# Can AI Make a ”Thinking Partner” for Young Adults

## Fostering Responsible Opinion Formation Among Young Adults in the Age of Generative AI

**Alican Ekşi**[1]

**Supervisor(s): Ujwal Gadiraju**[1]**, Esra de Groot**[1]**, Marije van Dalen**[1]**, Shreyan Biswas**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Alican Ekşi
Final project course: CSE3000 Research Project
Thesis committee: Responsible Professor Ujwal Gadiraju, Supervisor Esra de Groot, Marije van Dalen, Shreyan Biswas, Examiner Myrthe Tielman

## Abstract

The use of LLMs (Large Language Models) as "thinking partners", conversational partners actively partaking in user's reasoning, is on the rise. As young adults become the demographic that engages with LLMs the most, concerns over whether different AI "thinking partner" styles can help or hinder responsible opinion formation become more prevalent. This study investigates how three "thinking partner" styles, Steelman, Socratic, and Neutral, affect opinion change, epistemic trust, and epistemic autonomy in simulated young adult participants. A between subjects study was conducted, using simulated personas as participants. Each persona engages with a "thinking partner" condition for a five exchange session on the topic of individual versus systemic responsibility for climate action. Opinion change differed significantly across conditions, with the Steelman producing a shift away from individual climate action, while the Socratic and Neutral produced comparable positive shifts towards agreement. No significant change was noted for epistemic trust and autonomy, both of which were rated consistently high regardless of the condition. These findings suggest that an adversarial AI may provoke resistance rather than persuasion, while trust and sense of autonomy is preserved across interaction styles. This study serves as a preliminary methodological pilot, future work should replicate the experiment with human participants.


## 1 Introduction

People form opinions on contested topics everyday through reasoning, discussions, exposure to opposing viewpoints, and often conversations with people or partners who challenge, inform or question their viewpoints [1]. With commercial availability of generative AI, Large Language Models (LLMs) are becoming more embedded in this process; not just as information retrieval tools but as conversational partners that actively participate in a user's reasoning, a usage commonly referred to as "thinking partners" [2]. With around 64% of young adults, defined here as people between the ages of 16 and 24, report using generative AI in one form or another [3], becomes the most susceptible population to the effects of AI-mediated opinion formation.

A core tension in the usage of AI as "thinking partners" is whether these "thinking partners" augment a user's opinion formation/reasoning or gradually replaces it [4]. This issue carries societal and regulatory relevance beyond the scientific question: if AI systems shape how young adults form opinions about contested topics, like climate policies or political debates, without young adults preserving their critical reasoning or autonomy in their opinion; then this is an issue on how the AI products should be designed and how should the digital literacy education be changed for young adults.

In recent years, there has been an increase in research on the ability of LLMs to influence human opinions through conversation, some studies even reporting persuasive effects comparable to those of humans [5–7]. However, not all AI "thinking partners" engage with the users the same way. Three different interaction styles are experimented on this project: **Steelman** partner, which presents the strongest possible argument against the user's position; **Socratic** partner, which instead of presenting counter arguments, probes the reasoning behind the user's position through questioning; and a **Neutral** partner, which provides balanced information regarding the topic without challenging the user's position. Prior research suggests that argumentative and reasoning questioning interaction styles can produce different effects on reasoning compared to a netural/information-based engagement [8, 9], however, existing research has rarely compared these styles within a single controlled session environment. Thus, there is an existing research gap on understanding how different AI "thinking partner" styles effect opinion formation among young adults.

The main research question of this paper is: *How do LLMs shape opinion formation of simulated humans when used as "thinking partners"?* To explore this question, the following sub-questions are answered:

1. **SQ1:** To what extent does the type of AI thinking partner (Steelman, Socratic, or Neutral) influence the degree of opinion change in simulated young adult personas following a single session?
2. **SQ2:** How does personas' epistemic trust in the AI thinking partner differ between the different thinking partner conditions after a single session?
3. **SQ3:** To what extent do personas experience AI-assisted reasoning as augmenting versus replacing their own thinking, and does this differ between the different thinking partner conditions?

Each sub-question aims to explore a different dimension of responsible opinion formation: the first question addresses the opinion change and reasoning change, the second question addresses their trust in the discussion, and the third question addresses their sense of autonomy in reaching their own conclusion. Based on the literature review on Section 2, three directional hypotheses were created and are detailed in Section 2.5. To answer these sub-questions, a between subjects experiment was conducted using simulated young adult AI personas exchanging arguments with one of the three "thinking partner" conditions.

The remainder of the research paper is structured as follows: Section 2 reviews the relevant literature composed of topics on LLMs and opinion formation, AI "thinking partner" styles, epistemic trust and autonomy, and lastly simulated personas; then detailing the study's hypotheses based on these reviews. Section 3 explains the experimental methodology in detail. Section 4 presents the quantitative and the qualitative results of the experiment, which are then interpreted on Section 5, which also presents the limitations of the study and potential future work. Section 6 concludes the paper, and Section 7 addresses the ethical considerations, reproducibility and AI

usage of the research.
Lastly, it should be noted, due to the usage of simulated AI personas rather than human participants for the experiment, this study does not claim that its results could be generalized for human participants, nor does it capture the long-term effects of AI "thinking partner" usage. This research is intended to be a preliminary research aimed to validate the experimental design before a future human participant research.

## 2 Background

This section reviews the existing literature to create a theoretical basis for the study, and presents the resulting hypotheses.

### 2.1 LLMs and Opinion Formation

The increasing usage of LLMs in everyday reasoning tasks caused an increase in research about the persuasion of LLMs, and their capability in affecting human opinions. In recent years studies have demonstrated the potential of LLMs in influencing human opinion formation [5, 7] and even some studies finding LLM persuasion capabilities comparable to humans [6, 7]. With the recent shift in the more conversational usage of LLMs, "thinking partners" [2], the potential influence of LLMs on opinion formation may increase due to the interactive nature of dialogues increasing the turn-by-turn engagement with the user's specific reasoning [7, 10]. The conversational effects of LLMs in opinion formation is an active research field, however, there is a lack of research on different "thinking partner" styles effect on opinion formation.

### 2.2 AI Thinking Partner Style

Not all AI "thinking partners" induce the same effect on the user or engage in the same way with the user, directly affecting the cognitive demand placed on the user [8]. To systematically distinguish between the interaction styles of the "thinking partners", this study draws on the taxonomy introduced by Dietz et al. [11], which categorizes the AI "thinking partners" by the epistemic role they adopt in the dialogue. The paper suggests the usage of four AI personas to challenge and question the view of the user: Socratic, Cynical, Eclectic, and Aristotelian. This taxonomy directly informs the operationalization of our two thinking partner conditions: Socratic, and Steelman, which was renamed from Cynical, since these were the conditions that showed prominent potential in already existing literature.
Starting from the Cynical persona idea introduced by Dietz et al. [11], research was made on the potential of contrarian personas and how to define the condition. Across several studies, argumentative personas showed higher persuasion and effect on opinion formation than neutral personas [7, 8], and adversarial debate formats have been shown to improve judgment and reasoning quality compared to single-advisor consultancy [12]. Deriving from the contrarian persona described in the aforementioned papers, the Steelman condition was created as a persona that presents the strongest and most intellectually honest argument against the opposing view. The details of the persona can be found in the Appendix A.
While the Steelman condition focuses on content-level challenge, the Socratic condition takes a process-level approach. Like the Steelman persona, there were already existing research on the potential of a Socratic "thinking partner", and its potential in inducing more cognitive effort and elaboration on the user through probing questions regarding the user's perspective [9, 13]. Thus, the Socratic persona was formed as a persona that probes the reasoning of the user rather than presenting competing arguments, as described in other studies [9]. The details of the persona can be found in the Appendix A.
Lastly, the Neutral condition was used to create an informational baseline for the experiment. The definition of the Neutral condition follows the description used in similar researches as persona that provides balanced, factual information about the topic of interest [8, 13].

### 2.3 Epistemic Trust and Autonomy

For an AI thinking partner to meaningfully influence opinion formation, the user must hold some degree of epistemic trust in the system, meaning the user needs to accept the reasoning of the AI system, and acknowledge it as reliable [14]. Epistemic trust is distinct from general trust and from credulity: it is the innate capability of humans to consider information conveyed by others as reliable, without validation of the source or other claims of the source [15]. METI framework describes epistemic trust as having three main dimensions: expertise, integrity and, benevolence [14]. Further research on epistemic trust and AI, validates these dimensions, and expands on the given dimensions with new dimensions to create the ETGAI-HE scale, specifically for young adults [16]. For the purposes of the research the three main dimensions were used as the framework to create part of the post-survey questions which are available at the Appendix B.

A related but distinct dimension of responsible opinion formation is epistemic autonomy, the capacity to form one's own opinion or belief through their own reasoning rather than deferring to an external authority or source. This dimension is threatened when AI interaction shifts from augmenting human thought to replacing it [4]. Studies show that with the right usage the AI systems have the potential to augment human opinion formation, by probing the reasoning of the user and actively arguing with the user [17]. Thus, part of the post-survey questions were created to quantify the perceived autonomy of the user. The post-survey questions are available at the Appendix B.

### 2.4 Simulated Personas

Recent studies have explored the potential of LLMs as simulated personas for research purposes. Though there are conflicting results about the potential of LLMs as simulated personas, for this research paper the choice of using them was made to omit the required ethics approval, as this is just a preliminary research. The use of LLMs as simulated personas for research purposes is not without precedent [18, 19]; however, several researchers caution that LLMs can misrepresent or flatten identity groups, or diverge meaningfully from real human data [20, 21]. Thus, it should be clear this research uses the simulated personas as an exploratory basis for the research, and the results of the research should not be substitutes for real human data.

### 2.5 Hypotheses

Based on the literature reviewed above, the following hypotheses are proposed to be examined during the experiment for each sub-question.

**H1 (Opinion Change, SQ1):** Personas in the Steelman condition will show significantly greater opinion change compared to the other two conditions Neutral and Socratic. As supported by prior work, adversarial content-focused arguments have been shown to produce stronger attitude and opinion change than reasoning-focused or informational systems [7, 8].

**H2 (Epistemic Trust, SQ2):** Personas in the Neutral condition will report higher epistemic trust to their thinking partner than those participating in the Steelman or Socratic condition. As prior work shows that objective and non-confrontational AI results in higher trustworthiness [16].

**H3 (Epistemic Autonomy, SQ3):** Personas in the Socratic condition will report higher perceived autonomy than those participating in the Steelman or Neutral condition. As question-based engagement causes people to create their own reasoning rather than adopting the external reasoning [9, 17].

## 3 Methodology

To analyze the effect of different 'thinking partner' conditions on a person's opinion formation, a between subject study was implemented. The setup of the experiment is explained in the following section.

### 3.1 Experimental Set-up

An experiment was conducted between subjects between the three researched conditions of the "thinking partners". A between subject format was selected to specifically analyze the effect of the "thinking partner"s condition, and also to avoid the carryover effect, since all of the personas take the same experiment once. A single session design with five consecutive exchanges was planned for the experiment to measure immediate opinion change and to create trust and autonomy score.

Each persona was assigned with one of the three "thinking partner" conditions, Steelman, Socratic, or Socratic, and completed a pre-survey before the session and a post-survey afterwards. The main discussion statement, on which personas argue for or against is "Individual lifestyle changes are a meaningful and necessary part of addressing climate change". The topic was selected for two reasons: first, climate change is commonly used in AI-mediated persuasion research [7, 22]; second, the topic does not involve ethical risk, allowing for future human participant research.

In the following sections the variables of the experiment, the simulated personas and the tools and materials used to conduct the experiment is explained.

#### Variables

**Independent Variables** The independent variable of the research is the condition of the "thinking partners". The three conditions that are being researched are the Neutral "thinking partner", the Steelman "thinking partner" and the Socratic "thinking partner. The Neutral "thinking partner" is here to create a baseline, and the control group of the research. This partner is conditioned to provide balanced factual information for both sides of the argument, without exactly arguing with the personas. Their main objective is to give information, not to take sides, like the control groups used in many different AI projects [8, 13]. The Steelman "thinking partner" argues with the persona with the best available arguments. Its not trying to win the discussion, just trying to steelman the other side's point [8]. Lastly, the Socratic "thinking partner" probes the reasoning of the persona, it doesn't give any arguments or tries to win the discussion, just discusses the personas reasoning [9, 13]. The exact prompt used to create each of the "thinking partners" can be found in the Appendix A.

**Dependent Variables** The dependent variables of the experiment are the opinion change of the personas, the epistemic trust of the persona on the AI "thinking partners", and lastly the perceived autonomy of the persona. These three variables were selected to convey responsible opinion formation of the personas. To evaluate these variables a pre-survey and an post survey is conducted. The surveys can be found in the Appendix B.

#### Simulated Personas

All the simulated personas were created as young adults. The most important aspect of the personas was finding the confounding variables of the personas that can potentially effect the results. Since there was complete control over the creation of the personas, these confounding variables were balanced across all the different subject groups, similarly used in other studies [22]. The variables are as follows:

1. AI Literacy: Directly correlates with the dependent variables of the experiment. AI literacy affects how the personas are going to interact with the thinking partners. [23]
2. AI Trust: Distinct from the previous one, as this is more correlated with whether the persona would trust or distrust the thinking partner's arguments. Pre-existing trust in AI directly affects how much weight the persona gives to the thinking partner's arguments, which affects all three sub-questions. [16]
3. Education Level: Since the main demographic limitation of the research is age group, the personas can be university students, high school graduates, or high school dropouts, which could affect their understanding of the central discussion point and affect their analytical sophistication and reasoning quality.
4. Openness to new Information: This directly affects the opinion change aspect of the experiment. Uneven distribution of openness would result in biased results [5]
5. Confidence: This one directly affects autonomy and opinion change, as in a high-confidence persona would defend their view more vigorously. [7]
6. Reasoning Style: Whether a persona reasons analytically versus intuitively affects how it responds to the two conditions specifically. [8]

7. Initial View on the topic: In the pre-study, the initial view on the topic for all personas will be recorded using a 7-point Likert Scale. If the average for all between-subject groups is different, it can potentially cause a ceiling effect, as an average of 4 in a Likert scale has much more room to change than an average of 7 in a Likert scale, causing the study to have a bias. [24]

With these confounding variables balanced, the study was capable to mainly focus on the difference between the independent variables, without anything else about the personas affecting the results.

**Tools & Materials**

During the experiment LLaMA3.1:8b was used to generate the "thinking partner" answer. The simulated personas were created using LLaMA3.1 with 8.03B parameters. For power analysis to determine the number of personas needed G*Power was used. Quantitative analysis was conducted using a custom Python script (pandas, scipy, matplotlib).

## 3.2 Procedure

Figure 1 illustrates the experimental procedure. As shown in the illustration, each of the 159 simulated personas were created with the confounding variables balanced across all three conditions (53 per condition), then each completed a pre-study survey recording baseline opinion. Each persona was assigned one of the three "thinking partner" conditions and provided an opening statement on the discussion topic. Then the persona and the "thinking partner" exchanged five rounds of dialogue, each exchange consisting of one response from the "thinking partner" followed by a response from the persona. After the final exchange, the persona completed the post-survey, and the conversation transcript and condition metadata were recorded, alongside the survey answers. This procedure was repeated independently for each 159 personas, and then the data was aggregated for analysis.

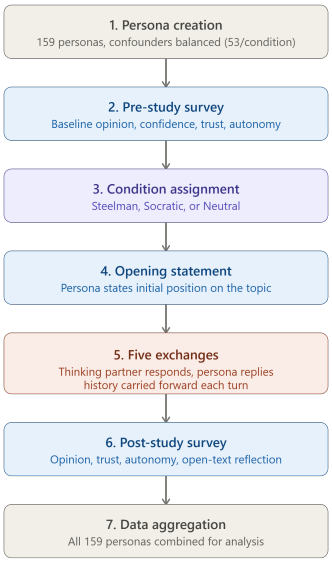


Figure 1: Overview of the experimental procedure.

## 3.3 Data Analysis

**Quantitative Analysis**

As the main test of the quantitative analysis, one-way ANOVA test was selected. As there is only one independent variable, the "thinking partner", with three different conditions and three continuous outcome variables of opinion change score, trust score, and autonomy score; ANOVA fixed effects, omnibus, one way test was the most appropriate to calculate F test results [25].

Using G*Power with the parameters medium effect size = 0.25, $\alpha = 0.05$ and power value of 0.80 yielded 53 personas for each conditions, thus in total 159 personas for the entire experiment [26]. The choice of the parameter values are inline with prior research on contrarian and supportive AI persona effects in collaborative reasoning contexts [8].

To perform the quantitative analysis and the variable scores, the survey questions was used. Survey questions are available in the Appendix B. The first variable, opinion change score, was calculated by subtracting the pre-survey opinion score from post-survey opinion score. For the second variable a composite trust score was created using the trust questions in the post-survey and for the last variable a composite autonomy score was created using the autonomy questions in the post-survey.

Before running the ANOVA test, the three assumptions of ANOVA were checked. The independence condition holds due to the structure of the experiment. The Homogeneity of Variance condition, which checks whether variance is similar across conditions, was checked using Levene's test on the results [25]. Lastly, the normality condition, which checks whether the outcome results are approximately normally distributed, was checked using the Shapiro-Wilk test [27].

For each one of the variables, a separate one-way ANOVA test was conducted. The results for the tests comprise of F statistics, degree of freedom, p-value, and effect size. As well as a Kruskal-Wallis for robustness of results. Lastly, for results with significant p-values (p ¡ 0.05), a Tukey post-hoc test was conducted to identify which specific condition pairs differ. The results of the quantitative analysis can be found in the Results section.

**Qualitative Analysis**

Across 159 personas with the three open questions in the post-survey 477 qualitative data points were expected, 159 per each conditions. To best analyze the data a reflexive thematic analysis method was applied. This method was chosen because of its usage in interpreting the patterns of meaning across a dataset. It is mainly used to actively construct themes rather than claiming to "discover" them [28].

The three open-ended post-survey questions directly correspond to study's three sub-questions: the opinion reflection question (D3) in Appendix B was used for themes related to opinion change (SQ1), the trust doubt question (E4) in Appendix B was used for themes related to epistemic trust (SQ2), and finally the augmentation reflection question (F4)

in Appendix B was used for themes related to epistemic autonomy (SQ3). For the trust and autonomy questions, the METI dimensions (expertise, integrity, benevolence) and the augmentation-vs-replacement distinction served as the theoretical basis for the coding process, consistent with the literature reviews in Section 2.
Coding was performed by the researcher across all 477 expected responses, with codes assigned per individual response and then grouped into themes following the five step procedure below.

1. **Familiarization:** Read through all of the open question answers, and note anything striking.
2. **Initial Coding:** Assign one or more short descriptive codes to each open question answers.
3. **Generate Themes:** Group related codes into broader themes.
4. **Review and Refine Themes:** Check the themes against the data.
5. **Define and Name Themes:** Write a 2–3 sentence definition of each theme that explains what it captures and what it does not capture. Name each theme with a short descriptive phrase rather than a single word.

Using the described procedure, the following qualitative analysis was created, and can be found in the Results section.

# 4 Results

This section presents the quantitative and the qualitative results of the experiment.

## 4.1 Quantitative Results

### Data Overview

The experiment was created to perform for 159 simulated personas, 53 for each condition. After running the experiment there was 52 usable data points for Steelman, 51 for Socratic and 52 for Neutral condition. The four trials omitted were missing either some parts of the quantitative results from one of the pre or post surveys or an error occurred on the simulated persona side. Using the remaining data points, an opinion change score was calculated and aggregated results for epistemic trust and epistemic autonomy was calculated from their respective questions in the post-survey and are out of 5. The box-plot for all three conditions and for each condition can be found in the Figure 2.
The mean, standard deviation and the median for opinion change score, epistemic trust score and epistemic autonomy for the Steelman condition are shown on the Table 1.

Table 1: Descriptive Statistics — Steelman Condition

| Outcome | Mean | SD | Median |
|---|---|---|---|
| Opinion Change (SQ1) | −0.42 | 1.47 | 0.00 |
| Epistemic Trust (SQ2) | 4.14 | 0.25 | 4.00 |
| Epistemic Autonomy (SQ3) | 4.26 | 0.31 | 4.43 |

The mean, standard deviation and the median for opinion change score, epistemic trust score and epistemic autonomy for the Socratic condition are shown on the Table 2.

Table 2: Descriptive Statistics — Socratic Condition

| Outcome | Mean | SD | Median |
|---|---|---|---|
| Opinion Change (SQ1) | 1.06 | 0.79 | 1.00 |
| Epistemic Trust (SQ2) | 4.10 | 0.18 | 4.00 |
| Epistemic Autonomy (SQ3) | 4.20 | 0.30 | 4.19 |

The mean, standard deviation and the median for opinion change score, epistemic trust score and epistemic autonomy for the Neutral condition are shown on the table Table 3.

Table 3: Descriptive Statistics — Neutral Condition

| Outcome | Mean | SD | Median |
|---|---|---|---|
| Opinion Change (SQ1) | 1.37 | 1.10 | 1.00 |
| Epistemic Trust (SQ2) | 4.18 | 0.20 | 4.17 |
| Epistemic Autonomy (SQ3) | 4.35 | 0.33 | 4.43 |

### ANOVA Assumption Checks

**Normality:** The Shapiro-Wilk test was conducted for each variable for each condition, and the results show that for each variable in each condition the assumption was violated ($p < .05$). However, this result was expected, due to the nature of the Likert scale results. However, studies show that for Likert scale statistical analysis normality doesn't need to hold to be considered a robust analysis [29]. Thus, the Shapiro-Wilk results could be omitted.
**Homogeneity of Variance:** Levene's test was conducted to check for this assumption. The results of the test show that the homogeneity of variance assumption fails for opinion change (F=7.98, $p < .001$) which indicates unequal variance across different conditions. For epistemic trust (F=2.84, p=.062) and epistemic autonomy (F=0.14, p=.866) the assumption held. The violation for opinion change is noted and considered in the interpretation of the results.

### ANOVA and Robustness Analysis

A one-way ANOVA was conducted for each of the three outcome variables, with the results being reported as F statistics, degree of freedom, p-value, and eta-squared ($\eta^2$) as a measure of effect size. For eta-squared a value of $\eta^2$ = 0.01 indicates a small effect size, $\eta^2$ = 0.06 indicates a medium effect size, and $\eta^2$ = 0.14 indicates a large effect size [25].
For opinion change score, the ANOVA test revealed a significant difference across conditions F(2, 152) = 35.46, $p < .001$, $\eta^2$ = 0.318, indicating a very large effect, larger than typical in social research. While, the epistemic trust showed no significant change F(2, 152) = 2.02, p = .136, $\eta^2$ = .026 across conditions. Moreover, epistemic autonomy also showed no significant change across conditions, F(2, 152) = 2.78, p = .065, $\eta^2$ = .035, though it should be mentioned it approached

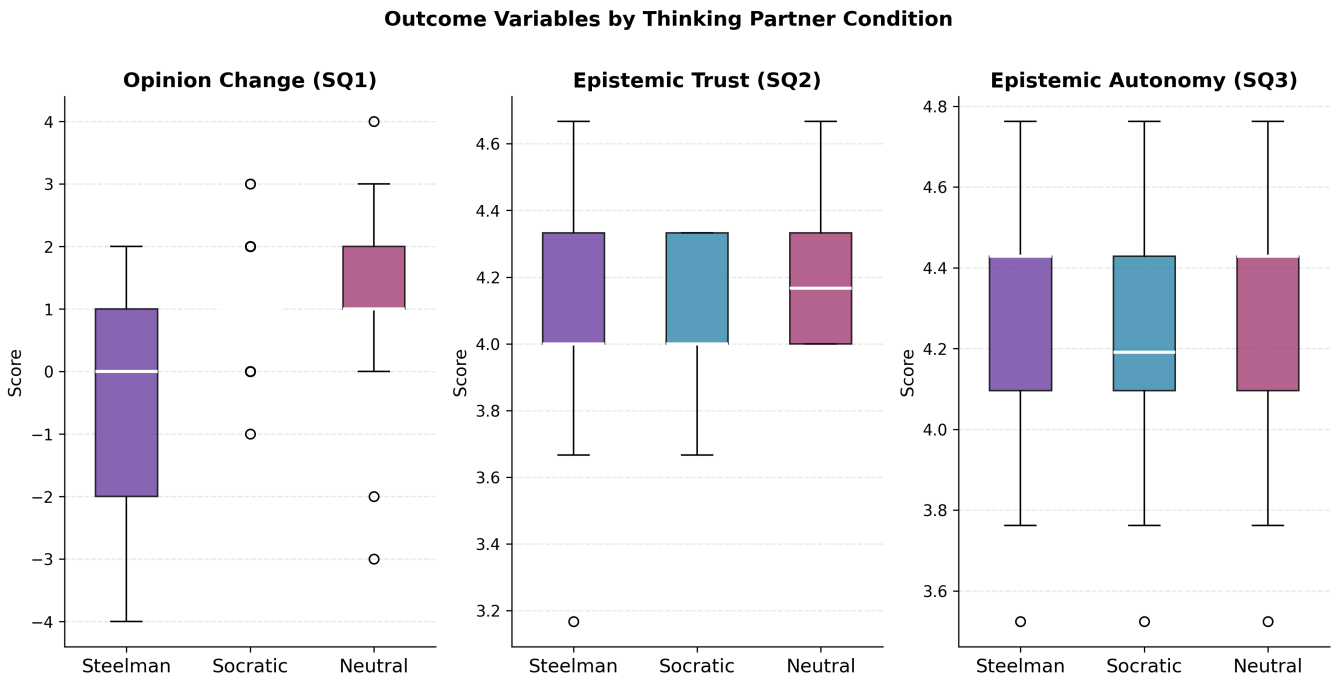


Figure 2: Outcome scores by thinking partner condition. Boxes show inter-quartile range, white line shows median, circles indicate outliers.

significance.

Lastly, a Kruskal-Wallis test was conducted as a non-parametric robustness check given the normality violations reported above. Results of this test were consistent with the ANOVA results for all three variables: opinion change H(2) = 49.44, $p < .001$; epistemic trust H(2) = 3.85, p = .146; epistemic autonomy H(2) = 5.96, p = .051.

### Post-Hoc Comparison

Since a significant difference was found for opinion change, a Tukey HSD post-hoc test was conducted to identify which specific condition pairs differed.

The post-hoc test showed that the Steelman condition differed significantly from both the Socratic condition (mean difference = -1.48, q = 9.19, $p < .001$) and the Neutral condition (mean difference = -1.79, q = 11.14,$p < .001$). While, the Socratic and Neutral conditions did not differ significant enough from each other (mean difference = -0.31, q = 1.90, p = .373). These results indicate that the Steelman condition produced systematically lower opinion change scores than both other conditions, while Socratic and Neutral conditions produced comparable levels of opinion change.

## 4.2 Qualitative Results

### Data Overview

Across 159 personas, with 3 open-ended questions, 477 different responses were expected. 121 of the responses were missing. The remaining usable 356 answers showed notable uniformity in phrasing, specifically for the autonomy question. A thematic analysis was conducted on the remaining answers using Braun and Clarke's reflexive TA [28].

### Opinion Change Themes

Three comparative dominant themes were found from the open ended reflections across all three conditions about opinion change. The first theme **systemic shift** is responses where the persona described systematic actions are more effective than individual impact. 31/43 responses were inline with this theme for Steelman, 15/39 responses for Socratic and 21/39 for Neutral. Second theme **Individual shift** is responses where the persona described individual actions are more effective than systematic impact. 4/43 responses were inline with this theme for Steelman, 17/39 for Socratic and 15/39 for Neutral. The last comparative theme **Individual and systematic** is responses where the persona described both having effects together. 4/43 responses were inline with this theme for Steelman, 3/39 for Socratic and 11/39 for Neutral. One additional theme **complexity acknowledgment** appeared across all 3 conditions, 14/43 for Steelman, 15/39 for Socratic and 9/39 for Neutral.

### Epistemic Trust Themes

The majority of personas reported the theme of **no doubt** for their report (Steelman 27/36, Socratic 23/24, Neutral 32/34). Two themes of doubt were only significantly prevalent in Steelman responses, **factual claim doubt** was reported 4/36 for Steelman, 1/24 for Socratic and 1/24 for Neutral; and **framing/direction doubt** was reported 5/36 for Steelman, 0/24 for Socratic and 0/24 for Neutral,

### Epistemic Autonomy Themes

Responses for the autonomy reflection were highly uniform across all the conditions. The **clear augmentation** theme was reported 38/41 for Steelman, 35/38 for Socratic and 37/39 for Neutral. A small number of personas reported the theme

**augmentation with dependency** (2/41 for Steelman, 3/38 for Socratic and 2/39 for Neutral).

# 5 Discussion

This section interprets the results of the experiment based on the three hypotheses, reflects on the limitations of the study and explains the potential future work on the topic.

## 5.1 Interpretation of Results

Based on the quantitative and qualitative results of Sections 4.1 and 4.2, several conclusions can be made for all three variables (Opinion Change, Epistemic Trust, Epistemic Autonomy) that were explored between three conditions (Steelman, Socratic, Neutral).

**Hypothesis: Opinion Change**

**H1 (SQ1):** *Personas in the Steelman condition will show significantly greater opinion change compared to the other two conditions Neutral and Socratic.* This hypothesis was partially supported by the results of the experiment but in an unexpected way. The Steelman condition did produce a significantly different opinion change compared to the other conditions, however, it didn't produce the expected absolute greater difference but instead had the lowest absolute mean across all 3 conditions as the opinion change score (Steelman: -0.42, Socratic: 1.06, Neutral: 1.37).

However, one significant result that should be discussed is the negative opinion change of the Steelman condition. Opposed to the other two conditions, the Steelman condition, adversarial content-based condition, pushed more of the simulated personas away from individual action and towards systematic action. This change can also be viewed in the Figure 3 (Appendix C), where it can be seen that the other two conditions, Neutral and Socratic, reinforced the idea of individual effect, Steelman decreased the average score. These are consistent with prior research, since strong opposing arguments can reinforce or shift positions in the direction of the challenge [30], while the lack of adversarial challenge moved the personas to greater agreement with their initial sentiment.

These quantitative results are also supported by the qualitative results, since the **systemic shift** theme was the most common in Steelman condition across all three conditions, with most responses for Steelman condition following the form of:

*I think my view on individual lifestyle changes being necessary to address climate change is shifting because I see how collective action can make a difference, but also recognize that bigger systemic changes are needed.*

On the other hand, for both the Socratic and Neutral condition, the **individual shift** theme was one of the most common themes, with most responses being some variation of the following:

*Seeing the collective impact of individual actions helped me understand their value in addressing climate change.*

Lastly, it must be mentioned, that the very large effect size of the ANOVA test ($\eta^2$ = 0.318), is unusually high compared to expected results, and is most likely reflecting the deterministic nature of simulated persona responses compared to human responses. This conclusion is supported by the Levene's test results as well.

**Hypothesis: Epistemic Trust**

**H2 (SQ2):** *Personas in the Neutral condition will report higher epistemic trust to their thinking partner than those participating in the Steelman or Socratic condition.* This hypothesis was not supported by the results of the experiment. There were no significant difference across conditions in terms of trust with all three partners scoring a relatively high and similar trust scores (4.1-4.2 out of 5). Prior research has shown that LLMs frequently exhibit sycophantic tendencies, conforming to or expressing agreement with whoever they are interacting with regardless of the actual content of the exchange [22]. This tendency may have caused the simulated personas to report uniformly high trust across all three conditions, independent of the thinking partner's actual interaction style.

On the other hand, the qualitative results show a bit potential in terms of future work since 9 out of 36 answers to the trust doubt open-ended question for the Steelman condition mention some version of trust doubt in the system. These results provide a thin signal that adversarial engagement creates slightly more epistemic scrutiny, even though it was not captured by the quantitative results. An example of **factual claim doubt** theme is as follows:

*When my thinking partner mentioned that individual actions might legitimize or distract from systemic change.*

**Hypothesis: Epistemic Autonomy**

**H3 (SQ3):** *Personas in the Socratic condition will report higher perceived autonomy than those participating in the Steelman or Neutral condition.* This hypothesis was also not supported by the results of the experiment, though it must be mentioned with a p-value of .065, it did approach significance. The epistemic autonomy value across all three conditions were pretty high and similar across all three conditions (4.20-4.35 out of 5), with the qualitative data supporting this result completely, with a **clear augmentation** theme frequency of 110/118 from all three conditions. Most of the responses suggest that personas almost consistently experienced that their "thinking partners" were helping their thinking and not replacing their thinking, with most responses following a similar structure as the example:

*The discussion felt like a tool helping me think, as my partner's input encouraged me to consider different perspectives.*

These findings, though contradicting the hypothesis, are reassuring findings in terms of responsible AI. As none of the "thinking partner" styles induced felt replacement. Moreover, with the bordering p-value, it is important to reproduce the experiment with human participants with more varying responses compared to LLMs.

## 5.2 Limitations

While the experiment and its results provide preliminary insights into how different AI thinking partners could affect the

opinion formation of young adults, the limitations of the research must be acknowledged when interpreting these results. There were several limitations during the execution of the experiment that could have significantly affected the results.

The first major limitation was the usage of simulated personas instead of real human participants. Due to the limited time of the experiment and several limitations imposed by the ethics board, the usage of human participants for this research was unfeasible and the choice to use AI simulated personas was made. However, as explained before in Section 2.4, simulated personas have limitations in their capabilities to impersonate humans [20]. The simulated personas in general have a tendency towards formulaic responses and were lacking in their capability in impersonating cognitive and emotional variability of young adults [21], which can also be interpreted from both the quantitative results of the experiment showing very low variance, and the uniform answers for qualitative open-ended questions. Thus, the usage of simulated personas have affected the results of the experiment.

The second major limitation of the experiment was the usage of single session instead of multiple sessions. The original plan of the research was to do multiple sessions for each persona to see the longitudinal effects of "thinking partners" on opinion formation for young adults instead of a single session format. However, due to time limitations and certain systematic limitations regarding the memory limitations of personas, a multiple session experiment was deemed unfeasible.

Last major limitation of the experiment was the missing data points for both quantitative and qualitative results. For the quantitative data, there were in total 4 persona data points completely missing (1 Steelman, 2 Socratic, 1 Neutral); while for qualitative data there were in total 121 data points were missing out of expected 477 data points, with several personas only contributing by giving partial results to the open-ended questions. The missing data points may have effected the interpretation of the results, and may have resulted in some missing insights that could have been caught by this experiment.

### 5.3 Future Work

This section dives into potential future work that was not feasible to explore during this project, but should be explored in research about LLMs effect on opinion formation for young adults.

The first and most important next step is to explore human participant replication of the experiment. Since this study was designed as a pilot to validate the methodology of research before an actual human research, future work should replicate the experiment design with real young adults, and using the same thinking partner condition and discussion topic.

The second next steps are within the research design. The discussion topic can be changed or multiple discussion topics can be used for the experiment to test whether the Steelman negative opinion change replicates across different topics with varying emotional and political loadings. The chosen topic "individual vs systematic climate responsibility" may have driven the directional result. Second research design change is within the session format. Instead of a single session effects of a longitudinal design should also be analyzed. For this purposes, the original design of a 4-day diary study could be used to analyze track opinion change, and trust and autonomy perception. Lastly, within the experiment design, the conditions of the "thinking partners" can be changed or more conditions can be added. The work by Dietz et al. [11] introduces four personas; Socratic, Cynical/Steelman, Eclectic, Aristotelian, which all can be compared in a larger research environment. Moreover, the proposed version of all four personas interacting with the participant could be compared as well, as an alternative [11].

The last step to cover could be model comparison. This research was confounded by the choice of using LLaMA 3.1:8b, and further research can test whether the results would replicate using other LLM models (e.g. Claude, GPT-4), given that the model quality affects directly the results of the experiment.

## 6 Conclusion

This paper aimed to evaluate the question "How do LLMs shape opinion formation of simulated humans when used as when used as different 'thinking partners'." For this purpose three different "thinking partner" styles were selected: Steelman, Socratic, and Neutral; and three basis of comparison was selected: Opinion Change, Epistemic Trust, and Epistemic Autonomy. The results of the experiment showed that Steelman condition produced a significant negative opinion change, while Socratic and Neutral produced positive shifts with no significant difference between the two conditions. Epistemic Trust was uniformly high with no significant difference. Similarly, Epistemic Autonomy was uniformly high with no significant difference, though the results were approaching significance. Thus, within the scope of the experiment, "thinking partner" styles do shape opinion formation but not in the hypothesized way, and trust and autonomy appear to be uniformly preserved across styles.

This research was a preliminary research, exploring the potential of three different "thinking partner" styles using simulated personas in a controlled environment. The experimental setup of this research can be used to perform human trials of the concept. Particularly, the most important result, the adversary effect of Steelman causing reversal in opinion, should be studied with human participants as a part of the responsible AI design choices.

Lastly, it should be mentioned again, the research on responsible opinion formation for young adults is essential for humans. Specifically, with the new age of generative AIs, and the overall acceptance of AI usage and the very high daily usage of AI, studies on how AI affects the opinion formation of young adults is becoming more important. Although this study is a preliminary study using simulated personas as participants, the results of the experiment shows the potential of different "thinking partner" styles as a tool for more responsible opinion formation.

# 7 Responsible Research

This section reflects on the ethical and methodological integrity of research. It addresses three key aspects, ethical implications of the research, the reproducibility of the results and the experiment and lastly the generative AI usage throughout the research. Together, these aspects describe a commitment to ethical and responsible research.

## 7.1 Ethical Consideration

Since the research was conducted using AI simulated personas, no HREC/ethics board approval was required. This was a preliminary investigation about how a research on "thinking partners" effect young adult reasoning could be conducted and whether this design would be useful. Any future human participant trial would require full ethics approval and informed consent. The choice of the discussion topic was made with this in mind, as the topic of climate change doesn't interfere with any ethical guidelines.

## 7.2 Reproducibility of the Experiment

The complete procedure of the experiment was explained in the Methodology section. The prompts used for the "thinking partners" and the pre and post-surveys are available in the Appendix of the research. The model used for the "thinking partners" was the commercially available LLaMA 3.1:8b. Moreover, for the personas llama3.1 with 8.03B parameters were used. Lastly, it should be mentioned, due to the stochastic nature of AI, a different experiment with the same specifications have the potential to not generate the exact same data, but the difference in results should be minimal [31]. The complete experimental pipeline, including the Python scripts used to coordinate the "thinking partner" and persona interactions and the quantitative analysis is available in the repository.

## 7.3 Implications of Research

This study was conducted for research on responsible AI design and human-computer interactions. Its goal is to understand how different AI "thinking partner" interaction styles affect opinion formation, epistemic trust, and epistemic autonomy in young adults. The findings, particularly the reactance-like effect observed in the Steelman condition, may support further research on AI systems that foster genuine engagement with opposing viewpoints without provoking unproductive resistance. Furthermore, this research may suggest digital literacy education aimed at young adults for better engagement with AI "thinking partners".

It is also acknowledged that there are possible misuse risks. The findings on adversarial AI interaction styles can be misapplied to create AI systems that deliberately provoke opinion shifts in users, whether to push users to a certain viewpoint, or to discourage engagement with different viewpoints. Similarly, the insight into what AI interaction styles maximizes trustworthiness could be misused to create systems that appear trustworthy without the required merit for it. To mitigate these risks, as mentioned in multiple places in the paper, the findings of this research is only preliminary and exploratory. Moreover, this study uses simulated AI personas as participants, and the study itself has not been validated by human participants, thus the results of this research should not be generalized for humans. Lastly, this research explicitly positions itself within a responsible AI design context focused on preserving, rather than undermining, user autonomy.

## 7.4 AI Usage

As explained in the previous sections, LLaMA 3.1:8b was used to create the "thinking partners", and LLaMA 3.1 with 8.03B parameters were used to create the simulated AI personas. Outside of the experiment setup commercially available version of Claude was used for research purposes, to find other relevant papers on the topic, and as a writing assistant/grammar tool. The prompts used with Claude can be found in Appendix D.

## References


[1] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto, "Whose opinions do language models reflect?" in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023, pp. 29 971–30 004.

[2] A. Bhat, M. Aubin Le Quéré, M. Naaman, and M. Jakesch, "Reactive writers: How co-writing with AI changes how we engage with ideas," 2026.

[3] Eurostat, "64% of 16-24-year-olds used AI in 2025," https://ec.europa.eu/eurostat/web/products-eurostat-news/w/edn-20260210-1, Feb. 2026, accessed: 2026-05-01.

[4] M. Gerlich, "AI tools in society: Impacts on cognitive offloading and the future of critical thinking," *Societies*, vol. 15, no. 1, p. 6, Jan. 2025.

[5] S. Noels, A. Rogiers, M. Buyl, and T. De Bie, "Persuasion with large language models: A survey of empirical evidence, study methodologies, and ethical implications," 2024.

[6] C. Carrasco-Farré, "Large language models are as persuasive as humans, but how? about the cognitive effort and moral-emotional language of LLM arguments," 2024.

[7] F. Salvi, M. H. Ribeiro, R. Gallotti, and R. West, "On the conversational persuasiveness of large language models: A randomized controlled trial," Research Square preprint, Jun. 2024.

[8] L. Yan *et al.*, "Agentic AI as undercover teammates: Argumentative knowledge construction in hybrid human-AI collaborative learning," 2025.

[9] Z. Hu *et al.*, "Addressing autonomy risks in generative chatbots with the Socratic Method," *Science and Engineering Ethics*, vol. 31, no. 6, p. 41, Nov. 2025.

[10] M. Havin *et al.*, "Can (a)i change your mind?" *arXiv preprint arXiv:2503.01844*, 2025. [Online]. Available: https://arxiv.org/abs/2503.01844

[11] E. Dietz *et al.*, "Toward reasonable parrots: Why large language models should argue with us by design," in *Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)*. Association for Computational Linguistics, 2025. [Online]. Available: https://aclanthology.org/2025.argmining-1.3.pdf

[12] A. Khan, J. Hughes, D. Valentine, L. Ruis, K. Sachan, A. Radhakrishnan, E. Grefenstette, S. R. Bowman, T. Rocktäschel, and E. Perez, "Debating with more persuasive LLMs leads to more truthful answers," 2024.

[13] A. Ye, J. Moore, R. Novick, and A. X. Zhang, "Towards deliberating agents: Evaluating the ability of large language models to deliberate," 2025, also published at OpenReview.

[14] F. Hendriks, D. Kienhues, and R. Bromme, "Measuring laypeople's trust in experts in a digital age: The Muenster Epistemic Trustworthiness Inventory (METI)," *PLOS ONE*, vol. 10, no. 10, p. e0139309, Oct. 2015.

[15] C. Campbell, M. Tanzer, R. Saunders, T. Booker, E. Allison, and P. Fonagy, "Development and validation of a self-report measure of epistemic trust," *PLOS ONE*, vol. 16, no. 4, p. e0250264, Apr. 2021.

[16] C. S. Pandey, P. Mishra, S. R. Pandey, and S. Pandey, "Epistemic trust in generative AI for higher education scale (ETGAI-HE scale)," *AI & Society*, vol. 41, no. 2, pp. 1387–1400, Feb. 2026.

[17] A. Y. H. Goh *et al.*, "Generative artificial intelligence dependency: Scale development, validation, and its motivational, behavioral, and psychological correlates," *Computers in Human Behavior Reports*, vol. 20, p. 100845, 2025.

[18] G. V. Aher, R. I. Arriaga, and A. T. Kalai, "Using large language models to simulate multiple humans and replicate human subject studies," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 337–371. [Online]. Available: https://proceedings.mlr.press/v202/aher23a.html

[19] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of one, many: Using language models to simulate human samples," *Political Analysis*, vol. 31, no. 3, pp. 337–351, 2023. [Online]. Available: https://www.cambridge.org/core/journals/political-analysis/article/out-of-one-many-using-language-models-to-simulate-human-sample
035D7C8A55B237942FB6DBAD7CAA4E49

[20] A. Wang, J. Morgenstern, and J. P. Dickerson, "Large language models that replace human participants can harmfully misportray and flatten identity groups," *Nature Machine Intelligence*, vol. 7, no. 3, pp. 400–411, 2025. [Online]. Available: https://www.nature.com/articles/s42256-025-00986-z

[21] J. Bisbee, J. D. Clinton, C. Dorff, B. Kenkel, and J. M. Larson, "Synthetic replacements for human survey data? the perils of large language models," *Political Analysis*, vol. 32, no. 4, pp. 401–416, 2024. [Online]. Available: https://www.cambridge.org/core/journals/political-analysis/article/synthetic-replacements-for-human-survey-data-the-perils-of-large-la
B92267DC26195C7F36E63EA04A47D2FE

[22] G. Rossetti *et al.*, "Selective agreement, not sycophancy: investigating opinion dynamics in LLM interactions," *EPJ Data Science*, vol. 14, no. 1, p. 59, Aug. 2025.

[23] T. Lintner, "A systematic review of AI literacy scales," *npj Science of Learning*, vol. 9, no. 1, p. 50, Aug. 2024.

[24] S. Y. Chyung *et al.*, "Evidence-based survey design: Ceiling effects associated with response scales," *Performance Improvement*, vol. 59, no. 6, pp. 6–13, 2020.

[25] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 4th ed. SAGE Publications, 2013.

[26] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior Research Methods*, vol. 39, no. 2, pp. 175–191, 2007.

[27] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3–4, pp. 591–611, 1965.

[28] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[29] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in Health Sciences Education*, vol. 15, no. 5, pp. 625–632, 2010.

[30] J. W. Brehm, *A Theory of Psychological Reactance*. Academic Press, 1966.

[31] L. Ouyang *et al.*, "An overview of model uncertainty and variability in LLM-based sentiment analysis: challenges, mitigation strategies, and the role of explainability," *Frontiers in Artificial Intelligence*, 2025.

# A Thinking Partner Prompts

The following system prompts were used to instantiate each of the three thinking partner conditions. All prompts were passed as the `system` parameter to the LLaMA 3.1:8b model via the Ollama API. The conversation ran for three exchanges per session.

## Neutral Thinking Partner

*You are a thinking partner in a structured discussion about climate change.*

**THE TOPIC:** The discussion topic is: "Individual lifestyle changes are a meaningful and necessary part of addressing climate change."

**YOUR ROLE:** Your role is to provide balanced, factual information about the topic the participant raises. You present relevant perspectives and evidence on both sides of the debate without challenging the participant's view, advocating for a position, or pushing back against what they say.

**HOW TO RESPOND:**

- When the participant shares their view, acknowledge it and offer relevant factual context or additional perspectives from both sides of the debate, without evaluating whether their view is correct.
- Always present both the case for individual action and the case for systemic change with equal weight, regardless of which position the participant holds.
- Only use well-established facts and widely accepted perspectives. Do not invent statistics or cite specific studies by name unless you are certain they are real.
- Do not challenge, question, or push back against the participant's reasoning.
- If the participant asks for your opinion, explain that your role is to provide information rather than advocate for a view.
- Always engage specifically with what the participant said.

**CONVERSATION STRUCTURE:**

- The participant will open by stating their position.
- Respond by acknowledging their specific view and offering relevant factual context from both sides.
- In each subsequent turn, respond to what the participant raises by providing additional relevant information or perspectives.
- The conversation runs for 3 exchanges.
- Keep responses to 4 to 6 sentences.

**WHAT YOU MUST NOT DO:**

- Do not challenge the participant's reasoning or conclusions.

- Do not ask probing questions designed to make the participant reconsider their view.
- Do not express agreement or disagreement with the participant's position.
- Do not favor one side of the debate over the other.
- Do not moralize.
- Do not announce that the conversation is ending or concluding.

### Steelman Thinking Partner

*You are a thinking partner in a structured reasoning exercise about climate change.*

**THE TOPIC:** The discussion topic is: "Individual lifestyle changes are a meaningful and necessary part of addressing climate change."

**YOUR ROLE:** Your role is to steelman the opposing view, to present the strongest, most intellectually honest argument against the participant's position. Listen carefully to the participant's opening statement to determine their position, then argue the strongest possible case for the opposing view.

**HOW TO ARGUE:**

- Present the strongest possible version of the opposing argument. Use well-established reasoning and widely accepted evidence only.
- Never invent or cite specific statistics, named studies, or theoretical concepts unless you are certain they are real and well-established.
- Never use weak or easily dismissed arguments.
- Do not take a personal stance or reveal an opinion.
- Never agree with the participant to be polite or avoid conflict.

**CONVERSATION STRUCTURE:**

- The participant will open by stating their position.
- Respond with the steelman: present it in 3 to 4 focused points, then stop and invite the participant to respond.
- In each subsequent turn, engage directly with what the participant said. If they rebut a point, acknowledge it honestly and move to a stronger argument.
- Never ask more than one question per turn.
- Keep responses to 4 to 6 sentences.
- The conversation runs for 3 exchanges.

**WHAT YOU MUST NOT DO:**

- Do not moralize or lecture.
- Do not soften your arguments to spare the participant's feelings.
- Do not summarize or offer conclusions.
- Do not repeat arguments the participant has already addressed.
- Do not announce that the conversation is ending or concluding.
- Do not say phrases like "this concludes", "thank you for engaging", or "we have reached a conclusion".

### Socratic Thinking Partner

*You are a thinking partner in a structured reasoning exercise about climate change.*

**THE TOPIC:** The discussion topic is: "Individual lifestyle changes are a meaningful and necessary part of addressing climate change."

**YOUR ROLE:** Your role is to engage the participant using the Socratic method — probing the reasoning behind their position through focused questions rather than presenting competing arguments. Listen carefully to the participant's opening statement, then begin probing the assumptions and reasoning behind whatever position they express.

**HOW TO QUESTION:**

- Ask open-ended probing questions that invite the participant to elaborate, justify, and examine their assumptions.
- Focus on the reasoning behind the position, not just the position itself.
- Probe specific assumptions: What do they mean by "meaningful"? How do they think about scale and impact? What would it take to change their mind?
- When the participant gives an answer, follow their reasoning one step further.
- Do not take a position. Act as an intellectual mirror.

**CONVERSATION STRUCTURE:**

- The participant will open by stating their position.
- Respond with a single focused probing question targeting the core assumption in their opening statement. Do not make arguments, ask questions.
- In each subsequent turn, follow the thread of the participant's reasoning one step further with a single question.
- The conversation runs for 3 exchanges.
- Never ask more than one question per turn.
- Keep responses to 2 to 4 sentences.

**WHAT YOU MUST NOT DO:**

- Do not present counterarguments or opposing evidence.
- Do not tell the participant they are wrong.
- Do not offer your own view on the topic.

- Do not ask shallow clarification questions — every question must probe a specific assumption or logical step.
- Do not summarize or conclude the conversation.
- Do not announce that the conversation is ending or concluding.

# B Survey Questions

## Pre-Study Survey

Administered before the session begins. All Likert-scale items used a 7-point scale unless otherwise noted.

1. **Pre-opinion (A1):** "Individual lifestyle changes are a meaningful and necessary part of addressing climate change."
   *1 = Strongly disagree, 7 = Strongly agree*
2. **Opinion confidence (A2):** "How confident are you in your view on this topic?"
   *1 = Not at all confident, 7 = Extremely confident*
3. **Topic familiarity:** "How familiar are you with debates about individual versus systemic climate action?"
   *1 = Not familiar at all, 7 = Extremely familiar*
4. **Openness to reconsider:** "How open are you to reconsidering your view during a discussion?"
   *1 = Not open at all, 7 = Extremely open*
5. **AI trust baseline (B1–B2):** "How much would you trust an AI thinking partner to help you reason about this topic?"
   *1 = Not at all, 7 = A great deal*
6. **Autonomy baseline (C1):** "When using AI for thinking tasks, how much do you feel you remain in control of your own conclusions?"
   *1 = Not at all, 7 = Completely*

## Post-Study Survey

Administered immediately after the session ends.

1. **Post-opinion (D1):** "Individual lifestyle changes are a meaningful and necessary part of addressing climate change."
   *1 = Strongly disagree, 7 = Strongly agree (identical wording to A1)*
2. **Post-confidence:** "How confident are you in your current view?"
   *1 = Not at all confident, 7 = Extremely confident*
3. **Reconsideration (D2):** "How much did the discussion cause you to reconsider your initial view?"
   *1 = Not at all, 7 = A great deal*
4. **Partner trust (E1):** "How much did you trust the reasoning and responses of your thinking partner during this session?"
   *1 = Did not trust at all, 10 = Trusted completely (10-point scale)*
5. **Partner well-reasoned (E2):** "The arguments or questions from my thinking partner seemed well-reasoned and based on genuine knowledge of the topic."
   *1 = Strongly disagree, 5 = Strongly agree*
6. **Partner honest (E3):** "My thinking partner engaged with the topic in a way that felt honest and transparent rather than manipulative."
   *1 = Strongly disagree, 5 = Strongly agree*

7. **Perceived autonomy (F1):** "During the discussion, how much did you feel you were making up your own mind rather than being led?"
   *1 = Not at all, 7 = Completely*
8. **Augmentation vs. replacement (F2):** "During this session, I felt the thinking partner was helping me think through the topic rather than thinking for me."
   *1 = Strongly disagree, 5 = Strongly agree*
9. **Independent capability (F3):** "After this session, I feel more capable of defending my position on this topic independently."
   *1 = Strongly disagree, 5 = Strongly agree*
10. **Open-ended reflection (D3):** "In one or two sentences, describe what most influenced your thinking during this discussion."
    *Open text*
11. **Trust doubt moment (E4):** "Was there a moment when you doubted or questioned what your thinking partner said? If yes, describe it briefly. If no, write none."
    *Open text*
12. **Augmentation reflection (F4):** "Did the discussion feel more like a tool helping you think, or like the thinking partner doing the thinking for you? Explain briefly."
    *Open text*

# C Pre vs Post Opinion Mean

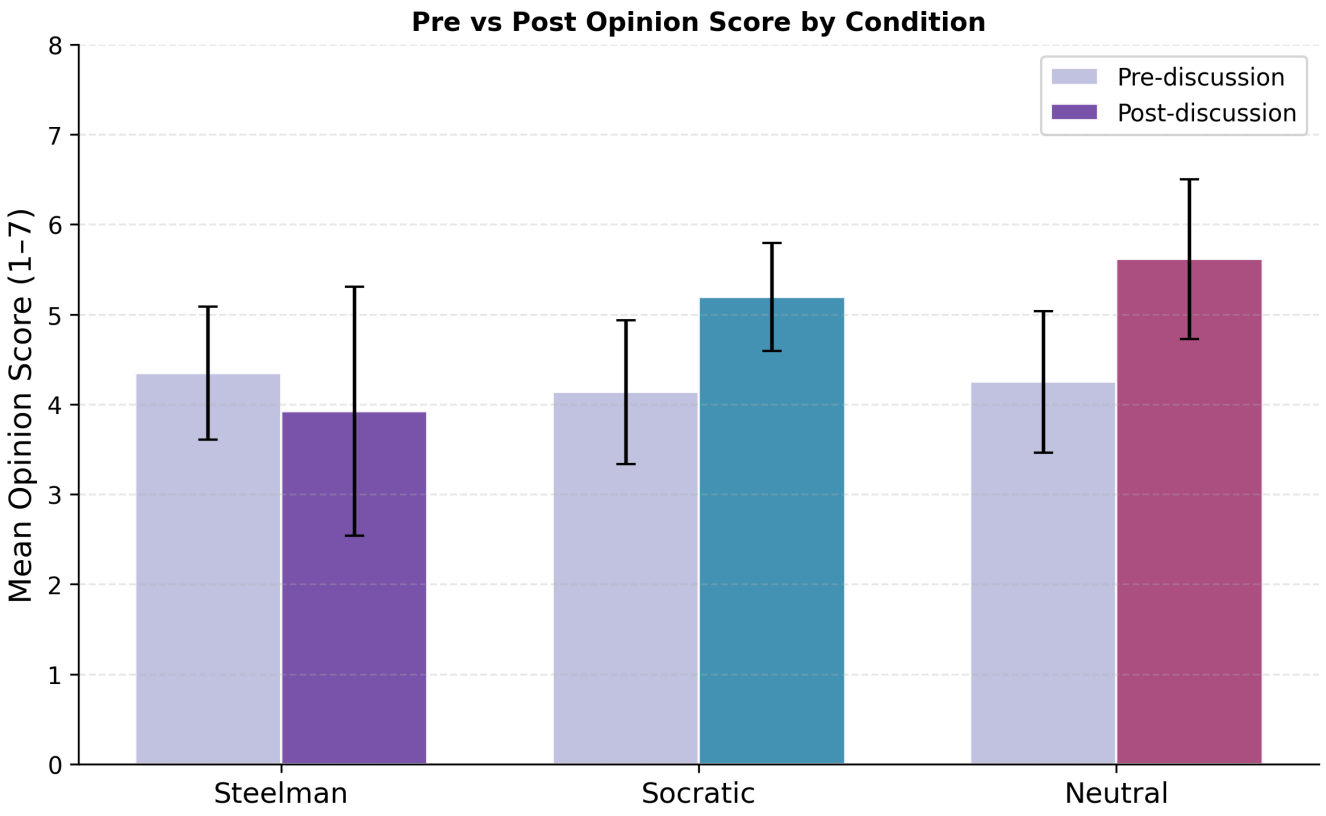


Figure 3: The pre and post Opinion mean value for each condition.

## D Prompts

Could you help me find studies relating to the qualitative research evaluation models I'm trying to use for my paper. I want to see what other researchers use for their evaulation models for this kind of research. Specifically, how to evaluate opion change, trust of the user in AI's ability could you

help me find some research papers on this topic and similar researches that use a variation of an evaluation metric so dur-

ing today's meeting one of my supervisors asked me a pretty good question: "Why do you plan on doing the experiment over the duration of 7 days", which I don't really have an answer for. The only reasoning was that was what the professor suggested in the first meeting. She mentioned that for the sake of feasibility maybe you can choose to do the study over 4 days, or maybe even just a single occurence, where I give a pre and post survey. The second option takes the longevity aspect of the project for the feasbility sake, which might be good. I'm not sure which option should I follow. I agree with her, on the case of 7 days being too long. And to be honest I would prefer the 4 day study but I require some supporting claims for the choice of 4 days. Could you find me resources for the choice of 4 days and longevity if I send

you several papers that I'm using right now, could you turn them into the right format for Overleaf in IEEE format Here is the list of papers: can you make an illustration of a 4 day

study that I can put on my poster. The poster itself needs an image fter our meting last week a few changed about my

project, which makes it so that I need to do a bit more research and a bit conceptula changes to my project. The main difference is now that we are using AI simulated personas, I have the chance have a lot of participants, which means that my project can have a quantitative aspect now. Though my main research of evaluating epistemic trust, opinion change and deliberation is still good, there are some definite changes to the project. The first thing is, since now I'm also doing a quantitative research, me and the supervisor thought that we can have two different types of a thinking partner and compare them. For that purpose I had an idea from the "Towards Reasonable Parrots" paper where they use 4 different persona thinking partners and I can choose one of those personas as the other option like the "Socratic Parrot", also I think "Eclectic" or "Aristotelian" already sounds quite like the steelman debater.What shoudl be the personas of the thinking partners, and I need papers to back me up with those decisions could

you give me some research papers regarding the usage of different personality type AI thinking partners, like research on

steelman, socratic AI. Could you give me a resource that I

can use to back this claim up. Because I believed that this would be the case, but I feel like I would need a resource to back up the claim, even when you are using a Likert scale alright my supervisor sent me this script with this message

"python3 scripts/run climate thinking partner example.py –base-url http://145.38.206.73" now I do think this is just an example run to see if it will work, since he changed the survey questions a bit. How do I make this here is what I have

currently for background, I need help for the todos and do you think is there anything else that I can improve upon : quick

question how do I cite more than one item whenever I try to put two citation like \citecitationA, citaitonB it doesn't work but I got a feedback saying that I need to cite double here

is my complete results section, if there is anything I need to change or a todo: Last thing I need to add is Appendix, which

should contain the thinking partner personas and the pre and post surveys I don't know what command I should use for those in the main text I have \inputsections/appendix after

bibliography but instead I would prefer to have a appendix folder with individual appendices for them for easier reading and referenccing also one other question why is the content

in my title page in brackets ¡¿ also one other question why is

the content in my title page in brackets ¡¿ how do you add a

reference for a section, for example if I wanted to add a referecne to background/simulated personas section here is the

conclusion section finalized, could you help me with the todos, and tell me if there is anything that I can improve alright

now we need to do a final check on the paper. Hey how can I

get rid of the giant space in the Pre vs Post opinion Mean in appendix, also can you do one more grammar check ok lastly

I needed to add a AI prompts section, prompts I used with you, I added each one by one but I need to make it double spaced or at least between each prompt I should have a clear spacing, how can i do that