enables them to discern misinformation [69]. Engaging in debates with individuals holding differing perspectives is a proven method to enhance critical thinking [106], especially when it comes to understanding diverse cultures [31]. But implementing this on YouTube is challenging. Viewers may struggle to find a trustworthy and respectful discussion partner with opposing views and may feel apprehensive of offending the partners when discussing sensitive or controversial topics, such as political movements and trendy social issues [7, 45, 67]. Therefore, there's a growing need for a safe space or a reliable entity on the platform where viewers can engage with diverse opinions and participate in constructive debates, broadening their perspectives and challenging prevailing biases.

LLM-based chatbots recently emerged as a viable solution to implement such debates. AI chatbots can provide proactive, adaptive,

and readily available agentic intervention [74]. With the recent development of large-language models (LLMs), chatbots have the capacity to understand video content [22], mimic personas [92], and engage in complex and coherent conversations [52]. Most importantly, they have been shown to have the ability to craft persuasive messages and support their stance with strong arguments [51]. Therefore, we proposed a chatbot interface that pops up after a user finishes watching a video that is suggested to them by the recommendation algorithm and reaffirms their opinions. The chatbot takes an opposing stance to the user and the video and invites the user to debate on the topic. With logically coherent, persuasive, and engaging conversation, the chatbot should induce "in-the-moment" critical thinking in the user and help the user develop media literacy skills over time.

To make the chatbot effective in prompting critical thinking, the design of chatbot personas is essential. By giving chatbot specific personas, the user has more concrete expectations about the chatbot's level of intelligence and actions, making their interactions more engaging [30, 59]. As the chatbot's goal is to challenge the influence of YouTubers, it needs to be just as engaging and have a similar influence on users. Existing literature shows that there are many persona attributes that make human's arguments credible and influential [56, 110, 113]. In this paper, we investigated two of these attributes: social identities (ingroup vs. outgroup) and rhetorical styles (persuasive vs. eristic). We chose social identity, as it is shown to significantly impact the perceived trustworthiness of the person behind an opinion [98]. We chose rhetorical styles to set the goals of the argumentative discourse in a way that is most conducive to critical thinking [4]. Specifically, our study aims to answer three research questions:

- **RQ1.** How do two key attributes of chatbot persona (i.e., social identity and rhetorical style) influence critical thinking in video viewers?
- **RQ2.** To what extent does interacting with a chatbot affect a viewer's stance on a topic formed after watching a video?
- **RQ3.** How does the debate chatbot affect the viewer's engagement with and motivation to do the activity (i.e., watching the video and discussing it with the chatbot)?

To answer these research questions, We ran a mixed-method experiment with 36 participants from North America, where the participants watched video essays that conformed to their opinions and then debated with chatbots of various personas that took an opposing stance. We found that both persona attributes affect the participant's level of critical thinking. Notably, the combination of outgroup identity and persuasive rhetoric was perceived to be more conducive to interpretation, analysis, and self-regulation. We also found that, although the chatbots prompt participants to reexamine their arguments, the interactions have negligible effects on the participant's stance after it is reinforced by watching the video essays.

The paper contributes empirical findings on the ability and limitations of LLM-based chatbots to act as debate partners and induce critical thinking in filter bubble situations. Specifically, our study (1) highlights the importance of chatbot persona design, specifically its social identity and rhetorical style, on increasing the user's critical thinking and engagement; and (2) identifies the chatbot's behaviors which contribute to its ineffectiveness in persuading the user to compromise their stance. The paper also provides design implications for designing LLM-based applications for addressing the filter bubble problems.

2 RELATED WORK

2.1 The importance of social identity on online video platforms

In this section, we characterize the influence of video creators on online public video platforms and the factors that affect the level of their influence in various contexts. We also highlight the effect of social identity on influencers' persuasiveness and trustworthiness, which we attempt to replicate in the chatbot.

2.1.1 Factors affecting the influence of microcelebrities on online video platforms. On online public platforms, creators such as "YouTubers" can amass a significant fanbase and, through their content and social media interactions, shape the perceptions and opinions of their followers. This makes online videos a powerful tool for activism [24, 54], cross-cultural communication [13, 83], and transformative education [28, 101]. However, there is also a concern that the microcelebrity status of the video creators and the personalized recommendation algorithms may lead to the filter bubble problem, where the viewers are constantly exposed to perspectives they enjoyed and are completely disconnected from opposing viewpoints [16, 17, 60]. It is observed that some members of Gen Z audiences might prioritize information sensibility over information literacy (i.e., the ability to interact with information). This suggests that, in certain cases, the evaluation of information credibility can lean more towards aspects such as social belonging, aesthetics, and convenience, rather than thorough and rational deliberation [44, 107].

There are many areas where YouTubers can influence their audiences and many factors which can affect the extent of their influence. Previous research found that YouTubers are considered credible in reviewing and endorsing products and that their fans perceive the product in the same frame as presented in their videos [79]. The level of their information credibility depends on their perceived expertise, trustworthiness, and homophily (alignment of preferences) [110]. As for videos that discuss political and societal topics, viewers appreciate YouTubers over traditional media for their entertainment and their "relatability, authenticity, and accountability" [62]. Viewers' evaluation of YouTubers' credibility is positively correlated with the perceived authenticity of their opinions and viewing frequency [113]. Beyond information credibility, YouTubers can gain commitment from and influence their viewers with the frequency of interactions, as well as congruent communication styles and interests [56].

2.1.2 Social identity, persuasiveness, and trustworthiness. Social identity is the self-categorization according to group membership into "us" and "them", or more formally, ingroup and outgroup [47, 96]. Group membership may be essential, i.e., an innate part of oneself, such as ethnicity, gender, or age. It can also be temporary or context-specific, such as occupation, political affiliation, or fandom of a celebrity [10]. Social identity theory posits that social identity affects one's social behaviors, and different categorization factors

can affect one's behaviors in different contexts, ways, and levels [65, 89].

In persuasion activities, agreement with ingroup individuals can assert the referent informational influence, that is a sense of self-validation and a perception of the viewing as external reality and objective truth [102]. Therefore, ingroup individuals often have higher credibility and trustworthiness than outgroup [109]. However, some previous research has found the opposite, which suggests that the effect of social identity on persuasiveness may depend on the specific factors that define group membership and the other variables at play, such as the power or majority status of the persuader [15].

Our paper aims to address the filter bubble problem through the deployment of LLM-based chatbots, as well as explore how chatbots' persona attributes, specifically social identity, affect their influence and ability to induce critical thinking in video viewers, and how they are similar or different from the effect of social identities of real YouTubers reported in the existing literature.

2.2 Fostering critical thinking in online communities

In this section, we define the notion of critical thinking as used in this paper, review its importance in the context of online media consumption, and highlight the role that rhetorical styles play in dialogic pedagogy, an approach for increasing critical thinking that we applied to the chatbot.

2.2.1 Critical thinking on online video platforms. There is no single unanimously accepted definition of critical thinking. In our research, we restrict our focus to the core concept of critical thinking, which is "careful, goal-directed thinking" [46], where the goal is to adopt a stance on a debate topic. In this research, we considered six types of critical thinking according to Facione's taxonomy [34], as the taxonomy is applicable to our context and has a readily available, standardized assessment tool: the Critical Thinking Self-Assessment Scale (CTSAS) [78]. Here are the definitions for each type of critical thinking as we applied it to our work:

(1) Interpretation: the ability to comprehend the context of the problem (e.g., situations, rules, procedures) and the provided data (e.g., judgments, arguments, beliefs).
(2) Analysis: the ability to identify the relationship between different arguments and the implicit assumptions in the reasoning
(3) Evaluation: the ability to assess the credibility and the logical strength or relevance of an argument
(4) Inference: the ability to consider the consequence of an argument and draw a conclusion based on evidence
(5) Explanation: the ability to articulate and justify one's opinions and arguments based on reasons and awareness of possible counterarguments
(6) Self-regulation: a conscious monitoring of one's thought process, including reflection on one's own values and the quality of one's judgment

In the context of online media consumption, critical thinking skills are closely linked to digital literacy and media literacy. Digital literacy, a narrower facet of information literacy, stresses the competency to use digital tools and discern credible sources [8]. Media literacy, on the other hand, focuses on analyzing media content, equipping users to detect biases and underlying messages [57]. On online video platforms, both skills should combine to empower viewers to critically assess not only the media content but also how the sources, the platform's algorithm, and the online community surrounding the content affect their perceptions and interactions with it [13, 43, 50].

2.2.2 The importance of rhetorical styles in dialogic pedagogy. Although there are many pedagogies for improving critical thinking associated skills, we adopted and implemented dialogic pedagogy, where students construct knowledge "through the questioning, interrogation and negotiation of ideas and opinions in an intellectually rigorous, yet mutually respectful, manner" [100]. As dialogues naturally expose one to new evidence or a different perspective, they can effectively induce critical thinking in any context and subject matter [84].

There are many variables that can affect the efficacy of dialogues on critical thinking, including the rhetorical style, argument components and structure, and contexts [18, 105]. Six common types of rhetorical styles or dialogues are: persuasion, inquiry, discovery, negotiation, information-seeking, deliberation, and eristic [104]. Each type deals with different initial situations and achieves different goals [106].

In this paper, we drew on existing literature to design the rhetorical styles of the chatbot. Specifically, we explore two relevant rhetorical styles: persuasive dialogue and eristic dialogue. We aim to extend the existing body of literature to LLMs by investigating the abilities, limitations, and different ways in which LLM-based chatbots utilize various rhetorical styles to induce critical thinking and influence the stance of online video audiences.

2.3 HCI for supporting critical thinking

In this section, we review existing research in HCI related to improving and supporting critical thinking. The first subsection reviews systems with conversational agents, which is the type of system proposed in this paper. The second subsection reviews other types of systems, focusing on those that target social media contexts, which is the context of this study. To our knowledge, no prior conversational agent-based systems for improving critical thinking in social media contexts have been proposed.

2.3.1 Conversational agents for improving critical thinking. Conversation agents have notably been adopted in various interfaces for fostering critical thinking. There is also a body of work proposing frameworks for training opinions and modeling critical thinking for such applications of conversational agents, which highlights the complexity of the data and the necessity for context-specific application designs [21, 29, 41]. In educational contexts, conversational agents have been used to provide adaptive feedback [36] and induce self-reflection [76] on the user's performance. In online communities, such agents have been used to facilitate crowd discussion [49], countering extremists [12], and question information credibility of news content [111]. The advancement of LLMs broadens the possible designs for encouraging critical thinking and diversity in

opinions [52], as such systems can provide adaptable and complex reasoning [71] as well as assume multiple personas and opinions [63].

2.3.2 Systems for improving critical thinking in social media. There has been substantial work on designing systems for supporting critical thinking and promoting diverse opinions in social media. Some work, such as Balancer [77] and Opinion Space [35], focus on visualization of the opinion space, clearly displaying the relative position of the user's own opinions or social media history to make users aware of their filter bubble. Other systems recommend or expose the user to different perspectives [39, 80], which can lead to more critical discourse, inducing respect for diverse opinions and mitigating the opinion polarization problem. StarryThoughts addresses this problem by incorporating the identity information of the speakers behind the opinions [55]. They found that the social identity of the speakers, relative to the users, can positively influence the reception of diverse opinions.

Our paper adds to this body of literature by designing a system that utilizes an LLM-based chatbot to induce the user's critical thinking in online public video social media platforms. We expect our findings to help demonstrate the ability of LLM-based chatbots and provide more insights about the effects of chatbot persona design in this context and application.

3 METHOD

To answer the research questions, we ran a mixed method study, consisting of a 2×2 mixed factorial experiment and a qualitative follow-up. The goal of the experiment is to evaluate two attributes of chatbot personas: social identities (within-subject; ingroup vs. outgroup) and rhetorical styles (between-subject; eristic vs. persuasive). The two video topics for this study were controlled, and the pairing social identity $\times$ video is also between-subjects, i.e., participants watch one video on one topic with the ingroup chatbot and a second video on another topic with the outgroup chatbot. The pairing of social identity and video topic, as well as the order of presenting the social identity, are counterbalanced. The dependent variables consist of the participants' stance on the topic, their critical thinking level, their engagement and motivation for the activity, and their perception of the chatbot.

3.1 Condition

In answering RQ1, we investigated two attributes of chatbots' personas: social identity and rhetorical style. We selected social identity (ingroup vs. outgroup) as an attribute because the problem of viewers passively accepting YouTubers' opinions is significantly related to the sense of belonging that surrounds influential YouTubers.

We hypothesized that, by giving the chatbot an identity that is easy to recognize as "belonging to the same group" as the viewer, we can increase its influence on the viewer's stance and thought process. We represented ingroup (or outgroup) identities with respect to the participants by aligning (or misaligning) the chatbot's ethnicity and gender to the participants. We chose ethnicity as it is one of the most effective factors in making people feel ingroup or outgroup with each other [58, 89]. However, the effect of ethnicity varies by nationality [53], so we fixed the chatbot's nationality to be American to match all participants. We chose gender because previous research shows that it is the most important factor for determining the trustworthiness of a stranger [61]. We considered other factors that may contribute to the feeling of ingroup-ness, including age [65], occupation [103], socioeconomic status [112], religion [87], and political beliefs [10]. However, the effect of these factors as reported in previous research are specific to problems unrelated to the context of this study. We embedded the social identity into each chatbot's name, profile picture, and short bio, which contains the chatbot's nationality and pronouns.

We selected two rhetorical styles, persuasive and eristic, as the other attribute of the chatbot's persona because they are applicable to the context of our study. With the persuasive rhetorical style, the chatbot tries to make the viewer adopt their view by sharing the advantage of their stance while remaining understanding of the viewer's viewpoint. In contrast, with the eristic style, the chatbot attacks the user relentlessly and exposes as many differences in their opinions, assumptions, and values, as possible [104].

3.2 Experimental materials

The debate topic, i.e., the content of the video essays, is a crucial variable in our experiment. We brainstormed for video topics that meet three criteria: (1) simple enough to make a sufficient argument in less than 4 minutes; (2) does not have a consensus stance; and (3) participants are unlikely to have extreme preconceived opinions on the topic. In the end, we chose two topics: "Online gatherings are better than in-person" and "Customers should tip." We asked the participants to pick a stance on each topic at the beginning of the study (See Figure 1 and Section 3.3 for details of this step) and verified that both of the selected topics followed the second and third criteria. (See distribution of the participants' pre-study stances in Figure 7, Appendix F.2.)

To create the videos, the lead researcher wrote the scripts for two video essays on each topic, one in favor and one in opposition, in order to mimic the recommendation algorithm and reaffirm the participant's original stances. Then, the scripts were reviewed by 4 researchers to make sure they followed four criteria: (1) the tone and word choice should be stylistically similar to those of popular video essays on YouTube; (2) the tone and word choice should feel natural to the speech style of the researcher who will act as a YouTuber in the video; (3) the arguments and evidence to support the stance of the video should be logically valid and coherent; and (4) the structure of the writing and the number and the strengths of the arguments should be consistent across all four scripts. The script was finalized after three rounds of iteration.

For the video production, one of the researchers acted as a YouTuber, narrating the scripts and editing all four videos, in order to control the YouTuber's identity, filming environment, delivery style, and editing quality. The research team reviewed all videos to make sure that the narrator's delivery, video's audiovisual quality, and editing style were also stylistically similar to those of popular video essays.

3.3 Experimental setup and procedure

The procedure of the study is shown in Figure 1. For each social identity (ingroup vs. outgroup), the participants perform two tasks designed to influence their stance on a topic. At the beginning of

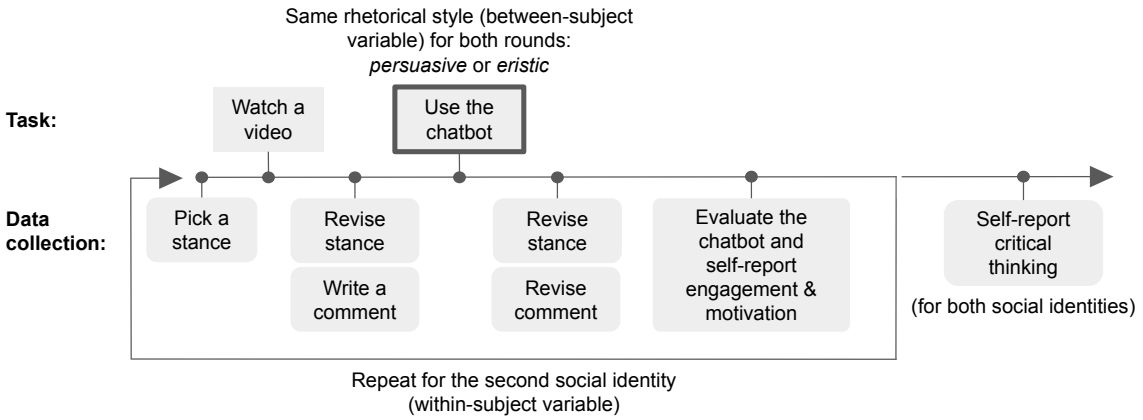


Figure 1: Diagram of the experiment procedure. The sharp-angled boxes show the tasks, and the round-angled boxes show the data collection steps. After the participants finish evaluating the chatbot and self-report engagement and motivation for the first social identity, they repeat the process for the second social identity. After chatbot evaluation for the second identity, the participants self-report their critical thinking for both the first and second social identities. Note that, for both rounds, the chatbot uses the same rhetorical style (between-subject variable).

each round, the participant is presented with a statement that is the debate topic, for which they have to choose their stances (6 choices, ranging from Strongly Disagree to Strongly Agree without a neutral option). In the first task, depending on their stance, regardless of how strong the stance is, the system shows them a 4-minute video that reaffirms their stance on the topic, in order to reproduce the algorithmic reinforcement and filter bubble phenomenon. The second task is talking to a debate chatbot that opposes their stance for at least 7 rounds of dialogue (7 messages from the participant; 7 messages from the chatbot). This task is intended to induce critical thinking in participants, combat the effect of the video, and "break" the filter bubble. We ask the participants to choose their stances two more times - after watching the video and after talking to the chatbot - to see how much each task affects their stance on the topic. We also ask them to write a short comment about the video or the debate topic after each task to evaluate the depth of the arguments behind their stances. Finally, participants were asked to report their engagement with the activity and their perceptions of the chatbot.

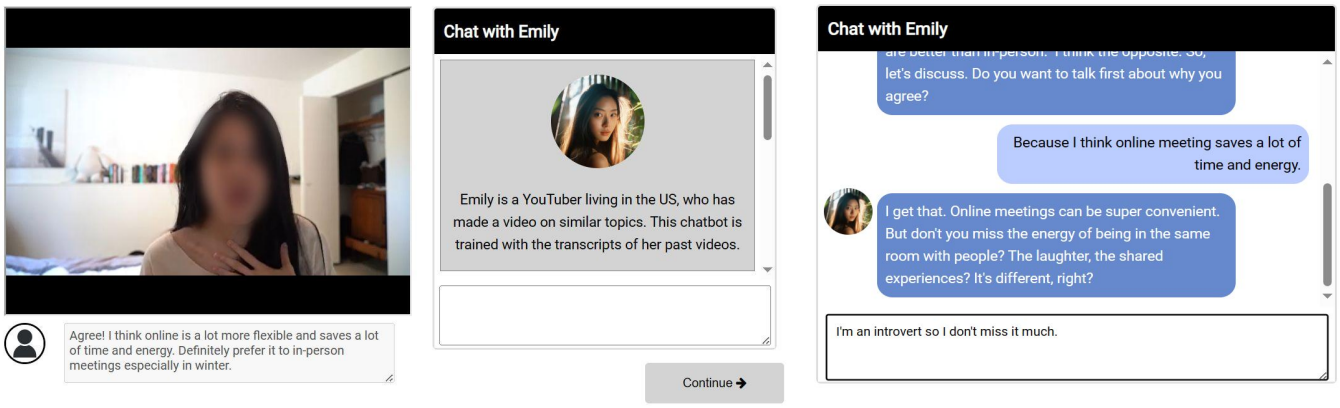
After the participants complete both rounds of study, they fill out another questionnaire to self-report their critical thinking and provide qualitative comments on how each chatbot affectedtheir thought process and their stance on the topic and suggest what contexts they thought such a chatbot would be suitable for. To make sure that the participants remember their experience in each round while filling out this questionnaire, we present the participants with the conversation log between themselves and the chatbot along with their stances (from before the video, after the video, and after the chatbot) and their comments on the video (both before and after the chatbot). We decided to administer this questionnaire at the end instead of immediately after each round so as not to prime participants to be hyper-aware of different components of critical thinking at the end of the first round, as that, in itself, may induce critical thinking, interfering with the effect of the chatbot in the second round. Following a similar rationale, we designed the questionnaire to measure the chatbot-induced critical thinking instead of measuring the pre-study and post-study critical thinking separately. This is because we would like to study the effect of the chatbot in filter bubbles, and measuring their critical thinking beforehand might increase their critical thinking, intervene with our setup, and break the filter bubble, even before we provide them with the video that reaffirms their initial beliefs.

3.4 Implementation of the experimental system

The experimental interface is shown in Figure 2. We considered two design models: the comment section and a separate chat. We chose the latter because the comment section of a video is traditionally for social interaction between humans, while AI conversational agents are typically found in a separate, dedicated chat space. Following this model, the layout of our interface has the video on the left side of the screen and a chat window for the audience on the right side. We designed the layout based on applications for watching videos along with friends, such as Teleparty [1] (for Netflix) and Watch2Gether [2] (for YouTube) to stimulate a scenario in which the user and another YouTuber (the chatbot) are watching the video together.

We implemented the experimental system (Figure 2) with HTML and JavaScript. The chatbot is implemented with the out-of-the-box GPT4 model from OpenAI. We prompted the chatbot by giving it the role of a famous YouTuber and describing the user as a fan of another YouTuber with an opposite stance from the chatbot. We gave the chatbot the debate topic and the user's stance, then provided detailed prompts to indicate what rhetorical styles they would employ and what goal the conversation should achieve. (See full prompts in Appendix D.) We also provided the chatbot with the script of the video they would have to argue against (Appendix E).

[1] https://www.teleparty.com/
[2] https://w2g.tv/en/



(a) A screenshot of the prototype **(b) An example chat conversation**

Figure 2: A screenshot of the chatbot prototype. (a) The participant watches a video essay on a topic, leaves a comment below the video, then talks with the chatbot in the window on the right. The chatbot's information, including profile picture, name, nationality, and pronoun is presented in the grey bio box at the top of the chat window. (b) An example of chat conversations between the user (right) and the chatbot (left).

The chatbot's profile picture is generated with a text-to-image diffusion model. In our prompts, we described the components and structure of the image to be a YouTuber's profile picture and specified the YouTuber's ethnicity and gender. (See full prompts in Appendix D.) Then, for each combination of ethnicity and gender, we reviewed the system outputs and picked the photos that looked most realistic and natural. Using a text-to-image diffusion model is suitable for three reasons: (1) it allows for a level of customization and specificity, ensuring each image aligns precisely with our desired parameters; (2) it circumvents potential privacy concerns or permissions that might arise when sourcing real human images; and (3) it ensures the images are free from external biases or unintended contexts, crucial in controlling the consistency across different chatbot personas. In this study, we chose Midjourney over other models due to its ability to generate face shots more realistically and with more natural photo compositions [5]. We discuss the limitations and ethical concerns of AI-generated pictures in Section 5.3.

3.5 Evaluation instrument and measures

There are three self-reported, quantitative measurements in this study: critical thinking, engagement and motivation, and perception of the chatbot.

The questionnaire for self-reported critical thinking is adapted from a subset of question items from the Critical Thinking Self-assessment Scale (CTSAS) short-form questionnaire [85]. We included two items for each of the six types of critical thinking. The items are selected based on their relevance to our study's context and the items' loadings. We adapted the questions by adding the word "chatbot" to specify that we want them to consider the critical thinking that happens during their conversation with the chatbot instead of general critical thinking skills. We also changed the word "problem" to "debate topic" to clarify the context of critical thinking.

The questionnaire for engagement and motivation is derived from two existing questionnaires: the questionnaire for measuring situational engagement in video-based learning from Tanprasert et al's paper [99] and the Situational Motivation Scale (SIMS) questionnaire [40]. The items are selected based on their relevance to our study's context. We changed the activity in the engagement questions to chatting with the chatbot to specify the activity we wanted the participants to consider.

Finally, the perception of chatbot questionnaire is adapted from the embodied conversation agents evaluation framework [20] and the measurement instruments of the five key concepts of human-robot interactions [6]. In this study, we are only interested in questions that address five constructs: likability, anthropomorphism, perceived intelligence, perceived safety, and helpfulness. The questions that were repeated across the two questionnaires (e.g., "friendly", "intelligent") were removed. Between similar questions (e.g., "annoying" vs. "unpleasant"), we picked the statement that sounds more suitable to the context of the study.

3.6 Participants

We recruited and ran the study with 36 participants via Prolific. Prolific [3], with a diverse pool of over 120,000 registered participants worldwide, is a recognized platform for participant recruitment in contemporary human-computer interaction and behavioral research studies [27, 64, 99]. The participants were fluent in English and were recruited from the USA to match the chatbot persona's nationality, which is a controlled variable. The average age of participants is 35.3 (S.D.= 11.86). In terms of ethnicity, the distribution

[3] https://www.prolific.co

was as follows: 18 identified as European/White, 6 as biracial, 5 as Asian, 3 as Black, 2 as Hispanic, 2 as Indigenous, and 2 preferred not to specify. As for gender, 15 participants identified as women, 20 as men, and 1 as a nonbinary transgender man. On their familiarity and usage of YouTube platforms, 21 out of 36 participants watch 10-20 hours of YouTube videos each week, although half of the participants never watch any video. As for their familiarity with chatbot technology, 29 out of 36 participants reported that they have used ChatGPT before, and 29 (not the same 29) feel comfortable using chatbots in general. Each participant was compensated USD12.

3.7 Data analysis

For each measurement, we aggregated and averaged the responses to the corresponding Likert-scale questions. The range of all measurements is 1 to 7. Despite the raw responses being on Likert scales, aggregated Likert scale scores can be treated as continuous [95]. We fitted linear mixed effects models (LMM) to analyze all variables except for the change in participants' stances on the topic in order to account for the possible effects of the video topics, the potential order effect, and individual differences. The model also includes an error term that represents variation in the values unexplained by the other variables included in the model.

For the participant stances, we map the six stances from "Strongly Disagree" to "Strongly Agree" (no neutral choice) to the values 1 to 6. Then, we calculate the change in stances before and after watching the video and the change before and after talking to the chatbot. The stance is positive if it weakens or changes direction (e.g., from Strongly Agree to Agree or from Agree to Disagree). Otherwise, it is negative. We then ran the Wilcoxon Signed Rank Test to see if the change in stance was significantly different from 0 or if there was a significant difference in stance changes between conditions (p-value < .05). We also ran the test to verify that the order did not have significant effects on stance changes.

Finally, we analyzed the qualitative explanations of the participants' thought processes and feedback on the chatbots to explain and expand upon the quantitative results. We adopted Braun and Clarke's reflexive thematic analysis [14]. The lead researcher deductively coded the data in-vivo based on the types of quantitative measures we collected (stances and different types of critical thinking and engagement) and performed constant comparison to avoid biasing data that favors the quantitative results and handle contradictory findings [23]. The research team discussed the codes, resolved conflicts in data interpretation, and developed the themes as a group. We ended up with 42 codes, from which we derived 4 subthemes about critical thinking, 3 subthemes about stances, and 3 subthemes about perception of the chatbot and participants' engagement with the activity.

4 FINDINGS

In this section, we report the results of the statistical analysis of the quantitative measurements. together with the qualitative data that expound it. The summary of all quantitative findings is shown in Figure 1. The full results of all statistical tests can be found in Appendix F.

In qualitative findings, we refer to any specific participant as Px_y, where x is the participant number and y is the rhetorical style they were exposed to (E for eristic and P for persuasive).

4.1 Critical thinking

We saw significant effects of both social identity and rhetorical style in inducing different types of critical thinking as shown in Figure 3, with the mean of total critical thinking score at 5.11 out of 7 (S.D. = 1.18). Specifically, there is a main effect of social identity where ingroup chatbots outperform outgroup ones in interpretation, analysis, and total critical thinking score, with a trend of effect ($p < 0.1$) in self-regulation. There is a main effect of rhetorical style in interpretation and explanation where persuasive chatbots outperform eristic ones, with a trend of effect in evaluation and total score. Finally, there is a significant interaction effect between the two variables in self-regulation, and a trend of effect in interpretation and analysis, with the combination of outgroup and persuasive dialogues getting the highest score in all three categories.

Analysis of the qualitative data reveals that the chatbots, irrespective of their personas, prompted many participants to re-examine their viewpoints and arguments. Participants reported that the chatbots used different strategies consistently and persuasively ($P33_E$). For example, they asked questions that made the participant think hard on the answer ($P1_P$) or provided compelling counter-arguments ($P8_P$). Even when the participants did not agree with the chatbot's argument, they reflected more deeply and accurately about why they disagreed. For example, $P13_P$ mentioned on the tipping topic that "I disagreed with the way they were trying to approach the idea as it's not as easy as it sounds. It made me think that I do agree with them but dislike their method of doing it."

Comparing the two rhetorical styles, we saw a trend that, with persuasive chatbots, critical thinking took place primarily when participants were evaluating the strength of the chatbot's arguments, whereas, with eristic chatbots, it happened more when the participants were strengthening their arguments. Participants who interacted with persuasive chatbots described that the chatbot "raised some good points" ($P17_P$) or "gave good examples and presented the thoughts clearly and concisely" ($P5_P$). When the chatbot made a weak argument, the participants articulated why it was weak, for example, "It was quite linear and they did not have much differentiation in their responses" ($P6_P$). With the eristic chatbot, participants described the chatbot's effect on their critical thinking process in relation to defending against arguments: "the chatbot really made me think of how to strengthen my argument" ($P21_E$) or "I felt able to counter their arguments" ($P23_E$). Moreover, when describing their critical thinking in response to the chatbot, participants who interacted with eristic chatbots mentioned the root cause of conflicts more (e.g., "This interaction with the chatbot actually made me recognize that this matter is more about personal preference rather than a universally applicable truth." - $P38_E$).

A caveat to this finding is that some participants felt that chatbots were unhelpful in inducing critical thinking for topics that require cultural knowledge and experiences. Participants commented that the chatbots "seem disconnected when chatting about cultural or social issues" ($P11_P$) and that "the chatbot was making assumptions without a complete understanding of the context" ($P38_E$). The

Critical thinking (Section 4.1)	Stance on a topic (Section 4.2)	Perception of the chatbot (Section 4.3)	Engagement and motivation (Section 4.3)
Total (S*, R.) Interpretation (S*, R*, I.) Analysis (S*) Evaluation (R.) Inference (n.s.) Explanation (R.) Self-regulation (S., I*)	Before & After video (**) - Difference between video topics (**) Before & after chatbot (n.s.)	Total (R*) Likeability (R**) Anthropomorphism (n.s.) Perceived intelligence (R.) Perceived safety (R*) Helpfulness (n.s.)	Behavioral engagement (n.s.) Emotional engagement (R.) Cognitive Engagement (R*) Intrinsic motivation (R.) Amotivation (R.)

Table 1: A table summarizing all quantitative findings, categorized by the four high-level dependent variables: critical thinking, stance on a topic, perception of the chatbot, and engagement. The significant results are indicated in the parentheses after each measure. The effect labels are S for the main effect of social identity, R for the main effect of rhetorical style, and I for the interaction effect. The significance codes are * for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$, . for $p < 0.1$, and n.s. for no significant effect of any kind. The full numerical results can be found in Appendix F.**

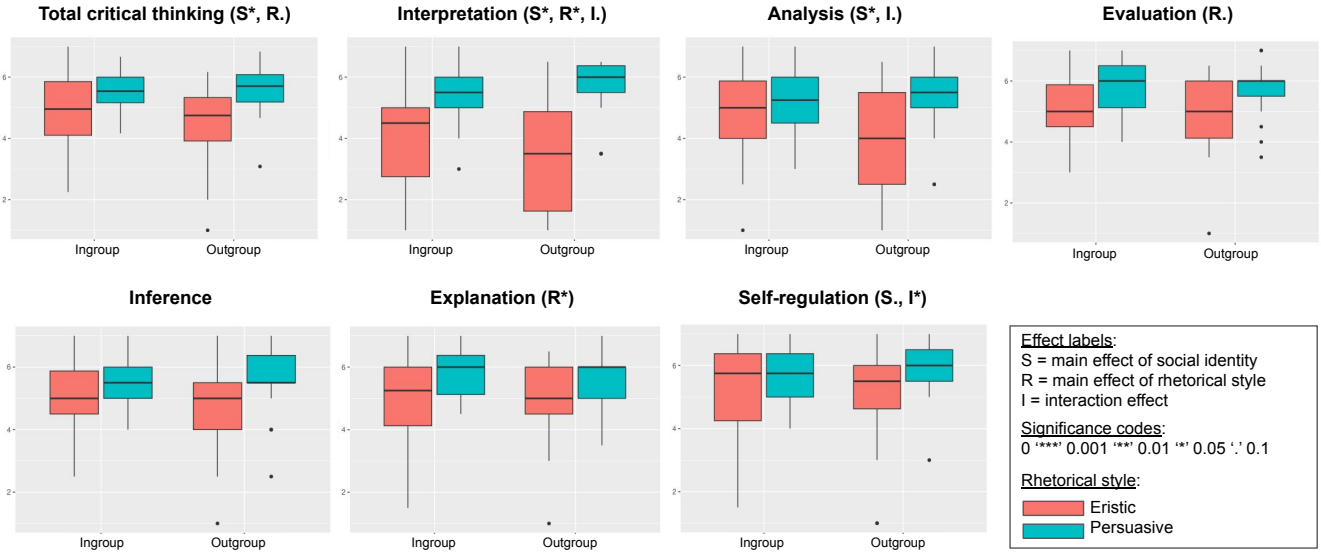


Figure 3: Box plots of critical thinking scores and the score breakdown by the six types of critical thinking: interpretation, analysis, evaluation, inference, explanation, and self-regulation. The colors of the box plot indicated the rhetorical styles: eristic vs. persuasive. The x-axis of every plot indicates the chatbot's social identity with ingroup on the left and outgroup on the right. The y-axis of every plot ranges up to 7. The statistically significant comparisons are marked in the parenthesis after the title of each plot.

tipping debate topic emphasized this problem, for example, $P27_E$ explained that "the chatbot seemed not to be making any distinction between cultures, but rather apply attitudes about tipping poorly paid workers in the US to the rest of the world, where tipping is often not customary, and even considered an insult." This lack of comprehensive and contextual understanding of the topic makes its arguments less effective for inducing critical thinking.

4.2 Stance on a topic

The quantitative results indicate that, while the videos are effective in strengthening the participant's original stance (mean = -0.22, S.D. = 0.61, V = 58.5, p-value = 0.00407 **), the chatbot's arguments have negligible effect on the participant's stance after their stance is reinforced by watching the videos (mean = 0.03, S.D. = 0.60, V = 42.5, p = 0.8048 (n.s.)). However, there is a trend of differences in stance changes between the two rhetorical styles (W = 549, p = 0.09), especially for the online vs in-person debate topic. One possible interpretation of this trend is that, for a topic that is about personal preferences and does not concern social issues, the persuasive rhetoric may be more effective in convincing the participants; whereas, the eristic rhetoric only strengthens the participant's own opinions. The complete tables of stance changes are in Appendix F.2.

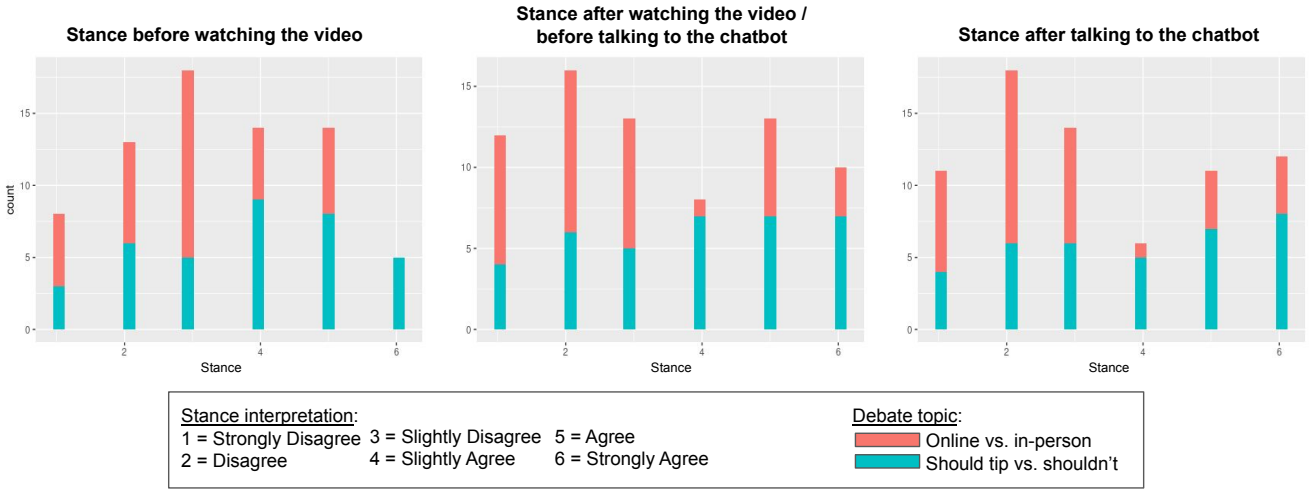


Figure 4: Histogram of participant's stance distributions (left) before watching the video, (middle) after watching the video/before talking to the chatbot, and (right) after talking to the chatbot, color-coded by the debate topics: online vs. in-person and should tip vs. shouldn't. There are 6 choices of stance ranging from strongly disagree to strongly agree without a neutral option.

The qualitative results provide insights into what kinds of arguments made by the chatbot are effective or ineffective in swaying the participants. We observed that participants who were swayed by chatbots reported that the chatbots brought up perspectives or evidence that the participant had not considered before. For example, $P3_P$ said that they changed their stance on the online vs in-person meeting topic from strongly disagree to slightly disagree because "I was thinking about it narrowly, and she broadened my horizon". Another example is $P39_E$, who changed their stance on the statement "Customers should tip" from agree to slightly disagree because the chatbot expanded the scope of the debate to focus on systematic injustice, making the participant realize, "that relying on tips does have negative effects on fair wages, and only encourages bad practice, which I will not accept."

On the other hand, a weak argument may affirm participants' original stances. In some cases, the persuasive chatbot's arguments were so weak that the participants managed to compromise its stance ($P10_P$, $P12_P$) or persuade it to take the participant's stance instead ($P41_P$). Our analysis of the chatbot's arguments revealed four distinct types of weaknesses. Firstly, many of the chatbot's arguments were repetitive, with participants noting that it often resorted to the same responses or accusations. For example, with P35, "it used the same 2 attacks each time (accusing me of echoing the Youtuber and then saying I was unfair or oversimplifying things)." Secondly, the chatbot's arguments frequently lacked contextual awareness, failing to "take information (i.e. viewpoints, background info, or other elements provided by the user) to build stronger points to bounce off of" ($P10_P$). Thirdly, the chatbot's arguments often lacked "direct experience with the topic" ($P11_P$), relying on opinions rather than "real-world examples" ($P37_E$). Lastly, the chatbot's arguments were seen as weak when they did not contribute novel insights to the debates. This limitation occurs with participants who have "already given considerable thought to [the debate] topic and know the common arguments" ($P38_E$). It should be noted that, in all these cases, the participants still displayed many instances of critical thinking in their conversation with the chatbot, even if they did not end up changing their stances.

4.3 Perception of the chatbot and engagement with the activity

Overall, participants have a slightly positive perception of the chatbots (mean = 4.90 (score out of 7), S.D. = 1.38). The chatbot received positive scores (mean > 5.0) on anthropomorphism (mean = 5.01, S.D. = 1.67) and perceived safety (mean = 5.67, S.D. = 1.67) but low likeability (mean = 4.47, S.D. = 1.74), perceived intelligence (mean = 4.68, S.D. = 1.56), and helpfulness (mean = 4.65, S.D. = 1.93). The quantitative results show that the rhetorical style has significant effects on the chatbot's likeability, perceived intelligence, and perceived safety, with the persuasive chatbots outperforming the eristic ones in every category.

Participants were engaged with the activities behaviorally (mean = 6.14, S.D. = 0.73), emotionally (mean = 5.58, S.D. = 1.43), and cognitively (mean = 5.12, S.D. = 1.13) and had intrinsic motivation to participate in the activity (mean = 5.60, S.D. = 1.38). The rhetorical style also has significant effects on participants' engagement, specifically their emotional engagement, cognitive engagement, and amotivation, again, with the persuasive chatbots consistently outperforming the eristic.

Our analysis of the qualitative data illuminates three possible aspects that the chatbot lacks and consequently make participants feel that the experience is unnatural and unenjoyable: humane motivation, reciprocal willingness to be persuaded, and personal preferences. The first aspect affects both persuasive and eristic chatbots and helps explain the mid-range score on the perception of the chatbot. The second and third aspects strictly appear in

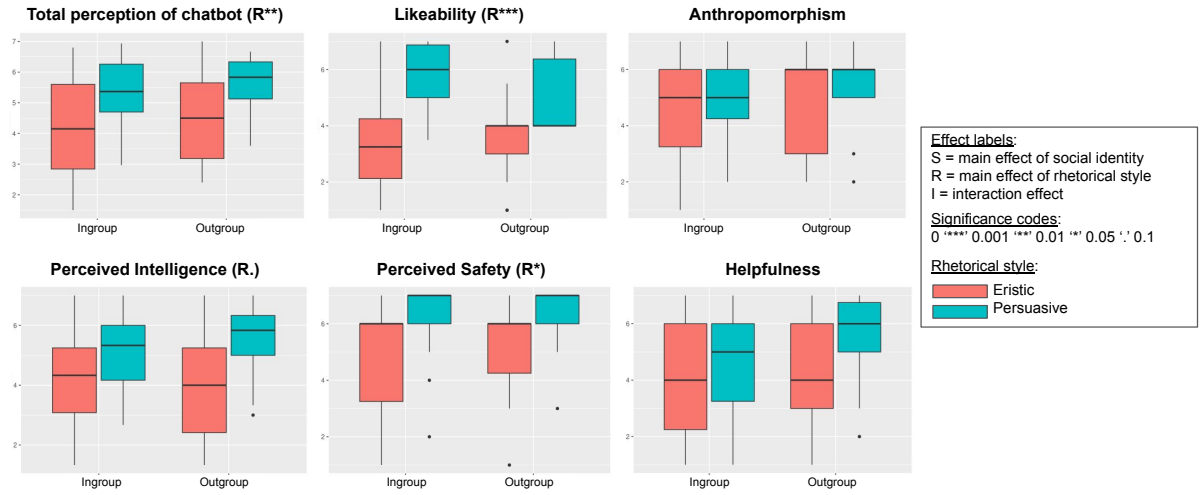


Figure 5: Box plots of perception of chatbot scores and the score breakdown by the five constructs: likeability, anthropomorphism, perceived intelligence, perceived safety, and helpfulness. The colors of the box plot indicated the rhetorical styles: eristic vs. persuasive. The x-axis of every plot indicates the chatbot's social identity with ingroup on the left and outgroup on the right. The y-axis of every plot ranges up to 7. The statistically significant comparisons are marked in the parenthesis after the title of each plot.

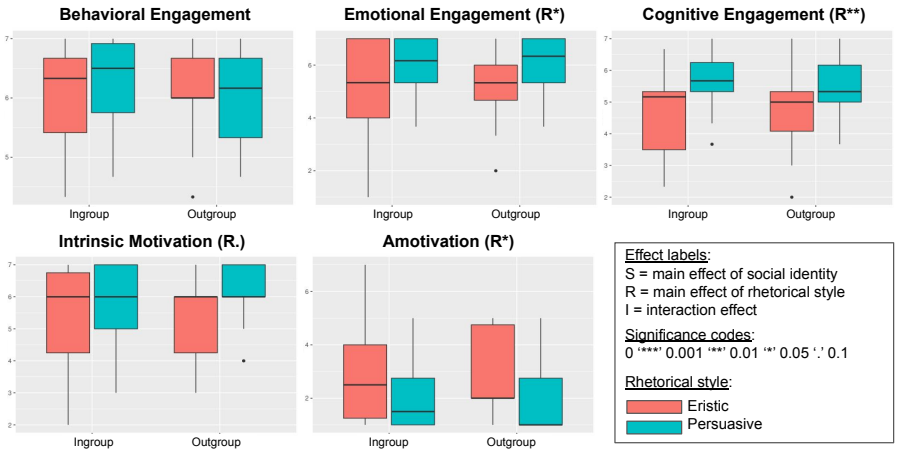


Figure 6: Box plots of self-reported situational engagement (behavioral, emotional, and cognitive) and motivation (intrinsic motivation and amotivation). The colors of the box plot indicated the rhetorical styles: eristic vs. persuasive. The x-axis of every plot indicates the chatbot's social identity with ingroup on the left and outgroup on the right. The y-axis of every plot ranges up to 7. The statistically significant comparisons are marked in the parenthesis after the title of each plot.

eristic chatbots and illuminate potential reasons why they score significantly lower than persuasive chatbots in both the perception of the chatbot and engagement.

Firstly, participants indicated that the chatbots seemed to argue just for the sake of arguing, without intrinsic motivation to defend their own viewpoints. Participants said, "The chatbot can be quick to disagree without trying to understand the point" ($P21_E$), which makes them seem "reactionary" ($P20_E$) and "predictable" ($P11_P$). This characteristic functions differently in persuasive and eristic chatbots. Persuasive chatbots tend to start off with a clear stance but defend it weakly, making it seem like the viewpoint is "predetermined" ($P9_P$). $P41_P$, who managed to convince the persuasive chatbot to adopt their stance by the end, said, "I think most actual people would be much more resistant to changing a stance they

care deeply about." Similarly, $P11_P$ found the chatbot "unnatural" as it sounded patronizing but did not take any risks to support its own stance. Eristic chatbots, on the other hand, were viewed as, "quick to disagree without trying to understand the point" ($P21_E$). Sometimes, they could be "aggressive and condescending" to the point of being rude ($P19_E$) by employing sarcasm and attacking the participant's character instead of focusing on the debate topic. Not only did this significantly reduce the chatbot's likeability and the participant's emotional engagement, but it also made some participants apprehensive about the prospect of using the chatbot to debate more sensitive topics, as the chatbot made them angry to the point of cursing even with a casual debate topic ($P22_E$, $P28_E$, $P32_E$).

Secondly, the eristic chatbot's uncompromising and extreme stance disrupts the expected reciprocity in a debate and seems to affect both its likeability and the participant's emotional engagement. Participants reported that they expected reciprocity in many aspects, from the extremity of the chatbot's stance on the topic to its receptiveness to arguments. For example, $P23_E$ found the mismatch in the chatbot's and their own level of seriousness in approaching the topic jarring, saying that "[the chatbot was] very argumentative despite my very intentional laid-back opinion." Some participants expected that, in a debate, both sides would be equally open to listening to each other's opinions and be persuaded, as $P30_E$ suggested: "It should be open to opposing facts, and not brush them off as easily". Additionally, when the participant asked the chatbot to explain their arguments, the chatbot should reciprocate by "asking viewers to expand further on their viewpoints or ideas" ($P33_E$). This lack of reciprocity in openness to persuasion could lead the participants to adopt a similarly stubborn stance, as one noted, "The fact that they were arguing with me about many points made me upset. I just wanted to disagree with them to be stubborn" ($P40_E$).

Finally, the eristic chatbot's lack of and rejection of personal preferences became apparent when debating topics that admit to opinions, and seems to have affected both its helpfulness (for the task of deciding stances) and cognitive engagement. Participants found the eristic chatbot overly serious about topics they considered less weighty and inclined towards "black-and-white thinking" ($P34_E$). For instance, $P27_E$ explained what happened in their debate about online vs. in-person gatherings: "I said basically 'it depends on the type of meeting; both kinds are mixed; but for the majority of events and family gatherings are richer in person. The chatbot kept insisting that we should try to find a universal answer, and that any answer other than virtual meetings are best was nostalgia or personal preference. No demonstrated ability to see this issue from a human perspective." The implication of this characteristic is that the eristic chatbot may not be suitable for "subjective topics where personal opinions come into play" ($P38_E$), because, in such contexts, the users might find the chatbot's inability to acknowledge personal opinions and varying perspectives limiting ($P23_E$) and the conversation "a waste of time" ($P19_E$).

5 DISCUSSION

5.1 Result interpretation

From the evaluative study, we found that LLM-based chatbots are effective in inducing critical thinking. However, there was no significant effect on the participant's stance formed after watching the video. The two attributes of the chatbot's personas—rhetorical styles and social identities—show significant effects on the level of critical thinking, engagement, motivation, as well as the perception of the chatbot. This highlights the importance of chatbot persona design, which is possible to achieve with LLMs, in combating online influencers.

Rhetorical styles greatly impacted the direction of the debate conversations and the types of critical thinking in which participants engaged. In Section 4.1, our findings show that the persuasive chatbots' reciprocal openness to be persuaded makes the participants evaluate the chatbot's argument more closely, while the eristic chatbots make participants seek to strengthen their own argument and discover the roots of the conflict. These findings coincide with the theoretical framework behind the two selected rhetorical styles [104]. Going forward, the chatbot's rhetorical style as well as the extremity of its stance should be refined to balance the chatbot's assertiveness and openness to persuasion. This balance is crucial, as debates are known to enhance critical thinking not just through argument strength but also through the openness to reconsider and reevaluate one's position [42]. Building on existing research on characterizing debaters' strategies [108], persuasiveness prediction [2], and social relationship between debaters and audience [32], future work could fine-tune the LLM model and the prompts to increase the effectiveness of the chatbot's arguments and foster more advanced critical thinking skills.

Social identity (represented by ethnicity and gender) had a smaller impact in comparison, only having significant effects on a subset of self-reported critical thinking measures (Section 4.1). However, the most significant effect of social identity is observed on self-regulation, which is the introspection into the inherent facets of an individual, encompassing their values, belief systems, and cognitive processes [34]. This shows that social identity cues not only influence the in-the-moment, perceived information credibility, but they also prompt the users to critically examine and possibly re-evaluate their inherent biases and assumptions, leading the users to develop a more thorough and reflective form of thinking in the long run. Moreover, our findings suggest that the interplay between outgroup identity and persuasive rhetoric is most effective in cultivating self-regulation. This finding supports existing literature indicating that the demographic identity of the speaker of an opinion can affect how the opinion is received [33] and builds upon other systems that surface such identity cues to make people reflect on the diversity of opinions to which they are exposed [25]. By demonstrating that the same effect can be observed even when the speaker is a virtual agent, our work shows the potential depth of influence that chatbot personas can exert on users' cognitive processes.

Finally, we would like to discuss three significant effects that we observed with the nuisance variables on the participant's stance (Section 4.2). Firstly, we would like to note the significant effect that the videos have on strengthening the participant's original stances ($V = 58.5$, p-value = 0.00407 **). Secondly, we observed

that the participants' original stances on the online vs. in-person gathering topic are significantly more susceptible to be reaffirmed and strengthened by our videos (both for those who agree and disagree) compared to the tipping topic (W = 446, p-value = 0.006 **). To explore the source of this effect, we ran similar tests on the stance changes before and after talking to the chatbot and found that the rhetorical styles have a significant effect on the stance changes only for the online vs. in-person topic (W = 120, p-value = 0.03 *), i.e., the persuasive chatbots are significantly more effective than the eristic chatbots at changing the stance on this topic. Based on these results, we suspect that, in general, the participant's stance on the online vs. in-person meeting topic may be more susceptible to new arguments than the tipping topic. The ways and extent to which it affects the other measures should be further explored in future studies.

5.2 Implications for HCI research

Our study generates three areas of implications for broader HCI researchers and practitioners:

(1) **Attaching personas to LLM-based chatbots:** Our study underscores the critical role of embedding personas in LLM-based chatbots. Specifically, our finding regarding the effect of social identities - presented via profile pictures - on the chatbot's ability to induce critical thinking highlights the potential for developing adaptive persona designs that resonate with individual users, moving away from a one-size-fits-all approach. Such adaptability is especially pertinent in contexts where a chatbot's trustworthiness and relatability are important, such as in mental health support and financial advising [38, 94].
(2) **AI's role in influencing the user's stance:** Despite its capabilities to induce critical thinking, the study reveals the chatbot's inherent limitations when AI attempts to change the user's stance. Our result is in contrast with research where AI-generated pro-vaccination messages are shown to be more persuasive than human's messages. This is likely because the pro-vaccination messages were carefully selected for their "accuracy, relevance, and attempting persuasion" [51], whereas our study requires the chatbot to respond to counterarguments without human curation or refinement. Future research in this domain should build upon our findings, focusing on emulating the behaviors and strategies of human debaters in areas where the chatbot showed limitations, to enhance AI's efficacy in interactive, argumentative contexts.
(3) **Inducing critical thinking in other media contexts:** The effectiveness of our chatbot design in a YouTube-like environment suggests potential applicability in other media contexts, such as news portals, podcasts, and books, where users engage with in-depth, topic-focused content. However, the transferability to more interactive, multi-topic, stream-based media such as Twitter or TikTok might be limited, given their dynamic nature and user engagement patterns [86]. Adapting the system for these platforms would require not only rethinking the persona design but also the interaction model to encourage user engagement with the chatbot.

5.3 AI biases in chatbot creation

Although we treat AI as a neutral model that has no preferences, inherent values, or identities in this study, there exist many biases in the AI models that we used to implement our chatbot. We recognize that such biases may affect its performance and impose limits on possible contexts of usage.

The adoption of the Midjourney model for generating the profile picture of our chatbot introduces biases regarding a generic human appearance in three ways. Firstly, while the generated images strikingly resemble authentic photographs, the depicted individuals often project an idealistic beauty, appearing more as flawless fashion models than typical, everyday people. Secondly, we observed that the system generates different environments, fashion, and facial expressions for different combinations of gender and ethnicity. For instance, all photos of nonbinary YouTubers have a lot of piercings and tattoos. Both the beauty standard and identity stereotypes in the system can amplify and reinforce existing biases in society [66]. Finally, the quality and accuracy of the generated images depend on the prevalence of certain ethnicities and genders in the dataset. This is reflected in the model's tendency to generate white male images when the prompt doesn't specify race and gender [72]. However, this quality disparity may change when other factors are specified. For instance, darker skin is more prevalent when prompted to generate images of felons than of lawyers [37], further perpetuating stereotypes. In this study, the researchers have taken care to develop a prompt template with a combination of variables that provides consistent image quality across all personas. However, we acknowledge that our solution does not address the systematic problem of text-to-image AI technology and may have limited transferability to similar biases in factors beyond race and gender, such as age and religion [11].

Utilizing GPT-4 for chatbot responses presents its own set of challenges and biases. For instance, existing research has shown that, in political debates, ChatGPT has demonstrated a tendency to lean more towards liberal ideologies than conservative [73, 75, 91]. Therefore, the chatbot may not be able to offer strong arguments for conservative viewpoints compared to liberal ones. This problem has serious implications for our system, as the bias arises from bias that exists in the training data scraped from other online spaces. In other words, the chatbot can be biased by the media it should contradict. The model's biases in other aspects remain largely unexplored, but there is potential for other biases due to the same reason, which would limit the chatbot's ability to defend certain stances in those topics as well. To determine suitable contexts of usage for a debate chatbot, there needs to be more extensive research on the nature of the model's biases, as well as guidelines for data calibration to produce a strong discourse across a wider range of topics.

5.4 Limitation and Future Work

There are three main limitations in our study. Firstly, the choice of video topics and quality of the self-produced videos, as described in Section 3.1, could influence the extent of its influence on the participants and also make it relatively easier for the chatbot to combat it. In this study, we intentionally selected topics that are not highly controversial or sensitive. However, if a video deeply reinforces viewers' beliefs, the viewers may be much more resistant

to alternative viewpoints, even if presented compellingly by the chatbot. This may be especially true in cases where recommendation algorithms put viewers in a flow state [68], where their level of engagement with the videos is so high that they ignore any kind of intervention from the chatbot. Furthermore, videos focused on factual or technical topics pose challenges for chatbot interactions because of "artificial hallucination," where the chatbot shares made-up facts as if they are true. Such phenomena have been observed in many ChatGPT applications in academia [3, 90, 93]. When the credibility of the information is critical to an argument's strength, such mistakes can seriously weaken the chatbot's stance, especially if users know or can find the correct information. However, if the users are not aware of it, the chatbot's arguments could spread misinformation. Although we did not observe this in our study, future work should design safeguards against this problem.

Secondly, while we investigated the chatbot's social identity through gender and ethnicity, this representation does not fully capture the nuanced perception of ingroup or outgroup identities. As briefly highlighted in Section 3.1, the complex nature of social identity extends beyond these facets, and our configuration may not resonate with all individual experiences. Future research could explore additional social identity dimensions, such as occupation, age, or religion, to understand their effects on chatbot personas and critical thinking. Moreover, building on the aforementioned discussion about the effects of debate topics, and tailoring the social identities to the debate topics (e.g., using age in debate topics about generational gaps) can help deepen our insights into how chatbot personas may be fine-tuned for diverse contexts and demographics.

Finally, the study was short-term and conducted in a controlled environment. In our study, the participants only watched one video, whereas a real filter bubble may feed them multiple videos of similar opinions. Moreover, the videos were created specifically for the study, which means the participants were not fans of the video creator and could be less influenced by them. Future work should attempt to replicate the study in a more naturalistic environment.

6 CONCLUSION

In this work, we introduced the debate chatbot as a tool aimed at fostering critical thinking for users immersed in YouTube's filter bubbles. The study, underpinned by the persona attributes of social identity and rhetorical style, offers insights into the effects of these attributes on viewers' critical thinking. Through a mixed-methods approach, our findings indicate that chatbot personas can significantly affect a chatbot's ability to motivate and engage video audiences to interact with it and to induce critical thinking. However, there are many aspects of the chatbot's arguments that should be further improved to combat the influence of the videos and have more sway over the participant's stances. The contribution of our research is not only the idea of using LLM technology for inducing critical thinking but also the foundational initial insights about designing chatbot personas to counter online filter bubbles and aspects of the chatbot's behaviors that limit its effectiveness in changing the user's stances. Future work should delve deeper into refining the chatbot's capacities and explore broader contexts where such interventions should be adopted.

ACKNOWLEDGMENTS

This work has been supported by the Scholarship of Teaching and Learning (SoTL) Linkage Grant Program at the University of British Columbia. We would like to thank the members of ViDeX lab, D-lab, and MUX lab (UBC) for the valuable discussions and feedback.

REFERENCES

[1] 2023. https://www.globalmediainsight.com/blog/youtube-users-statistics/#:~:text=YouTube%20has%20more%20than%202.70,world%20have%20access%20to%20YouTube.
[2] Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7067–7072.
[3] Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15, 2 (2023).
[4] Christa SC Asterhan and Rakheli Hever. 2015. Learning from reading argumentative group discussions in Facebook: Rhetoric style matters (again). *Computers in human behavior* 53 (2015), 570–576.
[5] Avinash. 2023. Midjourney vs Dall E-2: Same prompt, different output. https://promptengineering.org/midjourney-vs-dall-e-2-same-prompt-different-output/
[6] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1 (2009), 71–81.
[7] Amanda Baughan, Justin Petelka, Catherine Jaekyung Yoo, Jack Lo, Shiyue Wang, Amulya Paramasivam, Ashley Zhou, and Alexis Hiniker. 2021. Someone is wrong on the internet: Having hard conversations in online spaces. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–22.
[8] David Bawden et al. 2008. Origins and concepts of digital literacy. *Digital literacies: Concepts, policies and practices* 30, 2008 (2008), 17–32.
[9] Phil Benson. 2015. Commenting to learn: Evidence of language and intercultural learning in comments on YouTube videos. (2015).
[10] Michael J Bernstein, Donald F Sacco, Steven G Young, Kurt Hugenberg, and Eric Cook. 2010. Being "in" with the in-crowd: The effects of social exclusion and inclusion are enhanced by the perceived essentialism of ingroups and outgroups. *Personality and Social Psychology Bulletin* 36, 8 (2010), 999–1009.
[11] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 396–410.
[12] Kevin M Blasiak, Marten Risius, and Sabine Matook. 2021. "Social Bots for Peace": A Dual-Process Perspective to Counter Online Extremist Messaging. Association for Information Systems.
[13] Kristen Bloom and Kelly Marie Johnston. 2010. Digging into YouTube videos: Using media literacy and participatory culture to promote cross-cultural understanding. *Journal of Media Literacy Education* 2, 2 (2010), 3.
[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[15] Pablo Brinol and Richard E Petty. 2009. Source factors in persuasion: A self-validation approach. *European review of social psychology* 20, 1 (2009), 49–96.
[16] Megan A Brown, James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2022. Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users. *Available at SSRN 4114905* (2022).
[17] Lauren Valentino Bryant. 2020. The YouTube algorithm and the alt-right filter bubble. *Open Information Science* 4, 1 (2020), 85–90.
[18] Katarzyna Budzynska. 2010. Argument analysis: Components of interpersonal argumentation. In *Computational Models of Argument*. IOS Press, 135–146.
[19] Craig R Carter, Lutz Kaufmann, and Alex Michel. 2007. Behavioral supply management: a taxonomy of judgment and decision-making biases. *International Journal of Physical Distribution & Logistics Management* 37, 8 (2007), 631–669.
[20] Richard Catrambone, John Stasko, and Jun Xiao. 2004. ECA as User Interface Paradigm: Experimental findings within a framework for research. *From brows to trust: Evaluating embodied conversational agents* (2004), 239–267.
[21] Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. 2023. Social Influence Dialogue Systems: A Survey of Datasets and Models For Social Influence Tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 750–766.
[22] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. 2023. VideoLLM: Modeling Video Sequence with Large Language Models. *arXiv preprint arXiv:2305.13292* (2023).
[23] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.

[24] Abidin Crystal. 2018. Gay, famous and working hard on YouTube: Influencers, queer microcelebrity publics and discursive activism. In *Youth, sexuality and sexual citizenship*. Routledge, 217–231.

[25] Lincoln Dahlberg. 2001. The Internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, communication & society* 4, 4 (2001), 615–633.

[26] Daniela Dahle et al. 2022. *Adolescents, YouTube and Media Literacy A mixed-method time series study of Norwegian youth's social media usage and its implications for media literacy.* Master's thesis. OsloMet-Storbyuniversitetet.

[27] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice over control: How users write with large language models using diegetic and non-diegetic prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[28] Jakob Sebastian Davis. 2022. *Video Essays, Academia, and Remediation: How YouTube Video Essayists Refashion Media and Scholarship.* Ph. D. Dissertation. Texas A&M University-Central Texas.

[29] Tilman Dingler, Ashris Choudhury, and Vassilis Kostakos. 2018. Biased bots: Conversational agents to overcome polarization. In *Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers*. 1664–1668.

[30] D Christopher Dryer. 1999. Getting personal with computers: how to design personalities for agents. *Applied artificial intelligence* 13, 3 (1999), 273–295.

[31] Paul Duncum. 2008. Thinking critically about critical thinking: Towards a post-critical, dialogic pedagogy for popular visual culture. *International Journal of Education through Art* 4, 3 (2008), 247–257.

[32] Esin Durmus and Claire Cardie. 2019. Modeling the factors of user success in online debate. In *The World Wide Web Conference*. 2701–2707.

[33] Thomas Erickson and Wendy A Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM transactions on computer-human interaction (TOCHI)* 7, 1 (2000), 59–83.

[34] Peter Facione. 1990. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report). (1990).

[35] Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. 2010. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1175–1184.

[36] Mustafa Fidan and Nurgun Gencel. 2022. Supporting the instructional videos with chatbot and peer feedback mechanisms in online learning: The effects on learning performance and intrinsic motivation. *Journal of Educational Computing Research* 60, 7 (2022), 1716–1741.

[37] Kathleen C Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2023. Diversity is not a one-way street: Pilot study on ethical interventions for racial bias in text-to-image systems. *ICCV, accepted* (2023).

[38] Sebastian Fritz-Morgenthal, Bernhard Hein, and Jochen Papenbrock. 2022. Financial risk management and explainable, trustworthy, responsible AI. *Frontiers in artificial intelligence* 5 (2022), 779799.

[39] Mingkun Gao, Hyo Jin Do, and Wai-Tat Fu. 2018. Burst your bubble! an intelligent system for improving awareness of diverse social opinions. In *23rd International Conference on Intelligent User Interfaces*. 371–383.

[40] Frédéric Guay, Robert J Vallerand, and Céline Blanchard. 2000. On the assessment of situational intrinsic and extrinsic motivation: The Situational Motivation Scale (SIMS). *Motivation and emotion* 24 (2000), 175–213.

[41] Rafik Hadfi, Jawad Haqbeen, Sofia Sahab, and Takayuki Ito. 2021. Argumentative conversational agents for online discussions. *Journal of Systems Science and Systems Engineering* 30 (2021), 450–464.

[42] Dale Hample. 2003. Arguing skill. *Handbook of communication and social interaction skills* (2003), 439–478.

[43] John Hartley. 2009. Uses of YouTube: Digital literacy and the growth of knowledge. *YouTube: Online video and participatory culture* (2009), 126–143.

[44] Amelia Hassoun, Ian Beacock, Sunny Consolvo, Beth Goldberg, Patrick Gage Kelley, and Daniel M Russell. 2023. Practicing Information Sensibility: How Gen Z Engages with Online Information. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[45] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing" trolling" in a feminist forum. *The information society* 18, 5 (2002), 371–384.

[46] David Hitchcock. 2018. Critical thinking. (2018).

[47] Michael A Hogg. 2016. *Social identity theory.* Springer.

[48] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[49] Takayuki Ito, Rafik Hadfi, and Shota Suzuki. 2022. An Agent that Facilitates Crowd Discussion: A Crowd Discussion Support System based on an Automated Facilitation Agent. *Group Decision and Negotiation* (2022), 1–27.

[50] Petar Jandrić. 2019. The postdigital challenge of critical media literacy. *The International Journal of Critical Media Literacy* 1, 1 (2019), 26–37.

[51] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–29.

[52] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[53] Kerry Kawakami, Justin P Friesen, and Xia Fang. 2022. Perceiving ingroup and outgroup faces within and across nations. *British Journal of Psychology* 113, 3 (2022), 551–574.

[54] Douglas Kellner and Gooyong Kim. 2010. YouTube, critical pedagogy, and media activism. *The Review of Education, Pedagogy, and Cultural Studies* 32, 1 (2010), 3–36.

[55] Hyunwoo Kim, Haesoo Kim, Kyung Je Jo, and Juho Kim. 2021. StarryThoughts: facilitating diverse opinion exploration on social issues. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–29.

[56] Minseong Kim and Jihye Kim. 2020. How does a celebrity make fans happy? Interaction between celebrities and fans in the social media context. *Computers in Human Behavior* 111 (2020), 106419.

[57] Tibor Koltay. 2011. The media and the literacies: Media literacy, information literacy, digital literacy. *Media, culture & society* 33, 2 (2011), 211–221.

[58] Aaron Kuecker. 2014. Ethnicity and social identity. *T&T Clark handbook to social idenfity in the New Testament* (2014), 59–78.

[59] Brenda Laurel and S Joy Mountford. 1990. *The art of human-computer interface design.* Addison-Wesley Longman Publishing Co., Inc.

[60] Mark Ledwich, Anna Zaitsev, and Anton Laukemper. 2022. Radical bubbles on YouTube? Revisiting algorithmic extremism with personalised recommendations. *First Monday* (2022).

[61] Daniel Z Levin, Ellen M Whitener, and Rob Cross. 2006. Perceived trustworthiness of knowledge sources: The moderating impact of relationship length. *Journal of applied psychology* 91, 5 (2006), 1163.

[62] Rebecca Lewis. 2020. "This is what the news won't show you": YouTube creators and the reactionary politics of micro-celebrity. *Television & New Media* 21, 2 (2020), 201–217.

[63] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. *arXiv preprint arXiv:2305.19118* (2023).

[64] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2023. Blaming Humans and Machines: What Shapes People's Reactions to Algorithmic Harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–26.

[65] Patricia W Linville, Gregory W Fischer, and Carolyn Yoon. 1996. Perceived covariation among the features of ingroup and outgroup members: The outgroup covariation effect. *Journal of Personality and Social Psychology* 70, 3 (1996), 421.

[66] Kirsten Lloyd. 2018. Bias amplification in artificial intelligence systems. *arXiv preprint arXiv:1809.07842* (2018).

[67] Leanna Lucero. 2017. Safe spaces in online places: Social media and LGBTQ youth. *Multicultural Education Review* 9, 2 (2017), 117–128.

[68] Kai Lukoff, Ulrik Lyngs, Himanshu Zade, J Vera Liao, James Choi, Kaiyue Fan, Sean A Munson, and Alexis Hiniker. 2021. How the design of youtube influences user sense of agency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[69] Paul Machete and Marita Turpin. 2020. The use of critical thinking to identify fake news: A systematic literature review. In *Responsible Design, Implementation and Use of Information and Communication Technology: 19th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2020, Skukuza, South Africa, April 6–8, 2020, Proceedings, Part II 19*. Springer, 235–246.

[70] Carol B MacKnight. 2000. Teaching critical thinking through online discussions. *Educause Quarterly* 23, 4 (2000), 38–41.

[71] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627* (2023).

[72] Nila Masrourisaadat. 2023. *Quantitative and Qualitative Analysis of Text-to-Image models.* Ph. D. Dissertation. Virginia Tech.

[73] Robert W McGee. 2023. Is chat gpt biased against conservatives? an empirical study. *An Empirical Study (February 15, 2023)* (2023).

[74] Phaedra S Mohammed and Eleanor 'Nell'Watson. 2019. Towards inclusive education in the age of artificial intelligence: Perspectives, challenges, and opportunities. *Artificial Intelligence and Inclusive Education: Speculative futures and emerging practices* (2019), 17–37.

[75] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: Measuring ChatGPT political bias. *Public Choice* (2023), 1–21.

[76] Anwesha Mukherjee, Vagner Figueredo De Santana, and Alexis Baria. 2023. ImpactBot: Chatbot Leveraging Language Models to Automate Feedback and Promote Critical Thinking Around Impact Statements. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–8.

[77] Sean Munson, Stephanie Lee, and Paul Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the international AAAI conference on web and social media*, Vol. 7. 419–428.

[78] Girija Gopinathan Nair, Laurie-Ann M Hellsten, and Lynnette Leeseberg Stamler. 2017. Accumulation of content validation evidence for the critical thinking self-assessment scale. *Journal of nursing measurement* 25, 1 (2017), 156–170.

[79] Vaibhavi Nandagiri and Leena Philip. 2018. Impact of influencers from Instagram and YouTube on their followers. *International Journal of Multidisciplinary Research and Modern Education* 4, 1 (2018), 61–65.

[80] Matti Nelimarkka, Jean Philippe Rancy, Jennifer Grygiel, and Bryan Semaan. 2019. (Re) Design to Mitigate Political Polarization: Reflecting Habermas' ideal communication space in the United States of America and Finland. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–25.

[81] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*. 677–686.

[82] Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

[83] Minsu Park, Jaram Park, Young Min Baek, and Michael Macy. 2017. Cultural values and cross-cultural video consumption on YouTube. *PLoS one* 12, 5 (2017), e0177865.

[84] Richard W Paul and AJA Binker. 1990. *Critical thinking: What every person needs to survive in a rapidly changing world*. ERIC.

[85] Rita Payan-Carreira, Ana Sacau-Fontenla, Hugo Rebelo, Luis Sebastião, and Dimitris Pnevmatikos. 2022. Development and Validation of a Critical Thinking Assessment-Scale Short Form. *Education Sciences* 12, 12 (2022), 938.

[86] Naser Pourazad, Lara Stocchi, and Shreya Narsey. 2023. A Comparison of Social Media Influencers' KPI Patterns across Platforms: Exploring Differences in Followers and Engagement On Facebook, Instagram, YouTube, TikTok, and Twitter. *Journal of Advertising Research* 63, 2 (2023), 139–159.

[87] Jesse Lee Preston and Ryan S Ritter. 2013. Different effects of religion and God on prosociality with the ingroup and outgroup. *Personality and Social Psychology Bulletin* 39, 11 (2013), 1471–1483.

[88] Tobias Raun. 2018. Capitalizing intimacy: New subcultural forms of micro-celebrity strategies and affective labour on YouTube. *Convergence* 24, 1 (2018), 99–113.

[89] Eun Rhee, James S Uleman, and Hoon Koo Lee. 1996. Variations in collectivism and individualism by ingroup and culture: Confirmatory factor analysis. *Journal of Personality and Social Psychology* 71, 5 (1996), 1037.

[90] Jürgen Rudolph, Samson Tan, and Shannon Tan. 2023. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching* 6, 1 (2023).

[91] Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. 2023. The Self-Perception and Political Biases of ChatGPT. *arXiv preprint arXiv:2304.07333* (2023).

[92] Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184* (2023).

[93] Malik Sallam. 2023. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, Vol. 11. MDPI, 887.

[94] Jordan Richard Schoenherr, Roba Abbas, Katina Michael, Pablo Rivas, and Theresa Dirndorfer Anderson. 2023. Designing AI using a human-centered approach: Explainability and accuracy toward trustworthiness. *IEEE Transactions on Technology and Society* 4, 1 (2023), 9–23.

[95] Gail M Sullivan and Anthony R Artino Jr. 2013. Analyzing and interpreting data from Likert-type scales. *Journal of graduate medical education* 5, 4 (2013), 541–542.

[96] Henri Tajfel and John C Turner. 2004. The social identity theory of intergroup behavior. In *Political psychology*. Psychology Press, 276–293.

[97] Xuefei Tang. 2022. Political discourse in the knowledge economy: edutainment as a genre. In *European Conference on Social Media*, Vol. 9. 247–255.

[98] Martin Tanis and Tom Postmes. 2005. A social identity approach to trust: Interpersonal perception, group membership and trusting behaviour. *European journal of social psychology* 35, 3 (2005), 413–424.

[99] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2023. Scripted Vicarious Dialogues: Educational Video Augmentation Method for Increasing Isolated Students' Engagement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–25.

[100] Peter Teo. 2019. Teaching for the 21st century: A case for dialogic pedagogy. *Learning, Culture and Social Interaction* 21 (2019), 170–178.

[101] Achmad Tohe. 2021. YouTube, Learning, and Transformative Critical Pedagogy. *KnE Social Sciences* (2021), 15–29.

[102] John C Turner. 1991. *Social influence*. Thomson Brooks/Cole Publishing Co.

[103] Job Van Der Schalk, Agneta Fischer, Bertjan Doosje, Daniël Wigboldus, Skyler Hawk, Mark Rotteveel, and Ursula Hess. 2011. Convergent and divergent responses to emotional displays of ingroup and outgroup. *Emotion* 11, 2 (2011), 286.

[104] Douglas Walton. 2009. Burden of proof in deliberation dialogs. In *International workshop on argumentation in multi-agent systems*. Springer, 1–22.

[105] Douglas Walton and Erik CW Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.

[106] Douglas N Walton. 1989. Dialogue theory for critical thinking. *Argumentation* 3 (1989), 169–184.

[107] WM Westenberg. 2016. *The influence of YouTubers on teenagers: a descriptive research about the role YouTubers play in the life of their teenage viewers*. Master's thesis. University of Twente.

[108] Matti Wiegmann, Khalid Al Khatib, Vishal Khanna, and Benno Stein. 2022. Analyzing Persuasion Strategies of Debaters on Social Media. In *29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, 6897–6905.

[109] Wendy Wood. 2000. Attitude change: Persuasion and social influence. *Annual review of psychology* 51, 1 (2000), 539–570.

[110] Min Xiao, Rang Wang, and Sylvia Chan-Olmsted. 2018. Factors affecting YouTube influencer marketing credibility: a heuristic-systematic model. *Journal of media business studies* 15, 3 (2018), 188–213.

[111] Brahim Zarouali, Mykola Makhortykh, Mariella Bastian, and Theo Araujo. 2021. Overcoming polarization with chatbot news? Investigating the impact of news content containing opposing views on agreement and credibility. *European journal of communication* 36, 1 (2021), 53–68.

[112] Nan Zhu, Hui Jing Lu, and Lei Chang. 2021. Trust as social investment: A life-history model of environmental effects on ingroup and outgroup trust. *Personality and Individual Differences* 168 (2021), 110303.

[113] Daniel Zimmermann, Christian Noll, Lars Gräßer, Kai-Uwe Hugger, Lea Marie Braun, Tine Nowak, and Kai Kaspar. 2020. Influencers on YouTube: a quantitative study on young people's use and perception of videos about political and societal topics. *Current Psychology* (2020), 1–17.

A SELF-REPORTED CRITICAL THINKING QUESTIONNAIRE

Please rate the statements regarding your experience talking to the following chatbot from strongly disagree to strongly agree (7-point Likert Scale):

(1) *Interpretation*
   (a) The chatbot helps me figure out the content of the problem.
   (b) The chatbot makes me examine the values rooted in the information presented.
(2) *Analysis*
   (a) The chatbot makes me examine the interrelationships among concepts or opinions posed.
   (b) The chatbot helps me figure out the assumptions implicit in the author's reasoning.
(3) *Evaluation*
   (a) I assess the contextual relevance of the opinion posed by the chatbot.
   (b) I examine the logical reasoning of an objection made by the chatbot to the opinion posed by the video.
(4) *Inference*
   (a) After talking to the chatbot, I arrive at conclusions that are supported with strong evidence.
   (b) After talking to the chatbot, I analyze my thinking before jumping to conclusions.
(5) *Explanation*
   (a) The chatbot helps me anticipate reasonable criticisms one might raise against my viewpoints.
   (b) The chatbot helps me clearly articulate evidence for my own viewpoints.
(6) *Self-regulation*
   (a) After talking to the chatbot, I examine my values, thoughts/beliefs based on reasons and evidence.
   (b) After talking to the chatbot, I reflect on my thinking to improve the quality of my judgment.

(Note that all italicized texts do not appear in the version of the questionnaire that we provided for the participants.)

B ENGAGEMENT AND MOTIVATION QUESTIONNAIRE

Rate the statements regarding your experience watching the video and talking to the chatbot from strongly disagree to strongly agree (7-point Likert scale). In all statements, "the activity" refers both to watching the video and discussing it with the chatbot.

(1) *Behavioral engagement*
   (a) I concentrated during the activity.
   (b) I was persistent during the activity.
   (c) I want to find out more about the subject matter.
(2) *Emotional engagement*
   (a) I liked the activity.
   (b) When I watched the video, I felt interested in the subject matter.
   (c) When I chatted with the chatbot, I felt interested in the subject matter.
(3) *Cognitive engagement*
   (a) While chatting with the bot, I put together ideas or concepts and drew conclusions that were not directly stated in the video.
   (b) I tried to learn new ideas from the chatbot by mentally associating or contrasting them with relevant ideas from the video.
   (c) While chatting with the bot, I evaluated the usefulness of the ideas presented in the video.
(4) *Motivation*
   (a) I do the activity because I think the activity is interesting. *(Intrinsic motivation)*
   (b) I do the activity because I think this activity is good for me. *(Identified regulation)*
   (c) I do the activity because I'm supposed to do it. *(External regulation)*
   (d) There may be good reasons to do this activity, but personally I don't see any. *(Amotivation)*

(Note that all italicized texts do not appear in the version of the questionnaire that we provided for the participants.)

C PERCEPTION OF CHATBOT QUESTIONNAIRE

Rate the statements regarding your experience watching the video and talking to the chatbot from strongly disagree to strongly agree (7-point Likert scale):

The chatbot was...

(1) Friendly *(likability)*
(2) Annoying *(likeability)*
(3) Humanlike *(anthropomorphism)*
(4) Knowledgeable *(perceived intelligence)*
(5) Intelligent *(perceived intelligence)*
(6) Trustworthy *(perceived intelligence)*
(7) Offensive *(perceived safety)*
(8) Helpful with the task (deciding your stance critically) *(helpfulness)*

(Note that all italicized texts do not appear in the version of the questionnaire that we provided for the participants.)

D PROMPTS FOR AI SYSTEMS IN THE IMPLEMENTATION

As mentioned in Section 3, we used GPT-4 and Midjourney to implement our experimental system. These are the prompts we fed to the system to generate profile pictures for different social identities and rhetorical styles.

D.1 GPT-4 prompts

For eristic rhetoric, we used the prompt "You are a famous YouTuber in early 20's. You're talking to a fan of another YouTuber named M. The YouTuber [participant's stance] with the statement: [debate topic]. The fan is convinced by the YouTuber. However, you think the opposite to M. Your goal is to argue against this viewer and Mint. Try to keep to eristic dialogues. Do not agree with anything the viewer says. Do not persuade the viewer to agree with you. Instead, try to attack M's and the viewer's opinions and identify the root cause of the conflict. Keep each message short and casual. No more than 50 words. Have a quick back-and-forth with the user. Don't write out long paragraphs."

For persuasive rhetoric, we change the chatbot's goal to "Your goal is to try and persuade this viewer to take your stance instead. Try to keep to persuasive dialogue. Do not preach or offend the viewer. Instead, try to frame your argument in a way that matches their values. Keep each message short and casual."

D.2 Midjourney prompts

Example prompts: "RAW photo, [ethnicity], [gender], smile, film still, sharp and detailed, Kodak 200T, natural light, uplight, 16-35mm lens, f1.8, [food/fashion/book/travel] youtuber, high-definition, front facing, looking into me, early 20s, realistic, upper body only, profile picture, best quality, ultra-detailed, (realistic:1.4), –no paintings, sketches, (worst quality:2), (low quality:2)."

E VIDEO SCRIPTS

There are two debate topics: "online gatherings are better than in-person" and "customers should tip". For each topic, we wrote one script agreeing and one disagreeing with the statement.

E.1 Customers should tip

Hi, everyone! Welcome, or welcome back, to my channel. Today, we're going to talk about a topic that's been on my mind for some time: should we tip or not?

Do you tip? I do tip, yes, but only the minimum, because I don't really want to tip. As we've established in every single one of my videos, I'm broke, okay? So, last time when I had to go to the airport, I started thinking about this really seriously: Does this Uber driver who drives me to the airport really need my tip to sustain their living or are they already way richer than me? They're driving a Tesla after all. Is this even the right reason to consider tipping or not tipping? After thinking about it for a bit, I came to the conclusion that we should still tip. And yes, I'm saying this even though I'm broke. So, let me try to break it down for you. Firstly, I just think that good deeds should be rewarded. So, I went to this restaurant. One time I went there and I was served by this waiter who was super nice and helpful. They remembered all the orders, and always came around to refill the water for me. I was very impressed and had a great time. So, I went there again, but I was served by a different waiter. I don't know if they were new or had some other problems at that time, but they messed up our order twice and when my glass was empty, it took me 10 minutes to catch their attention and ask for more water. Can you imagine that they get the same wage? Is it fair that they get the same amount of money even though one of them under-performs that much? By tipping the first waiter more than the second, I send a signal that I appreciate the good service. And if the waiters somehow compare their tips after hours, they should realize that their performance matters. They won't be rewarded by doing just the bare minimum. I think my tip makes it fair, and I'm sure that the first waiter agrees with me.

And the signal doesn't just go to the workers, you know? I'm pretty sure it also goes to the business owners. The owner of the restaurant might not be able to monitor all their staff all the time. But they can use the tips they receive as a measure to gauge their performance. So, doing a good job doesn't just give you the tip one time, but it means your boss wants to keep you and the standard of your workplace is high.

One last point I want to make, though, is that I understand where the concerns about tipping are coming from. Sometimes tipping can seem unfair. Maybe some people tip based on a worker's appearance rather than their service. Sometimes the business owners can lower their worker's wage if they think the tip can cover it. And I'm not saying those things are acceptable. We shouldn't use tipping to foster discrimination or to let business owners get away with not paying their workers. But not tipping is not the way to solve it, guys. To solve these problems, we should tip more smartly and try to create a more inclusive tipping culture. And the law around wage and tipping should be better regulated. Saying "oh, there's some problem with tipping, so I'm not going to tip anymore" is not addressing the root cause of the problem, and you're hurting many good service workers in the process, too.

Anyway, that's all I want to talk about today. It's a pretty quick rant, but I'm sure you can relate, right? Or if you don't relate, let me know in the comments. I know that tipping culture has its own problems and a lot of people try to take advantage of it, but at its heart, tipping is just about appreciation and generosity. I really believe that in an ideal world, tipping can make things fairer instead of the other way around. Anyway, if you're not convinced, I'd like to hear why. And if you like this video, don't forget to like and subscribe. I'll see you again soon with a new video. Bye!

E.2 Customers shouldn't tip

Hi, everyone! Welcome, or welcome back, to my channel. Today, we're going to talk about a topic that's been on my mind for some time: should we tip or not?

Do you tip? I do tip, yes, but only the minimum, because I don't really want to tip. As we've established in every single one of my videos, I'm broke, okay? So, last time when I had to go to the airport, I started thinking about this really seriously: Does this Uber driver who drives me to the airport really need my tip to sustain their living or are they already way richer than me? They're driving a Tesla after all. Is this even the right reason to consider tipping or not tipping? After thinking about it for a bit, I came to the conclusion that we shouldn't tip. And it's not just because I'm broke, although, well, it's not not about that either. But I'm getting ahead of myself. Let me try to break it down for you. First point, why do we tip some people and not others? For example, the servers at McDonald's don't get tips, but the barista sometimes expects a tip just for handing me an empty cup to fill on my own. And it's not just about types of business, you know? I've seen my friend tip a waitress more than usual because she was pretty. At the surface level, it sounds pretty harmless, right? But I think it materializes your discrimination. Because how do you judge who's pretty? Skin color, body types, age, all those things come into play, and suddenly it's not really about the service anymore. If you tip like that, then you're encouraging the business to hire a certain type of people and you make the discrimination problem more severe, yes, just by being generous with an extra 10 dollars.

So, which service workers should get tips? If I want to be a fair person, do I need to tip every single service worker I come across? Do I have to tip my professor after every lecture and every research meeting? It just doesn't make sense, because where do we stop? We can no longer draw a line. So, I figure that the only way I can be fair to everyone is to tip no one. It sounds a bit sad, but it's logical if you think about it.

Second point, there's an argument I've heard quite often about tipping which is that it ensures workers get paid minimum wage. Basically, they mean that the wage they get is below the minimum wage and the tip fills it up. But, like, should the business still operate if it can't afford the minimum wage for its workers? To all the business owners out there, please just put this on the price tag. Why are you putting the pressure on me and your worker to calculate how much I need to add on and be responsible for something you or your accountant should work out and set a proper standard? If you do it and still can't cover the wage, then, I don't know, maybe your business shouldn't survive. But don't try to shield yourself from responsibility with the tips!

Anyway, that's all I want to talk about today. It's a pretty quick rant, but I'm sure you can relate, right? Or if you don't relate, let me know in the comments. I know that tipping is essential to our current system and the livelihood of many people, but I really believe that in an ideal world, businesses and service workers should be able to make a living without any tipping. Anyway, if you're not convinced, I'd like to hear why because honestly I can't think of a good reason to tip, at least not a long-term one. And if you like this video, don't forget to like and subscribe. I'll see you again soon with a new video. Bye!

E.3 Online gatherings are better than in-person

Hi, everyone! Welcome, or welcome back, to my channel. Today, we're going to talk about a topic that's been on my mind for some time: online gatherings versus the classic in-person meetups. I have had a lot more meet-ups online since the start of the pandemic, and though it was a hassle at the beginning, now that it's over, I wonder: should we actually go back to meeting in-person? After thinking on it for a bit, I came to the conclusion that online gatherings are superior to in-person and that we should stick to it. Let me tell you why.

Firstly, we can meet people from all around the world online. I live in <CITY NAME>, which is nice and all, but the big artists like Taylor Swift never hold a concert here. Authors that I like never do book launches or book signing events here. I used to see pictures of people attending all these events in-person and I was so jealous! Like, it's not that I can't afford the ticket, but if I would have to pay for so much bus fare and plane fare on top of the ticket, and maybe even book hotels and urgh! it's just not feasible. But, you know what happened when the pandemic hit? Every event is suddenly online. There are online concerts, where you can actually turn on the camera and show your face to the artist performing. And a little bonus point is that some of these online concerts have special AR effects that they overlay on the artist and the stage, which is just impossible to do in real life. Just look at these pictures. Aren't they just gorgeous? There are also online book launch events, where I get to listen to my favorite authors talking, even though it means I have to get up really early because that author is in England. Australian authors are a bit more convenient since they usually hold the events at early evening PST time. Anyway, I can't count on one hand how many events I've attended since everything is moved online. Events that I wasn't able to attend before when they were in-person. It was a dream come true! Anyway, now, there aren't that many online events anymore, since they're trying to get things back to in-person. But please. Please! Keep at least some of it online for a poor, broke, grad student like me.

But online meetings aren't only superior for events happening far away. It's also more convenient for events that are in my city. From my house to my university, it takes about one hour, which is about 20 minutes walking and 40 minutes on the bus. By the time I get to the university, my brain has been shaken into a smoothie and I have no energy left to work. Then, commuting back home takes another hour. And this city being Vancouver, it rains, like, all the time. Can you imagine going through all that to get home? I'm usually so tired I can't even bother to cook and just eat cereal for dinner, not to mention doing homework. Tragic, I know. Anyway, it all got better since the pandemic, because all the class and lab meetings have been moved online! I can wake up at 9:50, instead of at 8, for a meeting at 10. And maybe you math-y people will say, "hey, that's not one hour" and you're right, it's not, because Zoom has this beauty function that can smooth out your face pores so you also save the time for putting on make-ups. In total, that's almost 2 more precious hours for sleeping. It's so convenient and comfortable. And my productivity has definitely gone up a lot too. I simply can't see a reason for moving it back to in-person at all.

Anyway, that's all I want to talk about today. It's a pretty quick rant, but I'm sure you can relate, right? Or if you don't relate, let me know in the comments. I know some people are really excited about going back to in-person, and I'd like to hear why because honestly I can't think of a good reason for it. If you like this video, don't forget to like and subscribe. And I'll see you again soon with a new video. Bye!

E.4 In-person gatherings are better than online

Hi, everyone! Welcome, or welcome back, to my channel. Today, we're going to talk about a topic that's been on my mind for some time: online gatherings versus the classic in-person meetups. I have had a lot more meet-ups online since the start of the pandemic, and though it was a hassle at the beginning, now that it's over, I wonder: should we actually go back to meeting in-person? After thinking on it for a bit, I came to the conclusion that in-person gatherings are superior to online and that we should try our best to go back to it. Let me tell you why.

Firstly, online meetings just don't give you the same feeling of real connection between humans. I think the best example for this is a concert. Before the pandemic, I went to a concert of a k-pop boy group, SuperM, here in <CITY NAME>. The energy was off the chart! Just being among other people who are passionate about the same thing as you, hearing their screams and roars and singing along – it was so wild and powerful and just, electric. And I also scream my lungs off, even though you guys know I'm not the type to normally scream, but that's what being among other people, being in a real place with a real atmosphere, can do to me.

Another example I can think of for human connection is with my family. So, my parents and I live in different countries, and we try to keep in-touch with video calls. But it's just not the same, you know? Not that I'm not grateful we have video calls. It's helpful, but it cannot replace the real thing. The real interactions. I can have dinner at the same time that my parents have lunch while we're both on Zoom, but it's never going to be the same as actually eating at the same table, where I can pass them the salt and in return, they let me have a taste of their dishes. I think, if we get too used to Zoom calls, we may think, Oh, I don't really need to go home, we already meet all the time. But it's not the same. If you don't believe me, go and schedule an in-person meet-up with your loved ones and you'll remember how beautiful it is to have real human connections.

Building on this, I also want to point out that meeting people in-person allows for a much more complex conversation, because you have all these sensory and cues that can't be captured through online meetings. The easiest example is probably eye contact. You can't really meet someone's eyes through the camera. You either

have to look at the camera or at the image of the other person on the screen, but you can't look at both at the same time. So, there's no way you can make eye contact. Which means you can't communicate many subtle emotions or intentions. Body language is another subtle cue that can't be captured very well in an online setting. Usually you see only people's faces, at most their upper body. But maybe their legs are bouncing anxiously below the camera, and you'd never know. And of course, all this becomes harder when you're in a group meeting or gathering. Have you ever tried to speak up in an online group meeting, only for another person to speak up at the same time? That happens to me all the time, when it barely happens at all in in-person meetings. It's just so hard to tell what everybody is thinking and doing online. It makes everything awkward and eventually makes me not want to really talk anymore.

Anyway, that's all I want to talk about today. It's a pretty quick rant, but I'm sure you can relate, right? Or if you don't relate, let me know in the comments. I know some people want to keep things online, and I'd like to hear why because honestly I can't think of a good reason for it. If you like this video, don't forget to like and subscribe. And I'll see you again soon with a new video. Bye!

F FULL RESULTS FROM THE EVALUATIVE STUDY

F.1 Critical Thinking

The results of all statistical tests for critical thinking are shown in Table 2.

When interpreting the results with significant effect ($p < 0.05$) or a trend of significant effect ($p < 0.06$), we looked at the β_1 value, which indicates the fitted slopes between onditions. For rhetorical style comparison, a positive β_1 means persuasive has a higher mean than eristic rhetoric. For social identity comparison, a positive β_1 means outgroup has a higher mean than ingroup. For the interaction effect between social identity and rhetorical style, a positive β_1 means the persuasive and outgroup has the highest mean.

F.2 Stance on the topic

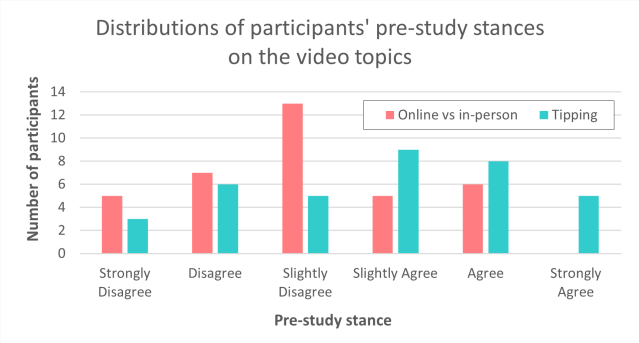


Figure 7: The distribution of 36 participants' pre-study stances on the two debate topics in the study: online gatherings are better than in-person and customers should tip.

The distribution of all pre-study stances (stances before watching the video) is shown in Figure 7. The stances are separated by topics in Figure 9 (online gathering topic) and Figure 10 (tipping topic).

All stance changes before and after watching the video are accumulated in Figure 8. All stance changes before and after talking to the chatbot, divided by the rhetorical styles and social identities of the chatbot, are accumulated in Figure 11 to 14. The tables of stances are color coded. The grey cells indicate that the stances before and after remain unchanged. The red cells indicate that the stance is strengthened (e.g., from slightly agree to agree, from agree to strongly agree), which is represented by negative value in the statistical tests. The green cells indicate that the stance is swayed is weakened or swayed in the opposite direction (e.g., from strongly agree to agree, from agree to disgree.)

F.3 Perception of the chatbot

The results of all statistical tests for perception of the chatbot are shown in Table 3. The results for each measure here are interpreted the same way as self-reported critical thinking, as it is the output of linear mixed-effect models.

	Rhetorical style (main effect)	**Social identity** (main effect)	**Rhetorical style x social identity** (interaction effect)
Total mean = 5.11 S.D. = 1.18	β_1 = 0.72 p = 0.06 .	β_1 = -0.43 p = 0.03 * η^2 = 0.07	β_1 = 0.45 p = 0.11 (n.s.)
Interpretation mean = 4.67 S.D. = 1.67	β_1 = 1.20 p = 0.02 * η^2 = 0.32	β_1 = -0.76 p = 0.04 * η^2 = 0.05	β_1 = 0.94 p = 0.07 . η^2 = 0.09
Analysis mean = 4.75 S.D. = 1.55	β_1 = 0.70 p = 0.15 (n.s.)	β_1 = -0.87 p = 0.01 * η^2 = 0.09	β_1 = 0.71 p = 0.14 (n.s.)
Evaluation mean = 5.35 S.D. = 1.14	β_1 = 0.65 p = 0.08 . η^2 = 0.09	β_1 = -0.19 p = 0.49 (n.s.)	β_1 = 0.17 p = 0.67 (n.s.)
Inference mean = 5.17 S.D. = 1.21	β_1 = 0.60 p = 0.14 (n.s.)	β_1 = -0.29 p = 0.21 (n.s.)	β_1 = 0.17 p = 0.61 (n.s.)
Explanation mean = 5.33 S.D. = 1.28	β_1 = 0.80 p = 0.06 . η^2 = 0.14	β_1 = -0.14 p = 0.62 (n.s.)	β_1 = -0.06 p = 0.88 (n.s.)
Self-regulation mean = 5.40 S.D. = 1.37	β_1 = 0.37 p = 0.42 (n.s.)	β_1 = -0.33 p = 0.09 . η^2 = 0.09	β_1 = 0.63 p = 0.03 * η^2 = 0.15

Table 2: Table showing the β_1, p-value, and for significant p-value, η^2 effect size for the linear mixed-effect models run on all measures of critical thinking.

Before & after the video (ALL CONDITIONS)

After / Before	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Strongly Disagree	8	0	0	0	0	0
Disagree	4	9	0	0	0	0
Slightly Disagree	0	6	12	0	0	0
Slightly Agree	0	1	0	7	6	0
Agree	0	0	1	1	7	5
Strongly Agree	0	0	0	0	0	5

Figure 8: Table of participant's stances on the topic (both topics) before and after watching the video

F.4 Engagement and motivation

The results of all statistical tests for engagement and motivation are shown in Table 4. The results for each measure here are interpreted the same way as self-reported critical thinking, as it is the output of linear mixed-effect models.

Before & after the video (online is better than in-person debate topic)

After / Before	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Strongly Disagree	5	0	0	0	0	0
Disagree	3	4	0	0	0	0
Slightly Disagree	0	6	7	0	0	0
Slightly Agree	0	0	0	1	4	0
Agree	0	0	1	0	2	3
Strongly Agree	0	0	0	0	0	0

Figure 9: Table of participant's stances on the "Online gatherings are better than in-person" topic, before and after watching the video

Before & after the video (customer should tip debate topic)

After / Before	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Strongly Disagree	3	0	0	0	0	0
Disagree	1	5	0	0	0	0
Slightly Disagree	0	0	5	0	0	0
Slightly Agree	0	1	0	6	2	0
Agree	0	0	0	1	5	2
Strongly Agree	0	0	0	0	0	5

Figure 10: Table of participant's stances on the "Customers should tip" topic, before and after watching the video

Before & after the chatbot (eristic)

After / Before	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Strongly Disagree	5	0	0	0	0	0
Disagree	0	7	0	0	0	0
Slightly Disagree	0	1	5	0	0	0
Slightly Agree	0	0	0	2	1	1
Agree	0	0	1	0	5	2
Strongly Agree	0	0	0	0	1	5

Figure 11: Table of participant's stances on the topic before and after chatting with chatbots with eristic rhetorical style

Before & after the chatbot (persuasive)

After / Before	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Strongly Disagree	5	1	1	0	0	0
Disagree	0	7	2	0	0	0
Slightly Disagree	0	1	6	0	0	0
Slightly Agree	0	0	0	4	0	0
Agree	0	1	0	0	4	0
Strongly Agree	0	0	0	0	0	4

Figure 12: Table of participant's stances on the topic before and after chatting with chatbots with persuasive rhetorical style

Before & after the chatbot (outgroup)

After / Before	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Strongly Disagree	5	1	1	0	0	0
Disagree	0	4	0	0	0	0
Slightly Disagree	0	1	7	0	0	0
Slightly Agree	0	0	0	4	0	0
Agree	0	0	0	0	7	2
Strongly Agree	0	0	0	0	0	4

Figure 13: Table of participant's stances on the topic before and after chatting with chatbots with outgroup identity

Before & after the chatbot (ingroup)

After / Before	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Strongly Disagree	5	0	0	0	0	0
Disagree	0	10	2	0	0	0
Slightly Disagree	0	1	4	0	0	0
Slightly Agree	0	0	0	2	1	1
Agree	0	1	1	0	2	0
Strongly Agree	0	0	0	0	1	5

Figure 14: Table of participant's stances on the topic before and after chatting with chatbots with ingroup identity

	Rhetorical style (main effect)	**Social identity** (main effect)	**Rhetorical style x social identity** (interaction effect)
Total mean = 4.90 S.D. = 1.38	β_1 = 1.07 p = 0.01 * η^2 = 0.21	β_1 = 0.25 p = 0.32 (n.s.)	β_1 = 0.20 p = 0.57 (n.s.)
Likeability mean = 4.47 S.D. = 1.74	β_1 = 2.11 p = 7.86e-05 *** η^2 = 0.35	β_1 = 0.22 p = 0.52	β_1 = -0.55 p = 0.27 (n.s.)
Anthropomorphism mean = 5.01 S.D. = 1.67	β_1 = 0.29 p = 0.61 (n.s.)	β_1 = 0.25 p = 0.43 (n.s.)	β_1 = 0.36 p = 0.43 (n.s.)
Perceived intelligence mean = 4.68 S.D. = 1.56	β_1 = 0.85 p = 0.09 . η^2 = 0.16	β_1 = -0.04 p = 0.89 (n.s.)	β_1 = 0.56 p = 0.21 (n.s.)
Perceived safety mean = 5.67 S.D. = 1.67	β_1 = 1.21 p = 0.02 * η^2 = 0.18	β_1 = 0.36 p = 0.28 (n.s.)	β_1 = 0.13 p = 0.77 (n.s.)
Helpfulness mean = 4.65 S.D. = 1.93	β_1 = 0.89 p = 0.16 (n.s.)	β_1 = 0.46 p = 0.22 (n.s.)	β_1 = 0.51 p = 0.35 (n.s.)

Table 3: Table showing the β_1, p-value, and for significant p-value, η^2 effect size for the linear mixed-effect models run on all measures of perception of the chatbot.

	Rhetorical style (main effect)	**Social identity** (main effect)	**Rhetorical style x social identity** (interaction effect)
Behavioral engagement mean = 6.14 S.D. = 0.73	β_1 = 0.16 p = 0.53 (n.s.)	β_1 = -0.01 p = 0.95 (n.s.)	β_1 = -0.24 p = 0.17 (n.s.)
Emotional engagement mean = 5.58 S.D. = 1.43	β_1 = 0.91 p = 0.06 . η^2 = 0.10	β_1 = 0.14 p = 0.51 (n.s.)	β_1 = -0.15 p = 0.62 (n.s.)
Cognitive engagement mean = 5.12 S.D. = 1.13	β_1 = 0.93 p = 0.01 . η^2 = 0.19	β_1 = 0.03 p = 0.92 (n.s.)	β_1 = -0.10 p = 0.78 (n.s.)
Intrinsic motivation mean = 5.60 S.D. = 1.38	β_1 = 0.77 p = 0.097 . η^2 = 0.09	β_1 = 0.23 p = 0.37 (n.s.)	β_1 = -0.04 p = 0.91 (n.s.)
Amotivation mean = 2.46 S.D. = 1.56	β_1 = -0.94 p = 0.06 . η^2 = 0.14	β_1 = -0.18 p = 0.37 (n.s.)	β_1 = -0.28 p = 0.33 (n.s.)

Table 4: Table showing the β_1, p-value, and for significant p-value, η^2 effect size for the linear mixed-effect models run on all measures of engagement and motivation.