Debate Chatbots to Facilitate Critical Thinking on YouTube

2024 Lab Seminar

인간중심컴퓨팅연구실 박사과정 류정우

Overview

- 필터 버블(Filter bubble)을 완화 시키는데 사용자를 **다양한 관점에 노출시키는 것**이 도움이 됨
- LLM 기반의 챗봇을 이용해 필터 버블 완화시키려면…?
  - 사용자가 영향을 받는지(critical thinking을 할 지) 여부는 **챗봇의 persona에 dependent**
  - 두 가지 persona attributes의 효과를 알아보자!
    - **Social Identity & Rhetorical Styles**

온라인 커뮤니티에서의 필터 버블(Filter Bubble)을 완화시킬 수 있는 챗봇 디자인

Paper

Debate Chatbots to Facilitate Critical Thinking on YouTube

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

Research Questions

- RQ1. How do **two key attributes** of chatbot persona **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?

Two attributes

Why social identity and rhetorical styles?

Social Identity

: ‘우리’냐 ‘그들’이냐가 그 사람의 ‘인지된 신뢰도’에 중요한 영향을 미친다는 것이 밝혀졌음

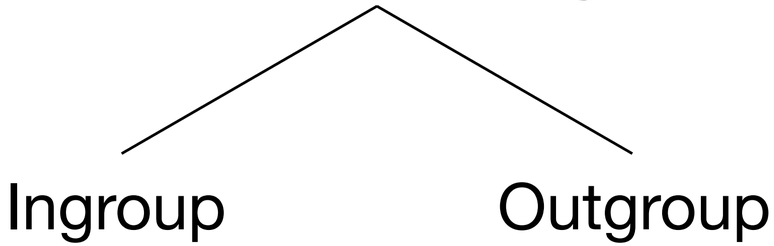
Rhetorical Styles

: 비판적 사고에 가장 도움이 되는 방식으로 논쟁적 담화의 목표를 설정하기 위해서

Two attributes

Why social identity and rhetorical styles?

Social Identity

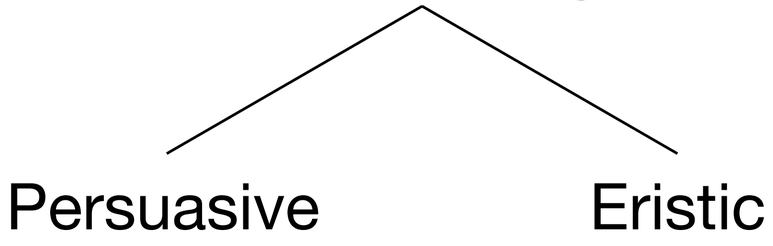


Ingroup 개인이 대체로 더 높은 신용과 신뢰

VS

설득자의 ‘권력’이나 ‘다수 지위’와 같은
여러 특정 요인에 따라 달라짐

Rhetorical Styles



Six common types:

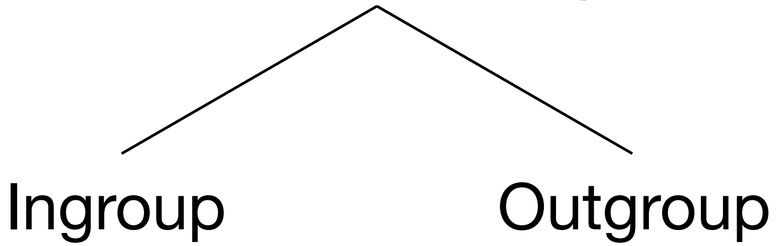
persuasion, inquiry, discovery,
information-seeking, deliberation, and **eristic**

Method

Controlled Variables

Within-subjects

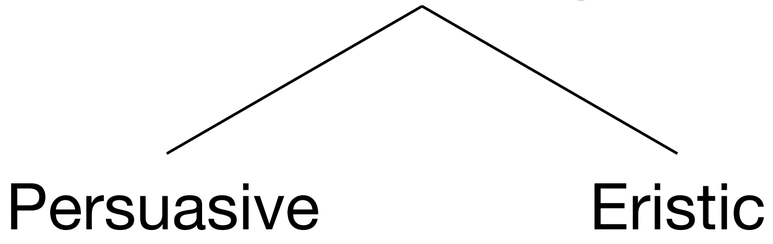
Social Identity



By aligning the chatbot's
ethnicity and ***gender*** to the participants

Between-subjects

Rhetorical Styles



By **sharing** the advantage of their stance
while remaining **understanding** of the viewer's viewpoint

By **attacking** the user relentlessly and **exposing** many differences in their opinion

Method

Controlled Variables

Within-subjects

Social Identity

Ingroup Outgroup

By aligning the chatbot's ***ethnicity*** and ***gender*** to the participants

Between-subjects

Rhetorical Styles

Persuasive Eristic

By **sharing** the advantage of their stance while remaining **understanding** of the viewer's viewpoint

F

By **attacking** the user relentlessly and **exposing** many differences in their opinion

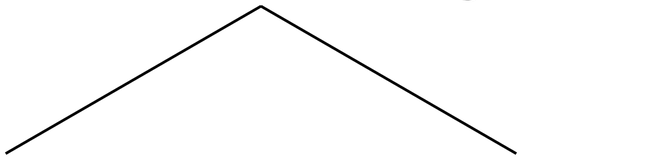
T

Method

Controlled Variables

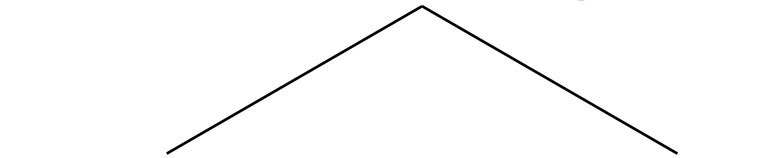
Within-subjects

Social Identity



Between-subjects

Rhetorical Styles



Between-subjects

Social Identity X 2 Videos(Topics)

: Ingroup으로 하나의 토픽을 보고,
Outgroup으로 나머지 하나 토픽을 보게끔

순서는 Counterbalanced

Method

Controlled Variables

Within-subjects

Social Identity

Ingroup　　Outgroup

Between-subjects

Rhetorical Styles

Persuasive　　Eristic

Between-subjects

Social Identity X 2 Videos(Topics)

: Ingroup으로 하나의 토픽을 보고,
Outgroup으로 나머지 하나 토픽을 보게끔

순서는 Counterbalanced

- Topic 1: 온라인 모임 VS 대면 모임
- Topic 2: 팁은 줘야한다

4 Videos - Favor/Opposition on each topic

Method

Dependent Variables

Stance on the topic

6 choices, Strongly Disagree ~ Strongly Agree

Critical Thinking Level

Critical Thinking Self-assessment Scale, CTSAS

Engagement and Motivation

Situational Motivation Scale, SIMS

Perception of the Chatbot

likability, anthropomorphism, perceived intelligence, perceived safety, and helpfulness

Method

Procedure

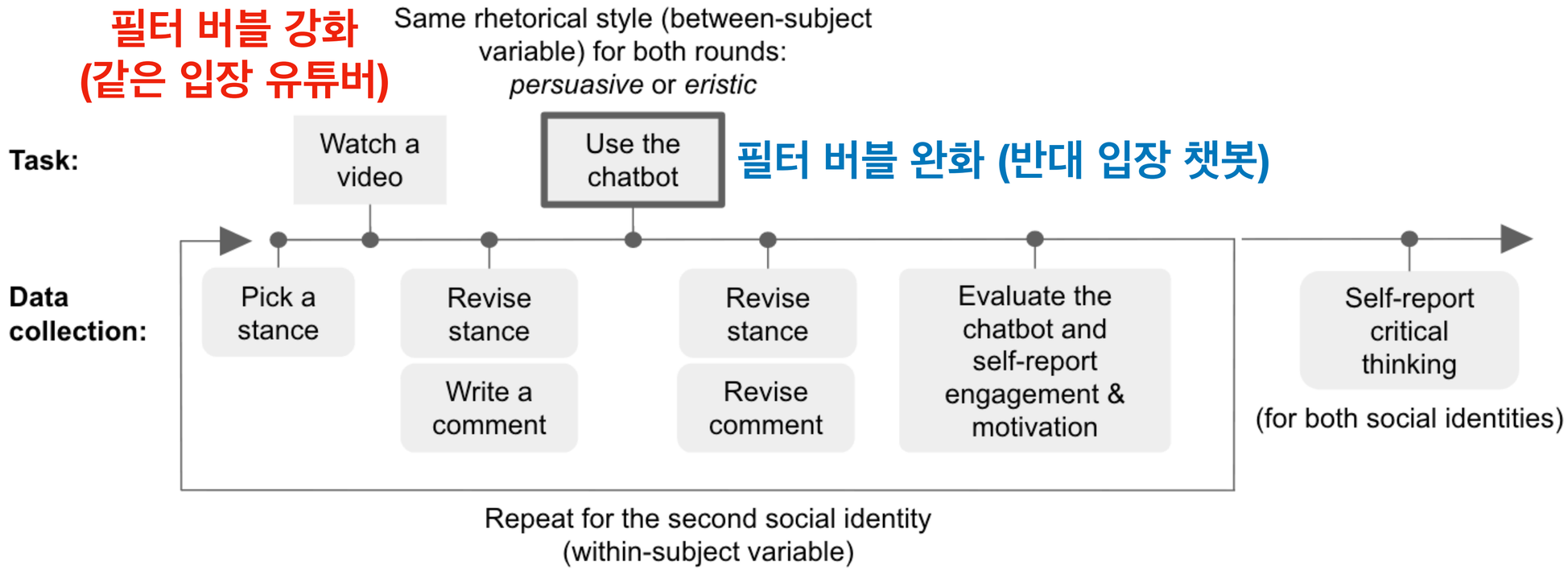
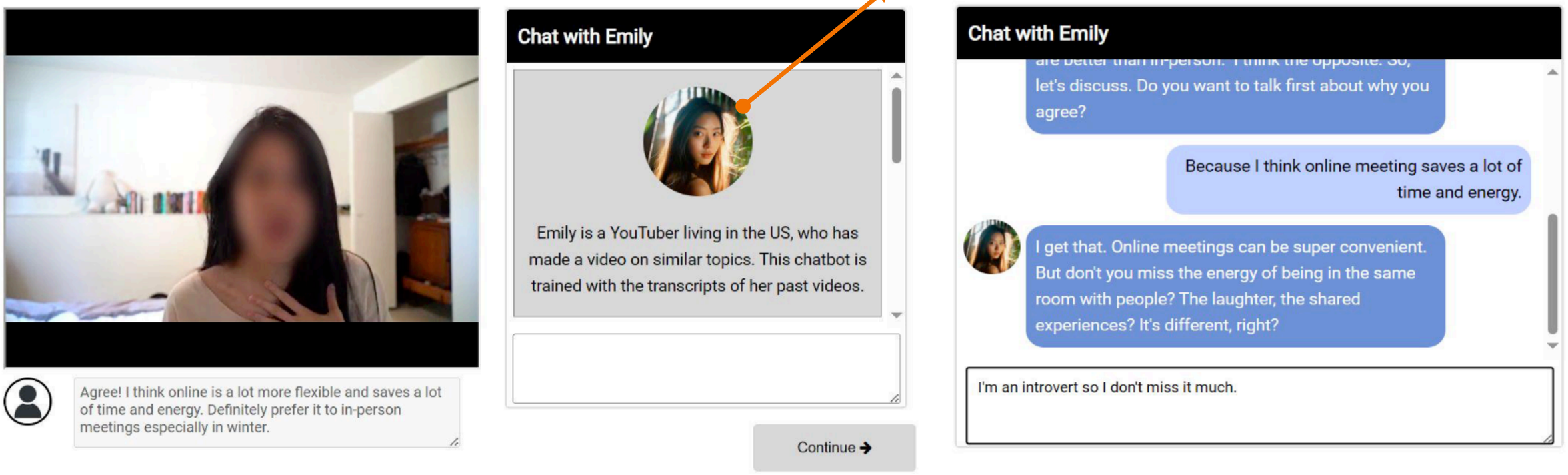


Figure 1: Diagram of the experiment procedure. The sharp-angled boxes show the tasks, and the round-angled boxes show the data collection steps. After the participants finish evaluating the chatbot and self-report engagement and motivation for the first social identity, they repeat the process for the 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).

Method

Chatbot Prototype



(a) A screenshot of the prototype

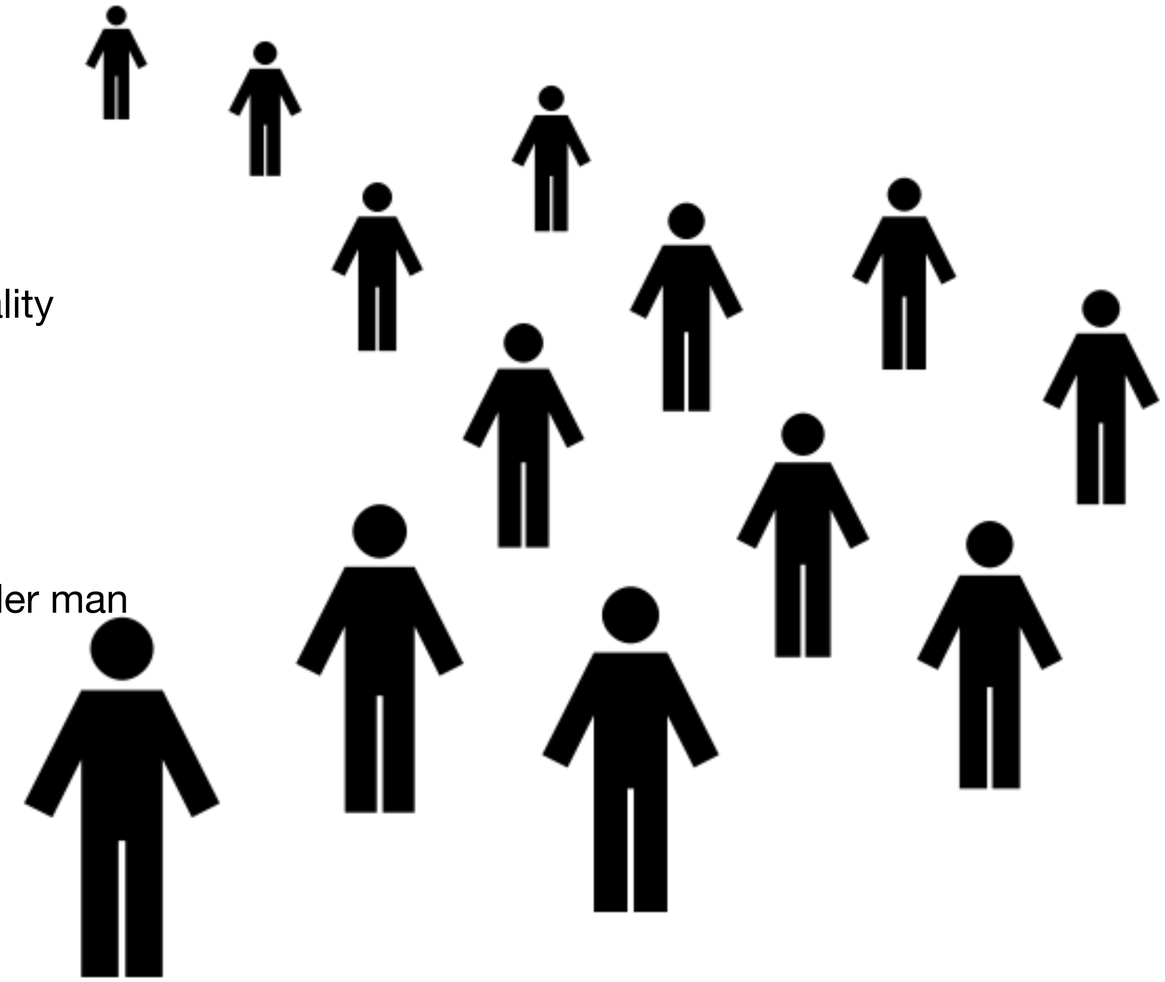
(b) An example chat conversation

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).

Method

Participants

- 36 participants recruited via Prolific
- English-fluent and from the USA to match the chatbot's nationality
- Average age: 35.3 years (S.D.= 11.86)
- Ethnicity distribution: 18 European/White, 6 biracial, 5 Asian, 3 Black, 2 Hispanic, 2 Indigenous, 2 unspecified
- Gender distribution: 15 women, 20 men, 1 nonbinary transgender man
- Chatbot familiarity: 29 participants had used ChatGPT; 29 comfortable using chatbots in general



Method

Data Analysis

Measurement Analysis(Stance 제외)

- Likert 척도(모든 측정문항이 7점 척도)로 응답 수집
- Linear Mixed Effects Models (LMM)

Stance Analysis

- “Strongly Disagree” 1점에서 “Strongly Agree” 6점까지 Mapping (비디오 시청 전/후, 챗봇과 대화 전/후)
- 입장 변화 측정: **점수차의 절대값에 방향이 바뀌거나 약해지면 +, 아니면 -**
- Wilcoxon Signed Rank Test를 통해 **입장 변화의 통계적 유의성 검증**

Qualitative Analysis(Thematic Analysis)

- 참가자의 사고 과정 및 챗봇에 대한 피드백 품질 분석: **Reflexive Thematic Analysis**
- 최종 결과: 42개의 Codes와 관련된 10개의 Subthemes(비판적 사고 4개 + 입장변화 3개 + 챗봇 인식 및 Engagement 3개)

Findings

Quant Summary

S: Social Identity / R: Rhetorical Style / I: Interaction Effect

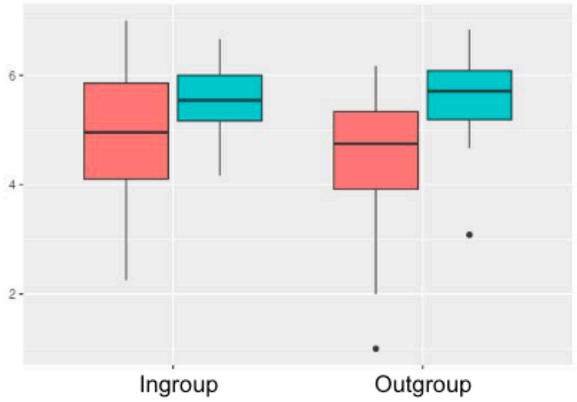
Critical thinking (Section 4.1)	**Stance on a topic (Section 4.2)**	**Perception of the chatbot (Section 4.3)**	**Engagement and motivation (Section 4.3)**
Total (S*, R.)	Before & After video (**)	Total (R*)	Behavioral engagement (n.s.)
Interpretation (S*, R*, I.)	- Difference between video topics (**)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

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.**

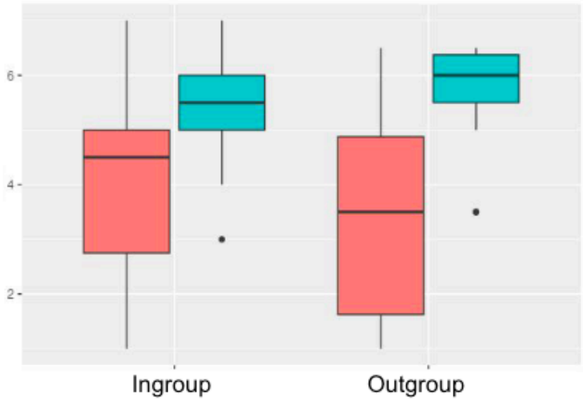
Findings

Critical Thinking

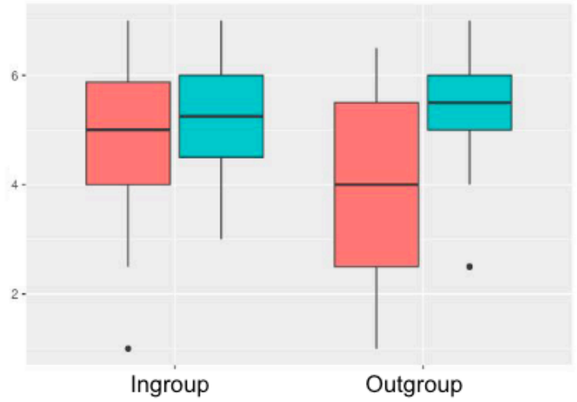
Total critical thinking (S*, R.)



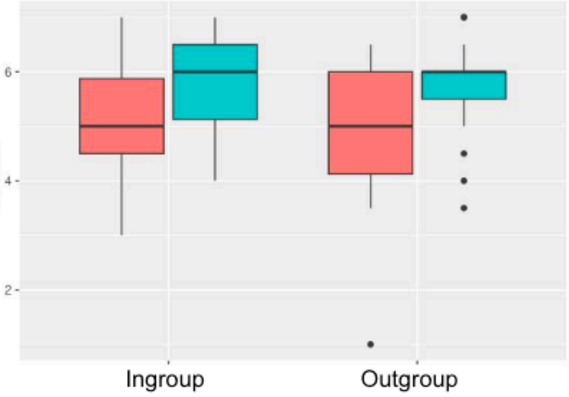
Interpretation (S*, R*, I.)



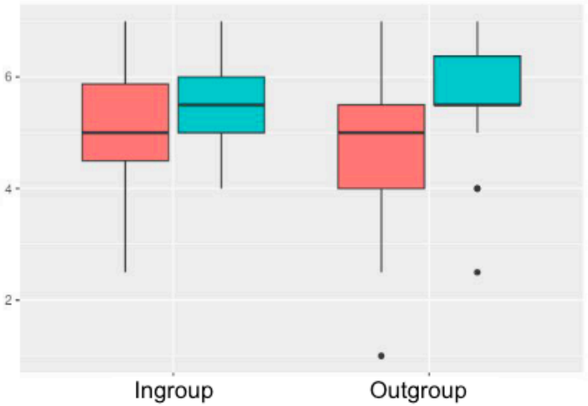
Analysis (S*, I.)



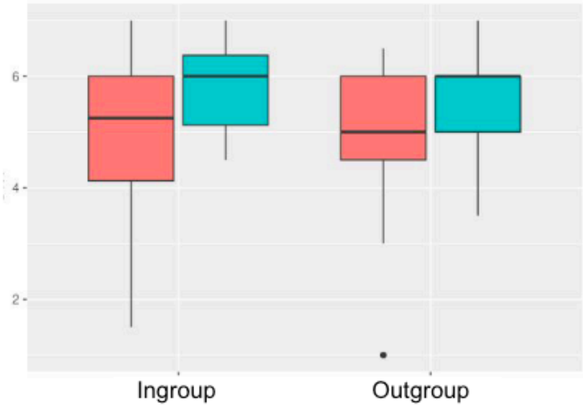
Evaluation (R.)



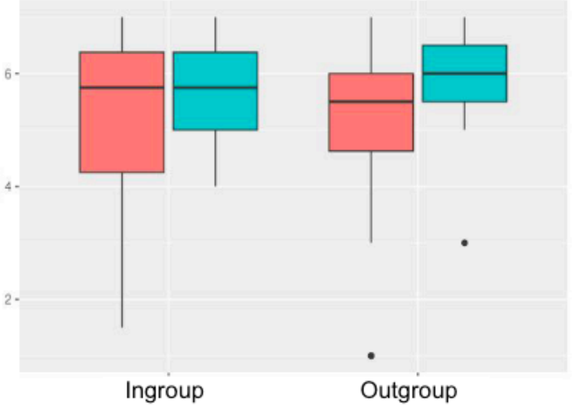
Inference



Explanation (R*)



Self-regulation (S., I*)



Effect labels:
S = main effect of social identity
R = main effect of rhetorical style
I = interaction effect

Significance codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Rhetorical style:
Eristic
Persuasive

Findings

Critical Thinking

챗봇의 Persona와 관계없이, 자신의 관점과 논리를 재고하도록 유도했음

- 챗봇의 주장에 동의하지 않더라도 왜 그렇게 생각하는지 더 깊이 있고 정확하게 성찰함

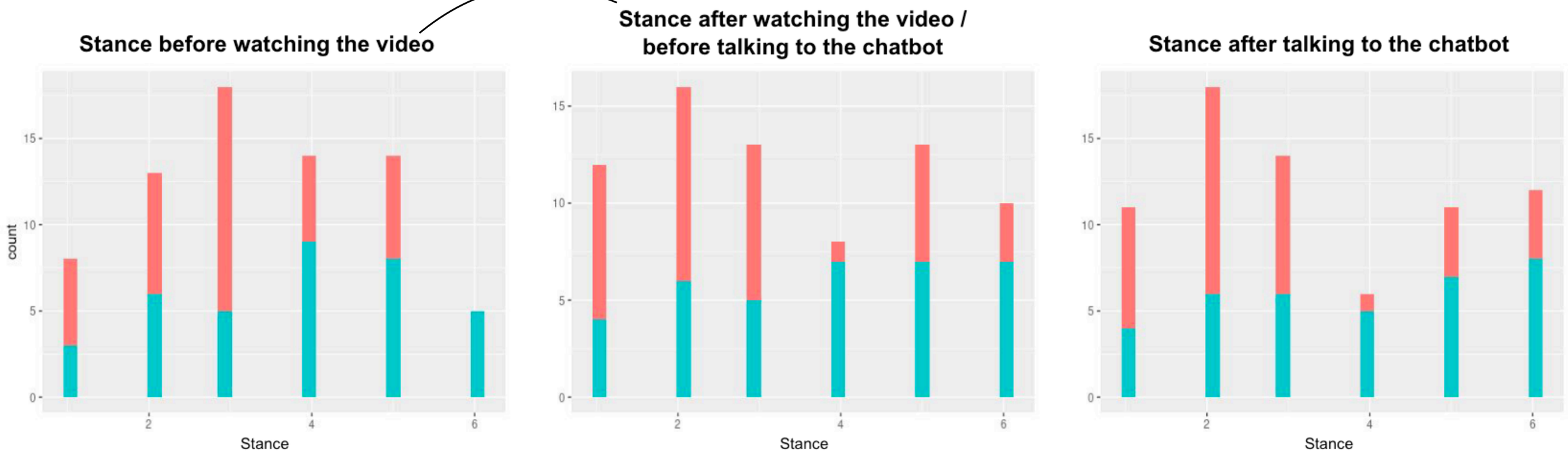
Rhetorical Style 측면에서 비판적 사고가 다른 시점에 일어남

- Persuasive 챗봇의 **주장 강도를 평가**할 때, Eristic 챗봇에서는 **자신의 논리를 강화할 때** 주로 일어난다고 함
- Persuasive 챗봇에 대해 참가자들은 “챗봇이 몇 가지 좋은 점을 제기했다”거나 “좋은 예시를 들었고 생각을 명확하고 간결하게 제시했다”고 **평가**

Findings

Stance on the Topic

**



Stance interpretation:
1 = Strongly Disagree
2 = Disagree
3 = Slightly Disagree
4 = Slightly Agree
5 = Agree
6 = Strongly Agree

Debate topic:
Online vs. in-person
Should tip vs. shouldn't

Findings

Stance on the Topic

참가자들은 챗봇이 제시한 새로운 관점이나 증거로 인해 자신의 입장을 변경하기도 함

- 챗봇이 제시한 **넓은 관점** 때문에 ‘온라인 대 대면’ 회의 주제에 대한 자신의 입장을 강하게 반대에서 약간 반대로 변경
- 챗봇이 시스템적 불의에 초점을 맞추어 **논의 범위를 확장**시킴으로써 ‘팁에 의존하는 것’이 공정한 임금에 부정적인 영향을 미치고 나쁜 관행을 장려한다는 것을 깨달아 동의에서 약간 반대로 입장 변경

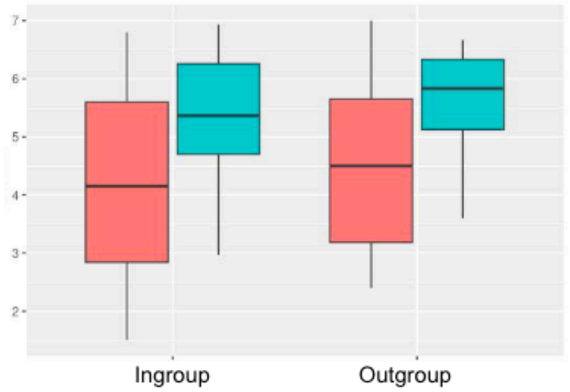
챗봇의 약한 논리는 참가자들의 원래 입장을 확실히 하거나 참가자의 입장으로 챗봇을 설득시키기도 함

- 설득력이 약한 챗봇의 주장으로 인해 챗봇의 입장을 타협키거나, 누구는 챗봇을 아예 자신의 입장으로 설득
- 챗봇 주장의 네 가지 약점: 반복적인 주장, 맥락 인식 부족, 실제 경험 부족, 새로운 통찰력 부족

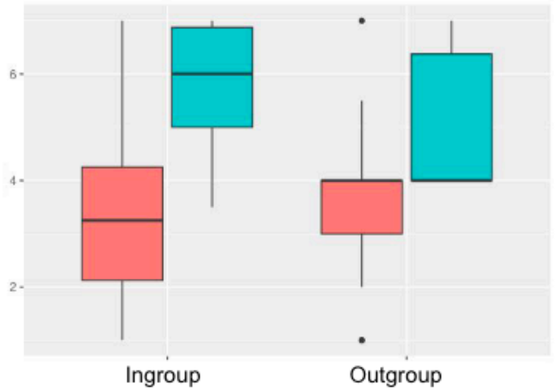
Findings

Perception of the Chatbot

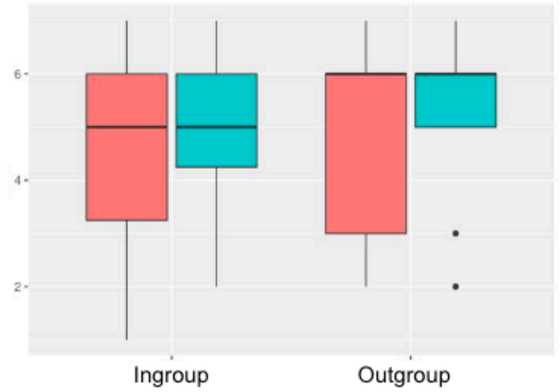
Total perception of chatbot (R)**



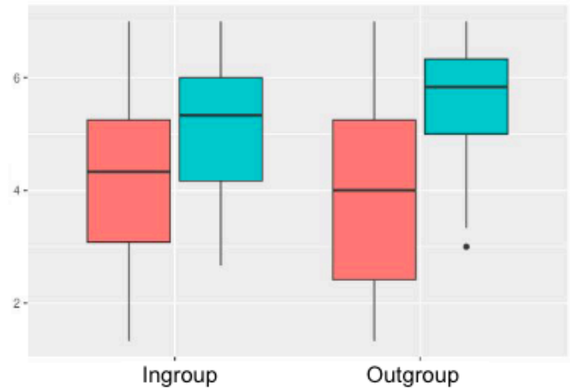
Likeability (R*)**



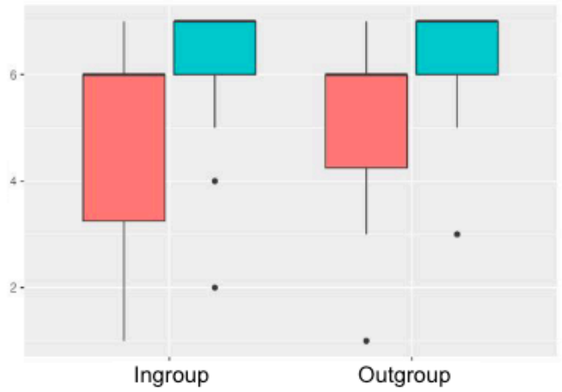
Anthropomorphism



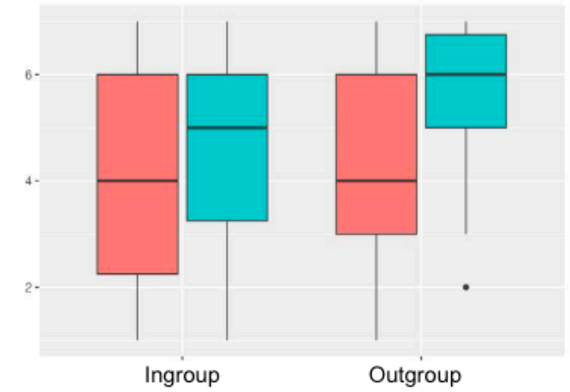
Perceived Intelligence (R.)



Perceived Safety (R*)



Helpfulness



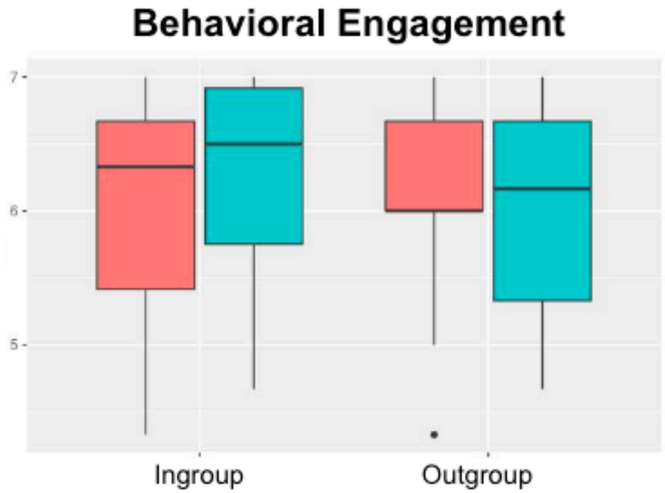
Effect labels:
S = main effect of social identity
R = main effect of rhetorical style
I = interaction effect

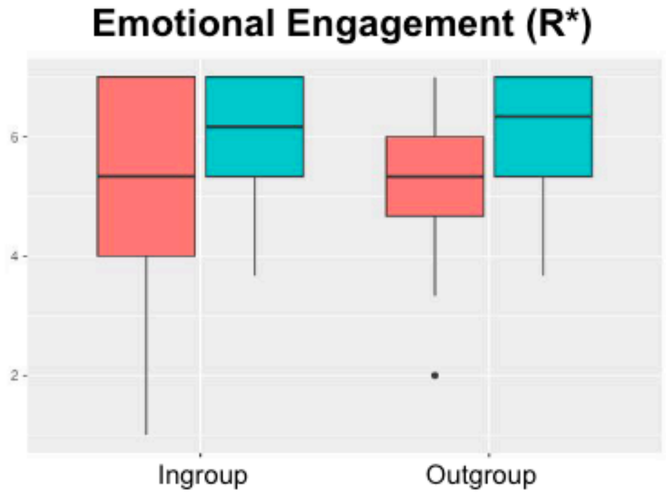
Significance codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

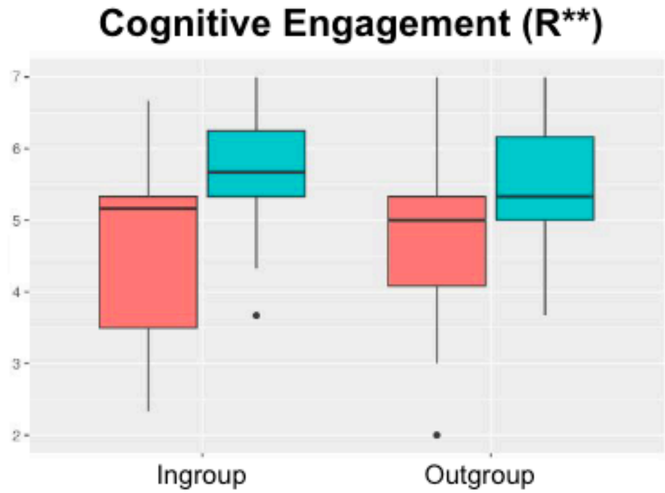
Rhetorical style:
Eristic
Persuasive

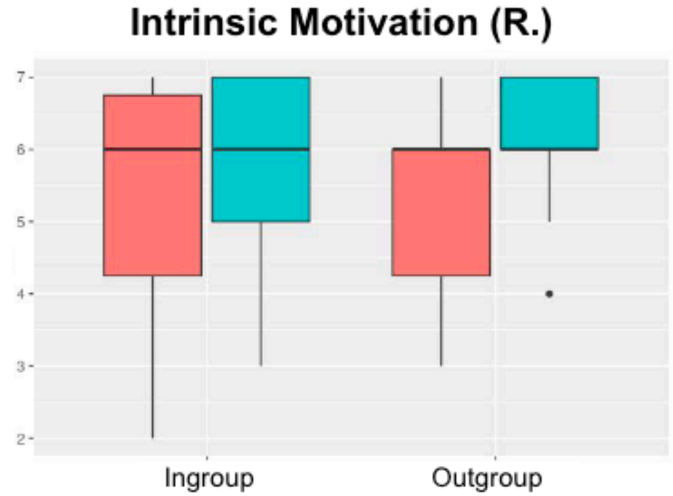
Findings

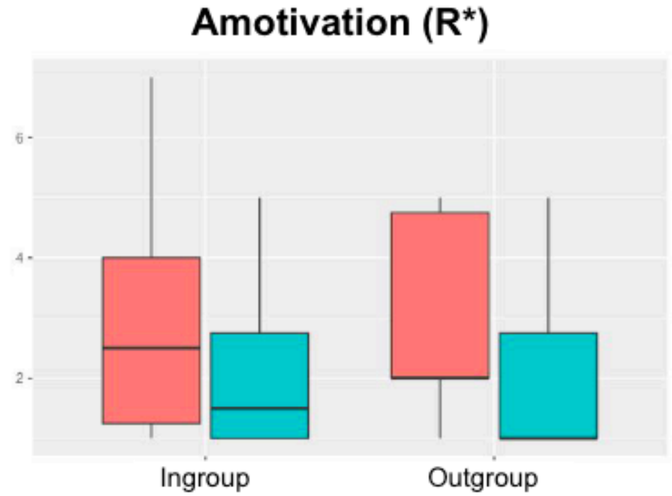
Engagement with the Activity











Effect labels:
S = main effect of social identity
R = main effect of rhetorical style
I = interaction effect

Significance codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Rhetorical style:

Findings

Perception of the Chatbot & Engagement

- 챗봇이 자신의 관점을 방어하려는 내재적 동기 없이, 그저 논쟁을 위한 논쟁을 벌이는 것처럼 보인다고 지적

- Eristic 챗봇의 무조건적이면서 과도한 자세는 논쟁에서 기대하는 상호작용 경험을 방해하고 챗봇의 likeability와 참여자의 감정적 engagement에 영향을 줌

- Eristic 챗봇이 개인적인 선호가 없는 태도는 의견이 중요한 논의에서 한계를 드러냈으며, 이로 인해 입장 결정에 있어서 helpfulness와 인지적 engagement에 영향을 줌

Discussion

Result Interpretation

- **LLM 기반 챗봇은 비판적 사고를 유도하는데 효과적이지만 직접적인 입장 변화에는 유의미한 영향이 없음**
- **Rhetorical Style은 토론의 방향과 참가자가 참여한 Critical Thinking 유형에 큰 영향을 미침**
  - Persuasive 챗봇은 **참가자가 챗봇의 논리를** 더 면밀히 평가하도록 함
  - Eristic 챗봇은 **참가자가 자신의 논리를** 강화하고 갈등의 근원을 찾으려고 함
- **Social Identity(인종 & 성별)는 Critical Thinking의 일부 영역에서만 유의미한 영향을 미침**
  - Self-regulation에 가장 큰 영향을 미침, 참가자가 **본인에게 내재된 편견과 가정을 재평가**하도록 유도
  - **Outgroup Identity X Persuasive Style의 상호작용**이 Self-regulation을 증진시키는데 가장 효과적임.

Appendix

Questionnaires

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.)

Thank you