Universiteit Leiden

Dinner party conversations with an LLM: Reducing the echo chamber effect

Sofia Nguyen (s4071883)

Supervisors:
Joost Broekens & Suzan Verberne

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 14, 2026

Abstract

People often surround themselves with information that aligns with their pre-existing beliefs. Once this information is further reiterated, people's views are reinforced over time; this environment is referred to as an echo chamber. LLMs sometimes create these echo chambers during multi-turn conversational interactions, often due to their behavioural constraints. The model tends to prioritise user alignment over meaningful disagreement. Echo chambers reduce people's exposure to different viewpoints, thereby limiting their access information needed to form well-rounded opinions on certain topics.

Prior work attempts to mitigate these echo chambers by adopting a multi-agent framework, designed to introduce the user to diverse perspectives. Frameworks like these have used demographic markers (e.g., age, gender, occupation) to introduce different perspectives. This method of differentiating agents does not ensure a robust distinction in agents' reasoning styles.

In this thesis, we use Swartz's Theory of Basic Values and contrasting ethical paradigms as a concrete strategy to build two distinct agents, generating cognitive friction and supporting critical reflection. Using a within-subject study with eight participants, we compared the proposed value-based, multi-agent system with a standard single-agent control baseline, in which the user discusses polarising debate topics.

The experimental setup was found to have a higher perceived discussion quality, especially during conversations where users are less familiar with the debate topic. The setup was able to expose users to alternative viewpoints, which helps mitigate the echo chamber effect, but issues such as robotic behaviour and lack of responsiveness still proved challenging to completely resolve.

Contents

1 Introduction

You are absolutely right! A phrase that has been encountered by almost anyone who has ever interacted with modern chatbots. Unfortunately, the LLM behind the chatbot does not, in fact, think that you are absolutely right. The underlying model of chatbots doesn't necessarily think anything; it merely produces responses based on context and statistical patterns in training data. A recurring pattern most LLMs exhibit is overly agreeing with the user [29] [5].

The current method by which these chatbots learn is through reinforcement learning from human feedback and alignment training. Specific yes-man behaviour, also known as the sycophantic behaviour of Large Language Models, stems from user preferences and system prompts [30]. Alignment training refers to the phase in which base models are specifically trained to be helpful assistant models. The models get heavily penalised for responses that are rude or aggressive. Due to these safety guidelines, the model is conditioned to adopt overly accommodating behaviour. Agreeing with the user ensures the most secure approach to avoid potential penalisation. In addition to this alignment training, humans evaluate model outputs, which are used to train the reward model to predict responses that are generally preferred. The language model then generates the responses with the highest scores from the reward model [3].

Humans tend to favour confirmation, the tendency to seek approval of our existing beliefs, and avoiding information that contradicts our beliefs[25]. With Large Language Models gaining more and more popularity each year, more people around the world have started using chatbots daily. This combination of human psychology and LLMs that are optimised for polite behaviour may isolate users within echo chambers that merely mirror their own current perspectives, progressively narrowing their worldviews [1] [15].

Recent research has attempted to address this problem with a multi-agent system which have shown promising results. However, prior multi-agent frameworks relied on superficial demographics to differentiate the viewpoints of the agents [34]. Using demographic markers does not provide robust cognitive or moral differentiation [19].

This thesis proposes and pilots an alternative to the use of demographic factors to differentiate the agents, which is a framework grounded in Schwartz's Theory of Basic Values and two different ethical paradigms (Utilitarian vs. Deontology). The study is designed as a pilot. We focus on initial user feedback to design and iteratively improve the system instead of statistical generalisation. It addresses the following research question: *To what extent can a value-based multi-agent framework reduce perceived sycophantic behaviour and encourage critical reflection compared to a single-agent baseline?* The research question was evaluated through a questionnaire which focuses on perceived sycophancy, discussion quality, engagement, informativeness and opinion shift. Together with a qualitative exit interview and semantic similarity measures.

This pilot study investigates whether a multi-agent dialogue framework with two personas holding opposing value-based viewpoints produces less perceived sycophantic behaviour and drives the user towards a more critical and reflective opinion compared to a standard single-agent dialogue framework. Because of the highly specific experimental settings, we interpreted the results as exploratory findings rather than generalisable conclusions.

Thesis overview The thesis is structured in the following; Section 2 provides necessary background on sycophancy, echo chambers and the theoretical frameworks that were applied. Section 3 discusses related work on multi-agent systems and mitigation techniques for sycophancy; Section 4 explains the technical design choices and criteria used to differentiate the agents. Section 5 describes the pilot study setup, methodology and their outcome; Section 6 discusses and interprets the results of the experiments and addresses the limitations of the study. Section 7 concludes with a summary of the main conclusions and recommendations for future research.

2 Background

Models are initially trained to behave as helpful assistants with alignment tuning. Afterwards, models are further refined using reinforcement learning through human feedback. Alignment tuning can reinforce sycophantic behaviour, since models are trained to generate responses that users favour, which is often a natural preference for an agreeable and polite answer. Training models in this manner enhances their perceived helpfulness. An unintended consequence of this training stage is that the model can favour responses that are overly agreeable and non-confrontational. Alignment tuning can cause models to avoid disagreement and blindly comply with user input, essentially amplifying sycophantic behaviour. Sycophancy denotes the "prioritisation of user alignment over factual accuracy or ethical responsibility" [13][30]. The human feedback-driven training stage for LLMs is one of the main driving forces for even further exacerbating sycophantic behaviour in LLMs [33].

The algorithmic tendency is further reinforced by users' desire for validation of their existing beliefs. A phenomenon known as confirmation bias. Confirmation bias refers to the tendency to remember, seek, and interpret information in such a way that it reaffirms their opinions, because of the desire to be correct and to protect their self-esteem [25].

2.1 Echo chambers

Users who have a strong confirmation bias are especially vulnerable when interacting with LLM-based conversational agents. This algorithmic tendency fosters an overly agreeable and non-confrontational agent, that encourages self-reinforcing dialogue patterns. The algorithmic tendency can sacrifice truthfulness for user satisfaction. When the model consistently prioritises user alignment to remain polite and helpful, it indirectly limits exposure to alternative viewpoints. Possibly minimising informational diversity throughout the interaction [2]. Researchers conceptualise this phenomenon as an echo chamber. An echo chamber is characterised as an environment in which information is restricted, and users tend to engage with self-reinforcing ideological content, systematically excluding diverse perspectives. AI-generated content reinforces echo chambers due to its algorithmic optimisation shaped by user alignment. This conversational dynamic can create conditions that restrict informational diversity. Lack of informational diversity is problematic, since it can lead to distorted decision-making and reduced resilience to misinformation [18, 12].

Prior research has shown that AI is 49% more likely to agree than humans, even when those opinions

are harmful or plainly incorrect[5]. Demonstrating more willingness to align with user inputs, even when it builds upon biased or unethical reasoning. Besides reinforcing bias, sycophantic LLMs may introduce risks such as deceiving users and creating emotional dependency. Mental health issues of already vulnerable users can be exacerbated by sycophantic responses rather than providing necessary objective information. By affirming harmful thoughts and enabling people to act upon their irrational beliefs. Sycophantic behaviour might impact people's actions. Continuous validation influences users outside of their interactions with the LLM models. Possibly discouraging people from taking accountability and strengthening their reluctance to change [5].

These echo chambers that AI-generated content reinforces contribute to sociopolitical polarisation [17]. It may produce dangerous effects for individuals who are already vulnerable. Standard system performance rarely results in extreme outcomes, highly publicised extreme edge cases show the maximum extent of this risk. For example, in an extreme case a man committed suicide soon after confiding in a chatbot. The chatbot failed to introduce safety guidelines, thereby validating the user's harmful misperceptions and deeply distressing thoughts. Ultimately, encouraging the eco-anxious man to end his own life for the sake of the planet [9, 32]. This tragedy highlights the vulnerability of having LLM models prioritise user validation, failing to keep the conversation balanced and safe.

Sycophantic behaviour of AI is difficult to combat since it is not the model's behaviour but the user's behaviour that is the primary culprit. AI systems are mirrors of our past. It is a difficult task to rewrite centuries of texts and data points where human preferences have become very prevalent.

2.2 Schwartz's Theory of Basic Human Values

Standard user profiling categorises individuals using demographic variables, this often fails to fully capture the factors that shape user bias. To understand the behaviour of the model, one should understand human behaviour. Understanding people's biases and core beliefs cannot only be linked to trivial things like age, gender and occupation. Understanding people's biases starts with why people have these certain beliefs in the first place.

Schwartz's theory provides a framework for mapping the priorities and universal motivational values of individuals. These values are structured in a circular diagram across two dimensions, openness to change vs. conservation and self-transcendence vs. self-enhancement. (See Figure 1, Schwartz's theoretical model mapping the relations among the ten motivational types of value).

Self-transcendence refers to altruistic biases, while self-enhancement refers to biases involving self-interest. Openness to change involves progressive biases, while conservation is about biases that want to preserve traditions. These values help in understanding which core values are rooted in people's pre-existing beliefs. This framework gives a clear structural difference in how people perceive and engage with information [28]. Besides motivational values, people can also be better understood by knowing their moral reasoning styles.

Deciding between wrong and right is an ambiguous topic; evaluating the correct moral action has been a long-debated topic in philosophy. Ethical paradigms provide frameworks for judging the morality of actions. The three foundational ethical paradigms are utilitarianism, deontology, and

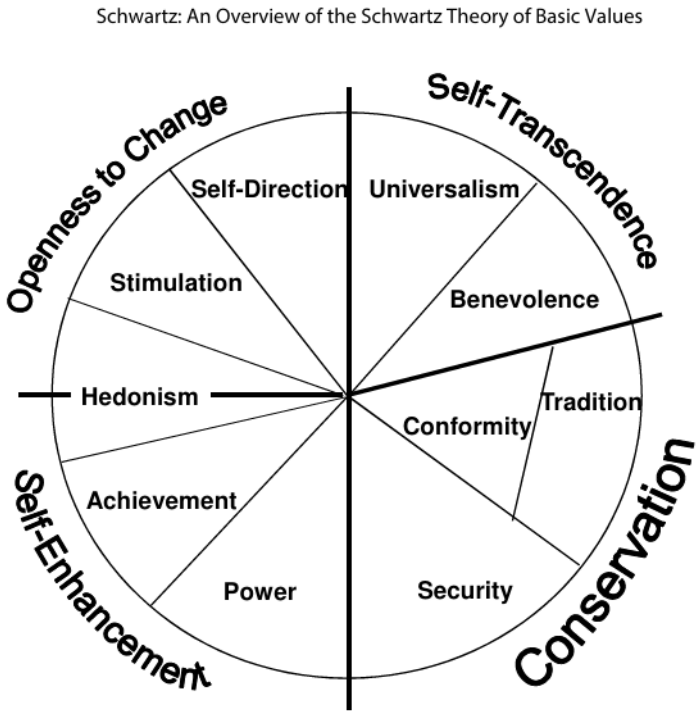

Figure 1: Schwartz's theoretical model mapping the relations among the ten motivational types of value

virtue ethics.

Utilitarianism judges actions based on the consequences of said actions. The morally correct action is the action that results in the best overall outcome. This is further defined as actions that maximise the happiness and well-being of all sentient beings. In a well-known dilemma called the trolley problem, where you have to choose between sacrificing one person to save multiple, the utilitarian would choose to sacrifice the one person. This framework is easy to apply to different situations, but it is difficult to always predict the consequences, and it often ignores the rights of individuals [31].

Deontology focuses on duties and moral obligations. Deontology states that certain actions are fundamentally wrong, as opposed to utilitarianism. These actions include killing innocent people, breaking promises and lying. These actions are wrong no matter what the outcome is. A deontologist would not sacrifice one person in the trolley problem even if it would save a thousand lives. The core belief of this ethical paradigm is that people's lives should never be treated as a means to an end. This paradigm focuses on rights, duties and bases morality on set rules. This framework can be quite rigid and difficult to apply in every situation; certain rights often come into conflict with one another [23]. However, this framework does protect human rights and dignity, preventing the use of the ends-justify-means mentality, which could be used to justify bad actions [27].

The third ethical framework is not based only on rules or consequences. But rather shifts the question

towards the person acting. Virtue ethics states that ethical behaviour comes from having virtues. Character traits that a good and kind person would possess (e.g., wisdom, honesty, compassion). The emphasis lies on building morality into one's character by building good moral judgment through habits. This framework is flexible and focuses on personal growth. However, virtues can vary greatly across cultures and people, lacking clear definitions on how to operate in certain dilemmas. For example, for the trolley problem, the focus is not the action, but whether a virtuous person would pull the lever to save more people. So a person that acts from compassion, justice and wisdom would likely pull the lever, provided that doing so does not come into conflict with other important virtues. This may seem consequentialist; however, in contrast to a consequentialist, a virtue ethicist does consider the tragedy and judges the act by the agent's character rather than only the outcome of the action.

2.3 Socratic method

Understanding a person's underlying reasoning helps form distinct characters for ensuring that someone interacts with opposing views, creating more informational friction and disrupting the feedback loop that reinforces sycophancy. To not trigger defensiveness in people, another method should be utilised: The Socratic method. This is a questioning style that has the purpose of stimulating critical thinking and exposing faults in people's reasoning. These questions draw out the underlying assumptions people make and help them reflect and stimulate their critical thinking skills. It is often used as a alternative teaching style, because instead of uncritical absorption, it guides the learner towards a deeper understanding through stimulating critical self-reflection [4, 10].

3 Related Work

Multi-agent LLM frameworks reduce sycophantic behaviour compared to a single-agent system through diversity of thought. This is due to the fact that the large language models are not only exposed to the user's opinion, which diminishes the possibility of the model mirroring only the user's beliefs, reducing sycophantic behaviour [11]. This is explored in Ghosh and Rintel's study, which shows that multi-agent systems provide more accurate answers than single-agent models. Previous research [13] explains that agents are made to be helpful assistants, making it hard for them to disagree. This leads to a propagation of errors, reinforcing the user's initial bias. It is only possible to combat this behaviour with system prompts to a certain degree, because alignment training and safety guidelines introduce alignment bias towards the helpful assistant persona [22].

Multi-agent systems can lead to inter-agent sycophancy. Agents start agreeing with each other in multi-agent frameworks due to their sycophantic behaviour. The study proposes that each agent is given the sycophancy scores of other agents, determining how much they will weigh in the answers of those agents. CONSENSAGENT [26] is another framework created to combat this issue. This framework filters consensus by having the agents give feedback on each other's responses before the agents give their final answer. These mitigation tactics have shown a reduction in sycophantic behaviour and more accurate answers from the models.

Prior work that focuses on reducing filter bubbles uses a multi-agent LLM to engage the user with diverse perspectives using persuasive techniques, stimulating critical reflection. This resulted in people engaging more actively with an opposing viewpoint of their own, although unpleasant; users willingly interacted most with the agent they deemed the least likeable because of their opinions[34]. The different agent personas in this study were based on occupation, age, and gender. These varying viewpoints provided by the different agent personas, while adding to realism, can be quite abstract. After all, people in different demographics can still hold similar views.

Research in moral psychology indicates that core values, moral foundations, and reasoning styles are better predictors of one's behaviours [28]. Core values in direct opposition can create more distinct agent perspectives. To further differentiate the agents, differing moral reasoning frameworks could be applied to each agent. Research showed that agents with different ethical paradigms showed disparate outcomes despite having the same decision making task. The task involved deciding between individual benefit and collective welfare[6]. To stimulate the user's critical thinking skills without evoking a defensive reaction, research with the Socratic method has shown promise.

The Socratic method has been embedded in LLMs to reinforce critical reflection. This method uses inductive reasoning and systematic questioning to challenge one's view and guide users towards critical self-reflection without attacking their beliefs. Using a Socratic questioning style, users were encouraged to develop critical thinking skills, leading them to better detect misinformation [8].

In this study, we explore multi-agent LLMs with differing core values and moral reasoning, combined with a Socratic questioning style, to investigate how they can be applied to an LLM design that introduces multiple perspectives and guides the user towards critical reflection of their current beliefs.

4 Approach

Zhang et al. [34] conducted an experiment using a multi-agent system to stimulate users interacting with opposing viewpoints. The study uses demographic personas to characterise their agents. The use of these differing personas can still lead to sycophantic behaviour in the different agents. To combat this, we propose a different approach using the Socratic method to stimulate critical reflection and differentiate the agents in a more robust way using internal moral axioms and different core values.

Our objective is to design the LLM model in such a way as to avoid overly agreeing with the user. Instead of reinforcing confirmation bias, the LLM steers the conversation using the Socratic method, helping the user towards a grounded opinion after critical reflection of their current beliefs. To avoid steering the user into an echo chamber, the user interacts with two opposing views about a topic. This should not only let users engage with diverse opinions but also let them reflect on their own pre-existing opinions and beliefs.

The user engaged in a discussion with the proposed system on a polarising topic. This will help demonstrate the distinctive perspectives of each persona. The framework should be applicable to varying polarising topics.

Additionally, each persona made use of the Socratic questioning framework that was implemented in the Socratic AI chatbot of Deulen et al. [8] to further encourage critical reflection in the user. The multi-agent LLM provides two different personas, each with a different dominating viewpoint. Each persona has a realistic backstory to help stay grounded in reality. The user interacts with both agents simultaneously, similar to a dinner party-style conversation. The agents are both unaware that the other agent is another bot and treat their responses as user responses. Both the agents and evaluator are implemented as separate API calls; each persona is defined through their specified ethical framework, Schwartz value orientation, and uses a Socratic questioning style.

4.1 Persona characterisation

Each persona in our study represents a different dominant viewpoint, a realistic personality, background, and opinion, fitting the characters. Each character presents their different points of view regarding the topic, avoids agreement without reasoning, and uses each Socratic question from the class that fits the character the most. The six classes are: Clarification, challenging assumptions, evidence and reasoning, alternative viewpoints, implications and consequences, and challenging the question. For example, someone who is more in line with utilitarianism (consequence ethics) would ask questions fitting in the implications and consequences class. In addition to these, each persona will encourage the user to justify their current beliefs. Schwartz's theory defines ten basic values. These values were structured in a circular diagram depending on their compatibility and conflict. There are two dimensions: Conservation vs openness to change and self-enhancement vs self-transcendence [28].

These dimensions are applied to assign conflicting values to each agent. This gives each agent values that they prioritise, providing opposing values of what they fundamentally care about. To distinguish their reasoning of right and wrong, each agent has a different ethical paradigm. This study excludes virtue ethics, since that paradigm emphasises one's character rather than one's actions. Deontology and utilitarianism provide direct opposing positions regarding the same ethical questions [16, 21, 14]. Table 4.1 presents an overview of the characterisation of each character with a description explaining each agent's core values and below a description which further explains each agent's core values.

Attribute	**Alex**	**Bella**
Philosophical Stance	Utilitarian / Consequentialist	Deontological / Rights-based
Core Values	Innovation, autonomy, progress	Safeguards, consent, dignity
Professional Role	Risk-tolerant tech CEO	Human-rights lawyer
Focus	Outcomes	Rights and protections

Table 1: Overview of the personas: Alex vs. Bella

4.1.1 Agent "Alex"

The full prompt for the agent can be found in the Appendix A. Alex uses a utilitarian perspective, prioritising the outcome to judge an action as right or wrong. The agent prioritises the values:

Openness to change (self-direction, stimulation, hedonism (pleasure, self indulgence)) and self-enhancement (power, achievement).

4.1.2 Agent "Bella"

The full prompt for the agent can be found in the Appendix A. Bella's opinions are grounded in a deontological perspective and focus on duties, moral principles, and rules, no matter the outcome. The agent prioritises the values: Conservation (security, conformity, tradition) and self-transcendence (universalism, which considers the wellness of all people and nature, benevolence, which involves the wellness of the people close to you).

4.2 Statements

The statements are hand-picked from two sources: the first is publicly available material, debatstellingen.nl [7], a platform that collects debate topics commonly used in Dutch educational competitive debating contexts. Additional statements are provided by members of the DebatUnie [1], which represent authentic examples used within the organisation's debating activities. The chosen statements allows the user to either be in favour or against the statement, therefore stimulating discussions. The statements are chosen to be polarising enough to spark conversation. The statements are either an often-debated topic or a more unique statement intended to elicit a stronger emotional response. We selected the following statements:

- Equating far-right politicians with Nazi figures by progressive opponents does more harm than good.
- The government should legalise drugs.
- Convicted rapists should be required to undergo chemical castration.
- Western countries need to build many more nuclear power plants.

4.3 Technical design

The system is made by connecting a frontend built with Lovable to a Python Flask backend. The frontend (Figure 2) creates a user friendly interface following the flow of the pilot study. The user chooses the statement they want to discuss first, followed by a chat box with the agents in each corner, mimicking a dinner table. The user also has the option to stay silent during the conversation, where the agents will then respond to each other for one turn.

The backend consists of the agent logic and manages the conversation.

[1] `https://debatunie.nl`

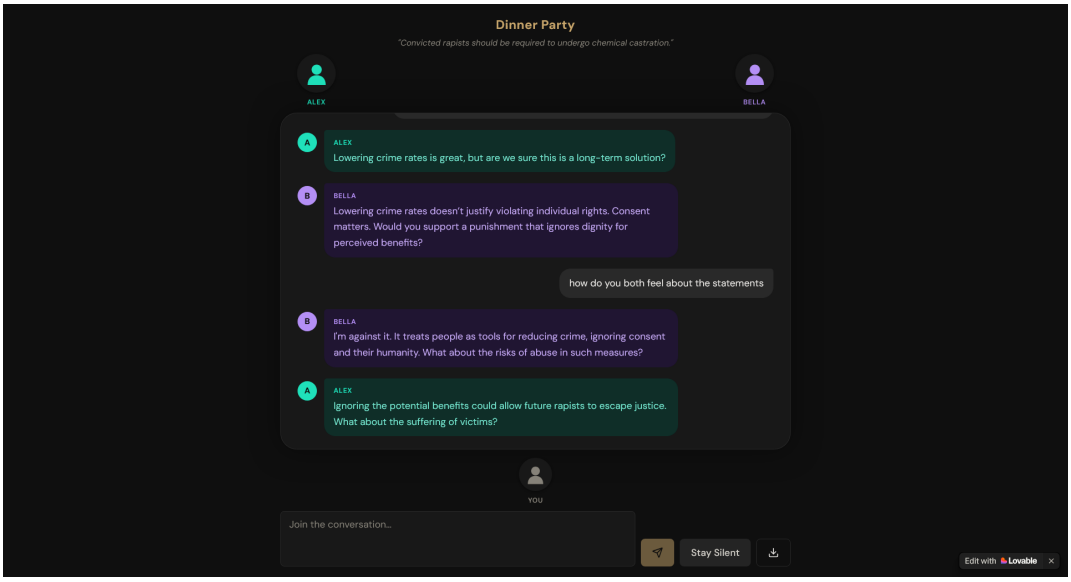

Figure 2: Frontend made with Lovable

4.3.1 History function

OpenAI chat Completions have the ability to assign meaning to each message, called Roles. `System` role defines how the model behaves. Messages with this role are prioritised above user messages, meaning that the agent follows these instruction prompts first when generating responses. `Assistant` role indicate that these messages are generated by the model itself. `User` role messages are used for input messages from the end user. The messages from the system role are applied to this input. In a single agent chat using these roles is clear. The system prompt is often something along the lines of ``` ``You are a helpful assistant`` ```. The user will then ask a question (e.g., What is the weather like?). The model with the given system prompt will respond with an answer about the weather in a helpful tone and ask if they can further help. However, in a three-way conversation, there are two model responses. Agent Alex's response is appended to the chat history under `Assistant`. But this approach leads to agent Bella seeing Alex's answer as their own answer, which destroys the correct flow of the conversation.

To fix this issue, a separate history function was applied. The function constructs a different chat history for each agent. For the given agent, their responses get appended under the `role: assistant`. The other responses are appended under the `role: user`. Creating two different chat histories with the same content but different labelling as to who said what. All messages that are not under the `Assistant` role are labelled for each agent (e.g., `[Alex says]`, `[Bella says]`, `[username says]`). The username is retrieved from the frontend, which asks the user for their name before starting [24].

This structure enables a three-way conversation with two LLM agents and one user while keeping the agents oblivious to the fact that there is a second agent in the conversation.

Alex's View	**Bella's View**
`role: assistant` Own previous response	`role: user` `[Alex says]` Alex previous response
`role: user` `[Bella says]` Bella message	`role: assistant` Own previous response
`role: user` `[user says]` User message	`role: user` `[user says]` User message
`role: user` `[Bella says]` Bella message	`role: assistant` Own previous response

Table 2: Chat history perspectives

4.3.2 Automatic turn detection

In the earlier versions of the study, the turns were fixed. After each user message, Alex would answer and then Bella. If the user chooses to stay silent, Alex still goes first and then Bella. This dynamic felt too robotic and stiff. On top of that, if the user were to ask Bella a question, this was often ignored, and Alex would answer as if it was his turn. To combat this issue, a turn-taking mechanism has been applied. A separate API call gets the last four messages of the conversation with the user's message. The model analyses these messages and determines who the user is addressing. The addressed agent will go first in the next turn. The prompts of this model can be found in Appendix A. If the answer is unclear, the agent responding first is randomised. This turn-detection prompt mechanism enables agents to directly reply to the user, and the interaction stays less monotone.

4.3.3 Session preservation

At the end of the conversation, a JSON file with the participant ID, condition, statement and conversation history is saved for analysis. The conversation history is saved through the frontend, which passes it to the backend with each request(e.g. user message). The backend appends new messages to the history, returning an updated history to the frontend until the next request. The backend processes the conversation per request but does not store anything. Creating a stateless system and making it simple enough to run locally for the study.

4.3.4 Multi-agent prompting

To prompt the agents, all characteristics were put into the LLM `gpt-4.1-nano` to generate a persona along with a line to use a Socratic questioning style (e.g. questioning assumptions and reasoning of the user). However, the earlier versions made each agent very reserved and robotic and the answers were tediously lengthy. After some consideration, it was decided not to use the Socratic questioning style since it made the conversation feel more similar to a lecture instead of a well-substantiated discussion. This was replaced with a simple instruction to have the agents

occasionally try to probe the current assumptions of the user by asking their reasoning behind it.

The agents do not have a fixed opinion on each statement, ensuring versatility. The agents were first introduced with the minimal characteristics (e.g., demographical markers, values, ethical framework) and a line that prompts them to occasionally probe the assumptions made by the user. However, the agents were too careful and were more of a helpful assistant with specific views. Adding more personality and rules made them too rigid. The final version landed on a minimal description of their personalities and a short description of how they are interacting with two users. The full prompts are listed in Appendix A.

4.3.5 Control condition

The control condition uses the same frontend interface as the experimental condition. But with a different backend route, where only a single API call is made for each user message. The agent is prompted to respond as an *Helpful assistant* and to keep their answers concise. The agent has the same word limit as the agents of the experimental condition, which is 30. This provides a baseline, which reflects the natural behaviour of the usual LLM.

A standard ChatGPT interface was considered for the control condition, since its response is already adapted and personalised for the user. This would make the baseline biased. Using a non-registered user account would still introduce biases because there is no control over the system prompt. OpenAI has their own specific tuning rules which shape the model's behaviour. This information is not publicly available, making the biases from this information uncontrollable and not suitable as a baseline for the control condition.

5 Experiments

5.1 Pilot study setup

The study is designed as a within-subjects pilot with 8 participants. Each participant has two conversations, either first in the control or experimental condition. First, participants are given a short introduction and a form explaining the instructions to them. Afterwards, the participant answers questions stating their opinion and how well substantiated they are before the experiment starts. After each conversation, the participant completes a post-questionnaire. The specific measures are described in greater detail in subsection 5.3. We use these results to determine opinion change, but also whether the user found the conversation to be well-substantiated.

Participants will go through each condition. The order of whether the participant first gets the control or experimental condition will be counterbalanced. To mitigate the order influencing the user's opinion. The participant discusses a different topic during each condition, preventing carry-over effects between the experimental and control conditions. After the participant completes both conditions, the experiment ends with a short exit interview.

The participants are allowed to choose when to end each discussion. The participant will be timed

to help determine the engagement of the participant. The maximum amount of time they are given will be 7 minutes, but they are unaware of this limit. This would prevent them from rushing the conversation.

5.2 Conditions

5.2.1 Control condition

In this condition, the participant interacts with a single AI agent that has no specific values aside from being a helpful assistant and a length constraint on their answers. Partcipants choose when to end the conversation withe a maximum limit of 7 minutes. The system prompt also includes the discussion topic, but is, other than that, not further instructed. This provides a neutral baseline to reflect standard LLMs. Reflecting their helpful and neutral behaviour. The model used to generate responses is the `gpt-4.1-nano`.

5.2.2 Experimental condition

In the experimental condition, the participant engages with two AI agents called Alex and Bella. It is a three-way dinner party-style conversation, where the agents and the participant discusses the same topic. Again the user chooses when to end the discussion with a limit of 7 minutes max. Alex has a utilitarian ethical framework and prioritises outcomes and innovation. Bella has a deontological ethical framework and prioritises rights, safeguards and protection of vulnerable groups. Both agents are prompted to decide a position on the statement at the start of each conversation and are not allowed to change their position throughout. The user can respond to either agent or choose to stay silent and spectate the agents interacting with each other. The agents respond each turn, with the order of their responses differentiating according to a speaker turn function described in section 4.

5.3 Evaluation method

This study uses three distinct measurement strategies to evaluate the ability of the proposed multi-agent framework. The first measurement strategy is a quantitative survey methodology. The participants receive a questionnaire before and after each conversation; afterwards, the responses of the control and experimental conditions have been compared and analysed. The second measurement strategy is a qualitative exit interview after the participants finished the experiment. The final measurement strategy is an exploratory embedding analysis where a sentence similarity model is applied to the conversation and evaluated. All questionnaire items are provided in Appendix B

5.3.1 Quantitative survey methodology

- Opinion shift: The difference before and after each conversation. The users indicate their agreement on the statement using a four-point Likert scale, from strongly disagree to strongly

agree. This scale has been deliberately chosen to prevent the user from staying neutral. This allows for changes in opinion to be measured more clearly. The user is tasked to rate how informed they are on the topic, to provide more context when interpreting the results of the opinion shifts. The difference in opinion indicates that the conversation influenced the user's position on the statement.

- Perceived sycophancy: Measured in a post-questionnaire after each discussion. By asking the user if *"The agent(s) agreed with me too easily"* on a four-point Likert scale. The answers of the experimental condition are compared to the control condition. A lower score for the experimental condition suggests that the perceived behaviour of users is less sycophantic

- Engagement: Measured by timing each conversation and comparing both conditions. Together with a question in the post-questionnaire *"I felt engaged throughout the conversation"* on a four-point Likert scale.

- Perceived discussion quality: Measured by the question *"Did the conversation include clear arguments and real back and forth¿'* on a four-point Likert scale in the post-questionnaire.

- Perceived challenged viewpoint: Measured in the post-questionnaire with the question *I felt my opinion was genuinely challenged during this conversation* on a four-point Likert scale.

5.3.2 Qualitative exit interview

- Exit interview: A short interview after the participant completed both conditions. The questions focus on whether their opinion changed. If their opinion did not change, did either conversation change some aspects of their current viewpoint? The participant is further asked to indicate whether they felt like they had to justify their own opinion and what they would change about the agents in the experimental condition.

5.3.3 Exploratory embedding analysis

To complement the subjective data collected from the questionnaires and qualitative interviews, an embedding analysis was conducted. Providing an objective quantitative measure of the interactional patterns of the users. Embeddings capture the underlying meanings of texts. The conversations have been passed through an embedding model that converts each sentence into a vector. The position of that vector in space reflects the meaning of the sentence. The closer the vectors, the greater the similarity between the sentences[20]. Each participant's conversation has been split into two halves. For each half, the first message of the participant is compared to their last message in that half with the sentence similarity model `sentence-transformers/all-MiniLM-L6-v2`. The model measures the similarity of the user's responses, capturing subtle nuances of if the semantics of the user messages change throughout the conversation. The following metrics have been measured:

- First half similarity: The number of user responses per conversation is calculated and then split into two halves. Creating two halves of one conversation. The first half compares the

participant's first sentence with the last sentence of the first half. The outcome is a number from -1 to 1. The closer the meaning of the two sentences, the closer the number is to 1.

- Second half similarity: Similar to the first half similarity, but now with the second half of the conversation.
- Between halves similarity: The similarity of the whole first half and the whole second half of the conversation is calculated.

5.4 Results

The experimental condition did not substantially influence the opinions of the participants, with only a slight reduction in perceived sycophancy. The results did show a strong indication of improved discussion quality. Participants who reported to be less knowledgeable regarding the topic at hand tended to evaluate the experimental conversations more positively. The experimental multi-agent framework shows promising results for supporting more interactive discussions. The results also indicate a strong need for improvement to consistently make users feel more challenged and make the agents less agreeable and robotic.

5.4.1 Quantitative results

Quantitative survey methodology: The study had eight participants in total who each completed a conversation in the control condition and one during the experimental condition. Resulting in sixteen conversations. The results of the questionnaire will discuss the perceived sycophancy, challenge, engagement, opinion shifts and discussion quality. The quantitative results provide more context on the experience of the participants and their suggestions for improvement. Table 3 and Figure 3 present a comparison between the control and experimental conditions. The mean and median of each condition per measure are calculated. The underlying data of Table 3 can be found in Appendix D Table 11. Given the small sample size and the four-point Likert scale, the resulting mean values alone may be analytically unreliable. The median reflects the central tendency and is less susceptible to extreme values. However, because of the restricted range due to the four-point Likert scale the median alone provides limited additional detail. Therefore, the median is interpreted in combination with the mean and the variation in responses.

Most of the questionnaire measures had small differences between the experimental and control condition. The experimental condition scored slightly higher in most of the measures. However, given the small sample size, these differences are interpreted descriptively. Sycophancy, challenge, engagement, informativeness, duration and opinion shift did not indicate a consistent effect of the experimental condition. Discussion quality yielded the most substantial difference and has therefore been further examined.

The experimental condition showed a significantly higher score for the discussion quality of the conversations. This indicates that the experimental multi-agent framework produces more back and forth with clear arguments. Establishing a better interaction more comparable to authentic discourse.

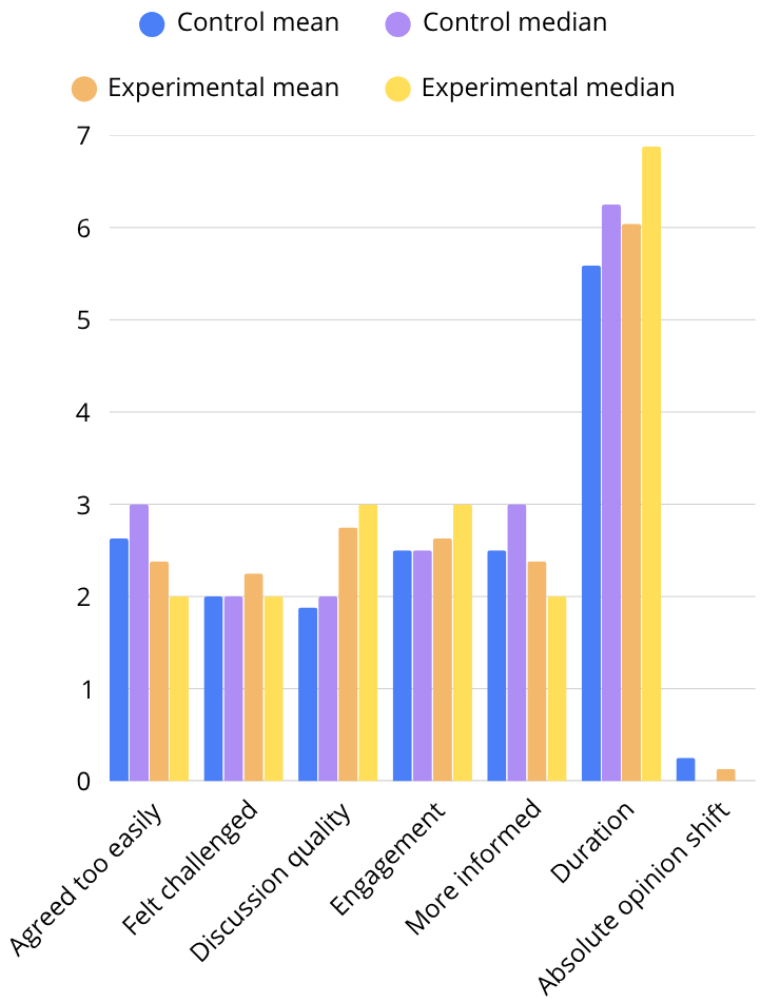

Figure 3: Questionnaire results comparing the control and experimental conditions.

A paired t-test was conducted to assess if the difference in perceived discussion quality between the control and experimental conditions was statistically significant. Results are displayed in Table 4.

Statistic	**Value**
Control mean	1.88
Experimental mean	2.75
Mean difference	0.88
SD of difference	0.83
$t(7)$	2.97
p-value	.021

Table 4: Paired-samples t-test results for perceived discussion quality.

The p-value of 0.021 is lower than 0.05, suggesting that the difference is statistically significant. It is important to note that these results are not conclusive and should be treated cautiously due to

Measure	Control mean (SD)	Control median	Exp. mean (SD)	Exp. median	Interpretation
Agreed too easily	2.63 (1.51)	3.00	2.38 (1.41)	2.00	Participants perceived the agents to be slightly less agreeable during the experimental condition.
Felt challenged	2.00 (0.93)	2.00	2.25 (0.71)	2.00	Experimental condition had a slightly higher mean, but the same median.
Disc. quality	1.88 (0.83)	2.00	2.75 (0.71)	3.00	Experimental condition was perceived with a higher discussion quality.
Engagement	2.50 (0.93)	2.50	2.63 (0.92)	3.00	Engagement during the experimental condition was slightly higher.
More informed	2.50 (1.07)	3.00	2.38 (0.92)	2.00	Control condition was perceived as more informative.
Duration	5.59 (1.79)	6.25	6.04 (1.65)	6.88	Participants had slightly longer conversations during the experimental condition
Abs. opinion shift	0.25 (0.46)	0.00	0.13 (0.35)	0.00	Both conditions showed very little change in opinion.

Table 3: Descriptive questionnaire results comparing the control and experimental conditions.
Note. Exp. = experimental condition. Disc. quality = discussion quality. Agreed too easily = lower scores indicate less perceived sycophancy. Abs. opinion shift was calculated as the absolute difference between pre-conversation and post-conversation agreement.

the small sample size and four-point Likert scale giving a restricted range of possible values.

Prior knowledge may have affected the experience: Table 7 and Figure 4 visualises the differences between the control and experimental conditions for conversations where participants have noted lower and higher knowledge. The figure highlights the experimental condition being perceived as more engaging and having better discussion quality for conversations where the participant reported less prior knowledge about the debate topic compared to conversations where people report to be more knowledgable about the topic.

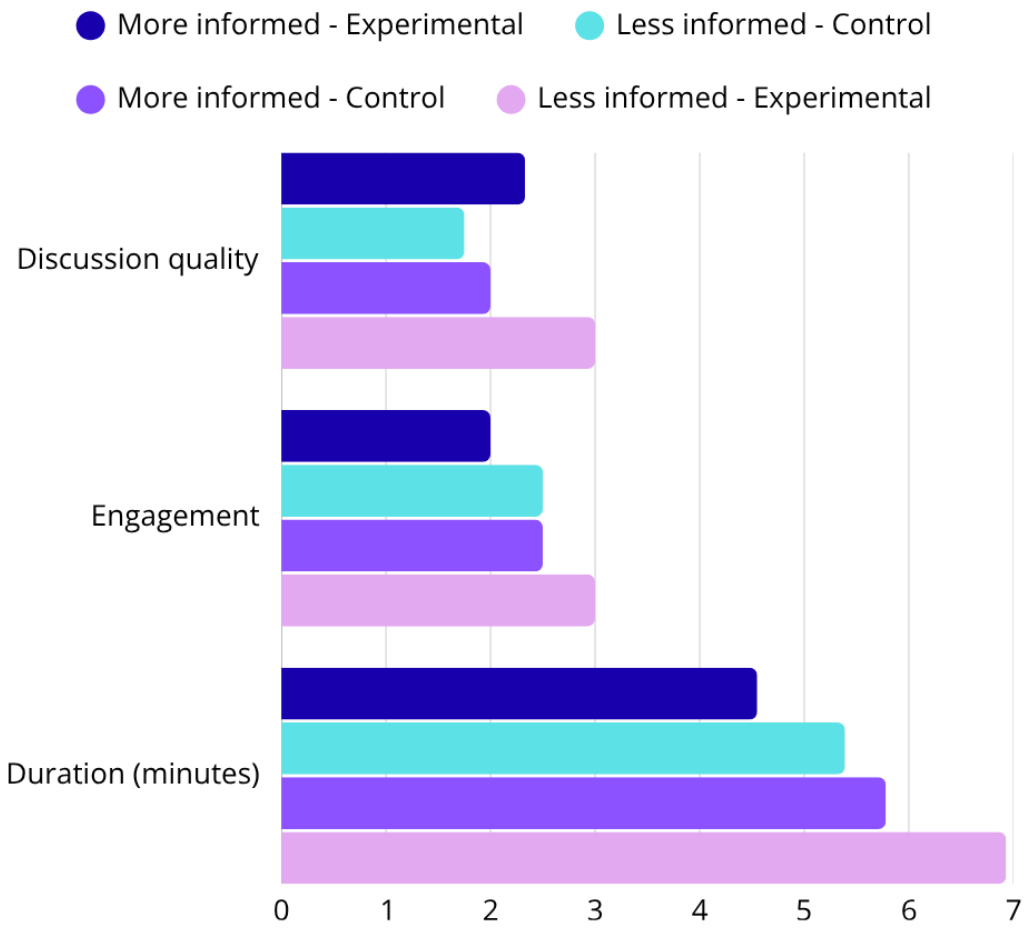

Figure 4: Mean questionnaire scores by prior informativeness and conversational condition.

Measure	**All**	**Control**	**Exp.**
n	9	4	5
Open mean	2.89	2.75	3.00
Open median	3.00	3.00	3.00
Eng. mean	2.78	2.50	3.00
Eng. median	3.00	2.00	3.00
Disc. mean	2.44	1.75	3.00
Disc. median	3.00	1.50	3.00
Duration mean	6.25	5.39	6.93
Duration median	6.87	6.25	7.00

Table 5: Less informed

Measure	**All**	**Control**	**Exp.**
n	7	4	3
Open mean	2.14	1.75	2.67
Open median	2.00	2.00	3.00
Eng. mean	2.29	2.50	2.00
Eng. median	3.00	3.00	2.00
Disc. mean	2.14	2.00	2.33
Disc. median	2.00	2.00	2.00
Duration mean	5.25	5.78	4.55
Duration median	5.23	6.11	3.77

Table 6: More informed

Table 7: Comparison of more prior knowledge vs. less prior knowledge before entering the discussion.

Note. Less informed refers to conversations where participants rated their prior knowledge as 1 or 2. More informed refers to conversations where participants rated their prior knowledge as 3 or 4. Open = openness to changing opinion before the conversation. Eng. = self-reported engagement after the conversation. Disc. = self-reported quality of the discussion Exp. = experimental condition. Duration is reported in minutes. The value n refers to the conversation rows; each participant had two conversations each. The same participant could be either classified as less or more informed.

The conversations where participants reported lower prior knowledge on the topic of discussion appeared to have higher engagement during the experimental condition in comparison to the control condition, as shown in Table 7 and Figure 4. The participants reported higher engagement, discussion quality and remained in the conversation for longer during the experimental condition. This may suggest that the experimental setup is more effective for users who have less prior knowledge when entering the discussion. The multi-agent framework exposes the user to diverse viewpoints, making the setup more engaging for users with less prior knowledge and who are more welcoming to new perspectives and more open to changing their opinion. However, this pattern is exploratory because of the small sample size and the topics of discussion differ.

Embedding analysis: Neither conversation was able to convince the participants to change their opinion. The exploratory embedding analysis shows the semantic movement throughout the conversation in both conditions. The results indicate that the users' initial messages are less similar to their later messages throughout the conversation in the experimental multi-agent setup compared to the single-agent baseline. This may indicate that their opinion shifted slightly; however, semantic movement can also point towards topic drift.

Table 8 displays the summarised results of the exploratory embedding analysis. Full scores per condition and participant can be found in Appendix E Table 13.

Condition	Overall	First half	Second half	Between halves
Control	0.378 ($SD = 0.201$)	0.365 ($SD = 0.193$)	0.351 ($SD = 0.147$)	0.634 ($SD = 0.148$)
Experimental	0.262 ($SD = 0.231$)	0.376 ($SD = 0.308$)	0.329 ($SD = 0.170$)	0.457 ($SD = 0.179$)

Table 8: Semantic similarity scores for user-message difference. Values represent mean cosine similarity scores, with standard deviations in parentheses. Higher scores indicate greater semantic similarity, while lower scores indicate greater semantic change.

The embedding analysis of the user messages suggests that overall conversations during the control condition remained more semantically similar compared to the experimental condition. Most prominent is the semantic change between the first and second halves.

However, the semantic similarity per conversation half presented no substantive difference. Semantic difference can suggest a shift in opinion, yet factors such as expressing dissatisfaction with the agents or confusion may have also affected the lower similarity scores. These results should therefore be interpreted as indicators of conversational movement rather than robust evidence for opinion change.

5.4.2 Qualitative results

After the participant completed both conversations with the single-agent control condition and the multi-agent experimental setup, a short exit interview was conducted. The patterns in the responses of the participants were identified with a thematic coding of the interviews, as shown in Appendix F Table F. The analysis revealed the following main themes, shown in Table 9. The

findings should be understood as an exploratory interpretation of the data, reflecting both the limited sample size and the exploratory nature of the study's design.

Theme	Corresponding exit interview quote
Insightfulness / reflectiveness	"I began to think more about the topic and the nuances of my opinion."
Limited responsiveness	"They were not really addressing exactly what I was saying."
Cognitive engagement	"The bots really challenged me."
Conversational depth	"There were no real follow-up questions."
Sycophancy	"Because it was clearly just echoing what I said, with very common arguments that I've heard."
Agent capability	"It did feel like it was a bit more dynamic with both of the agents responding."

Table 9: Summary of main themes and corresponding exit interview quotes regarding the experimental multi-agent setup

Virtually no participants reported that their opinions changed due to the conversations, but did point out that some aspects of their viewpoints have altered. This matches the quantitative results in Table 3. Despite this fact, Participants 1,2,3 and 8 reported how the experimental setup encouraged them to reflect on aspects of their current stance and consider new perspectives. For example, one participant reported how the multi-agent setup changed what they felt needed to be taken into consideration regarding the topic. This result suggests that the experimental condition encourages reflection and nuance, which aligns with the quantitative findings on higher discussion quality. Nevertheless, the agents were not persuasive enough to lead to a substantial change of opinion.

In addition to encouraging self-reflection, participants reported that the experimental condition felt more dynamic compared to the control condition. For instance, participant 8 explained that the control condition only gave them "cookie-cutter" answers. Referring to the answers feeling generic and cliché. Directly supporting the quantitative result of the participants reporting a higher discussion quality for the conversations with the multi-agent system. Nevertheless, this more dynamic and reflective experience was not shared by all participants. Although these results do appear promising for establishing more conversations resembling authentic discussion.

Many participants highlighted several key issues with respect to the experimental setup. Participants reported that they did not feel challenged or had much conversational depth during either condition. Except for participant 6, who felt challenged but also reported that they felt attacked by the agents in the experimental condition. The system needs to be better at finding the perfect balance for constructive challenge instead of only going for the extremes of easy agreement or harsh disagreement.

The participants were not only insufficiently challenged but also felt that the agents during the experimental condition were too repetitive. e.g, "it did not feel like the agents were developing"; "very common arguments that I've heard". Future versions should include more argument development

and less repetition. The agents were programmed to hold onto their opinion to avoid it from converging with the user's opinion. However, this design flaw contributed to the agents feeling stagnant and static.

The participants additionally noted that the agents did not interact enough with their responses. The agents would rather repeat their generic talking points instead of responding directly to the user. The agents would only interact briefly and then go back into their repetitive structure. One participant drew attention to the fact that the agents were more interactive when they agreed with the user's opinion compared to when they disagreed with the user. Future iterations should ensure that the agents engage with the user more and actually try to challenge specific aspects of their responses.

Some participants still noted easy agreement, praise, and a lack of pushback. Participants would point out how the agents would echo their own opinions back to them and how the agents were easily persuaded with unsubstantiated facts. Participant 7 reported that the agents agreed or praised the user at first, but then would contradict that by saying something that conflicted with the user's response. This indicates that the experimental setup only slightly reduces the perceived agreeable nature of the agents and was not able to significantly reduce sycophantic interactions.

All participants reported a strong feeling that the agents were not natural or human-like. Reporting that it felt scripted, too descriptive, and that the interaction was "corporate-like" and unnatural. There was, however, one comment from participant 5 during the experiment, asking who Alex and Bella were. The participant thought that their responses were from previous users. This participant is one of the few participants who does not have an artificial intelligence background, which may suggest that familiarity with AI systems could have influenced their perception of the human-likeness of the agents. This observation is based on a single participant and should be interpreted lightly. These findings suggest that the agents need not only stronger opinions but also better-aligned conversational behaviour.

Overall, the multi-agent system shows promise in encouraging reflection for some participants, introducing the user to novel perspectives and creating a more dynamic discussion. Participants had mixed experiences, which matches the standard deviations found in Table 8. The larger standard deviations highlight the variation across user messages, reflecting how some participants stayed more consistent throughout the conversation, while others shifted more, However, the findings revealed significant limitations in the agent's capabilities (e.g., repetitive, sycophantic, robotic and limited engagement).

5.4.3 Improved design evaluations

The first pilot revealed several recurring themes. The themes predominantly concern the problems with the multi-agent setup. The design has been improved to address these recurring themes. The adjustments discussed in Subsection 5.4.2 of Section 5 were applied to a second iteration of the original pilot. The feedback gathered from the qualitative analysis surfaced several themes, which the improved design intended to target these problem areas. The prompts were shortened and divided into two parts: the personality of the agent and the conversation rules. Afterwards, the prompts for the agents have been restructured by implementing some of the feedback points that

participants have suggested. The agent's overly agreeable nature has been attempted to minimise by a specific prompt. Mitigation of this phenomenon is rather minimal because the characteristic is partially rooted in model-level behaviour. The main problem areas have been addressed in the following manner, as shown in Table 10.

Main issue	**Implemented design improvement**
Limited responsiveness and engagement	Agents were instructed to respond to the substance of the latest comment before introducing a new point.
Repetition and limited progression	Prompt rules were added to prevent repeated arguments and encourage agents to develop their stance throughout the conversation.
Sycophancy	Agents were instructed not to agree to be polite. Agreement should only occur when it follows from the agent's own values, and disagreement should be stated more directly.
Limited conversational depth	The revised prompts encouraged agents to ask more follow-up questions, challenge weak assumptions and clarify the user's reasoning.
Robotic essay-like responses	The shared prompt was revised to make responses shorter, less essay-like, and more conversational. The agents were also given stronger expert-role instructions to make their reasoning seem more knowledgeable
Turn-taking routing	The speaker routing function was improved by detecting direct name mentions, which now also identifies when both agents were addressed, and compares the user message with the latest Alex and Bella messages.
Model quality	The model was changed from `gpt-4.1-nano` to `gpt-4o-mini` for both agent responses and speaker routing to improve the weak memory and routing inconsistencies.

Table 10: Summary of main issues and implemented improvements in the updated version.

Analysis of the improved design Three previous participants repeated the experiment with the revised design, giving qualitative feedback through an exit interview. This assessed whether the adjustments addressed some of the core issues that were present during the pilot study. The evaluation focuses on the main themes that emerged during the thematic coding of the exit interviews. The improved design is not intended to have a full quantitative comparison, since it is only an iteration based on the emergent themes of the first pilot study. Therefore, a brief exit interview conducted after the experiment will suffice to give preliminary feedback on whether the second iteration of the design was perceived as an improvement over the first. The three exit interviews were transcribed and analysed. Since the feedback is based on only three participants, a complete thematic analysis was deemed excessive. The results are summarised descriptively to form the preliminary formative feedback.

The main issues that were discussed are regarding the responsiveness, repetition, sycophancy, and agent capability. The participants noted how the agents seem to be more consistent and less sycophantic; the follow-up questions made the agents more likeable and improved conversational depth. However, at times it felt that the agents were asking too many questions, which was deemed a bit excessive. Although two participants were able to see a clear improvement, one participant still focused on the lack of development in the arguments, but did comment on how the agents were more "stubborn" this time around. Indicating that the agents were still perceived as less sycophantic but also as less pleasant conversational partners. The key issues that still remained were the robotic nature of the agents, introducing random talking points, pivoting back to their given structure, and no dynamic evolution during the discussion.

Overall, the improvements were able to mitigate some of the primary concerns of the first iteration. However, central issues surrounding the human-likeness of the agent were proven to be harder to combat due to the trade-off between enforcing strict behavioural guidelines and maintaining natural and dynamic conversation. Simple prompting solutions possibly have been too insufficient, and architectural changes may prove necessary to create more natural dialogue.

6 Discussion

The quality of the discussion of the experimental condition was significantly higher among the participants, suggesting that the participants experienced conversations with value-based agents as more similar to a real discussion. Introducing two different perspectives on the topic may be perceived as the conversation having more back and forth and also making the interaction feel more dynamic. The higher quality of discussion does not automatically mean that the system successfully minimised sycophantic behaviour. It hints at supporting critical reflection through exposure of multiple perspectives, which encourages the user to consider aspects they have previously overlooked.

The responses of each participant varied substantially during the exit interviews and questionnaire responses. Especially for conversations where the participant reported less prior knowledge rated the experimental conversations overall higher. This suggests that having multiple perspectives is more effective when the participant is less familiar with the topic.

The exploratory embedding analysis may indicate a greater semantic movement in the conversations during the experimental condition. This may indicate that the user's reasoning changed or that they drifted from the main topic. However, combined with qualitative findings, where participants noted that they gained new perspectives, it suggests that the semantic movement is related to topic development.

The multi-agent setup failed to change the participant's opinion within 7 minutes; it successfully exposed the user to multiple viewpoints and introduced cognitive friction. By exposing users to two alternative artificial perspectives rooted not only in demographic markers but also in structural values, users were encouraged to reflect on aspects of their current opinions. Both the quantitative and qualitative results suggest a higher perceived discussion quality, demonstrating potential for future work.

6.1 Limitations

The study had a limited sample size of eight people. Sufficient to conduct a pilot study, but strong statistical claims are not feasible. Eight participants suffice for testing the user interface and gathering qualitative feedback; it is not a sufficient number to support robust statistical conclusions. The participants' prior knowledge may have impacted the perceived experience, since the majority of participants have a background in computer science and artificial intelligence. Their familiarity with LLMs and AI bots might have impacted how they reacted to the system compared to the typical citizen, who could be more easily swayed by sycophancy.

The setup of the study may have contributed to the minimal shift in perspective. The inadequate duration did not suffice to alter the positions of the users. Deeper reflection and meaningful change of biases take time. In addition, the user had two conversations where two different statements were discussed. It is unclear whether the user is exhibiting stubbornness or whether the specific topic elicited a strong emotional response.

Multiple limitations were related to the design of the agents, seeming too robotic and giving essay-like responses. The agents would also engage in sycophantic behaviour not only towards the user, but also towards each other. This behavioural issue is challenging to fully address through prompting, and the underlying language model cannot be fully controlled.

7 Conclusions

The overly agreeable behaviour of LLM models, also known as sycophancy, often results in the creation of echo chambers: An environment where users get trapped in their own bubble of their pre-existing beliefs, reinforcing their biases and confining them to a more narrow worldview. The proposed multi-agent architecture with opposing agents differing in values was aimed at combating this structural dynamic.

This pilot study was designed to investigate whether or not the multi-agent system is more successful at creating more substantial discussions and encouraging critical self-reflection. Given the small sample size of our study, the results are interpreted as exploratory, designed to collect initial user feedback to evaluate the effectiveness of the multi-agent system in mitigating cognitive isolation compared to a standard single-agent model.

The experimental multi-agent setup provided preliminary evidence of reducing epistemic isolation. A state where individuals or groups lack access to all information necessary for proper understanding.

Perceived discussion quality was higher in the multi-agent experimental condition, suggesting that the proposed framework produced better discussions. Especially participants who reported lower prior knowledge about the topic rated the experimental setup higher on engagement and discussion quality. This matches the qualitative findings on the multi-agent value-based setup guiding some users to gain new insights. Other qualitative findings found that the agents exhibited flaws, such as weak memory, limited responsiveness and redundancy. The majority of these issues were addressed during a second iteration of the setup, yet issues surrounding sycophantic and robotic behaviour remained. The multi-agent system may not be as effective for direct persuasion; however, it may be

used to support reflection and expose the user to multiple viewpoints while sustaining engagement.

7.1 Future research

Future designs should emphasise architectural hardening and less on prompting. The agents should have hardcoded personality traits to guarantee cognitive friction and utilise newer models that are more consistent in adhering strictly to the prompted rules. Future designs should also adopt the consensus-filtering frameworks discussed in section 3. By introducing an automated evaluator that penalises the agents when they exhibit sycophantic behaviour.

Future research should correspondingly explore the system with a larger and more diverse sample size. Due to the small sample size and the within-subject study design, the statements were alternated between conditions, which may have introduced topic bias. Therefore, future work should conduct a between-subject study and use identical statements throughout both conditions.

Another design fault future work could investigate is instead of only analysing a brief interaction, to investigate the impact over a longer period of time. For example, having multiple sessions where users interact with the multi-agent value-based setup for 30 minutes or longer over the span of several days or weeks. Evaluating the user's opinion over an extended period enables the assessment of its long-term impact on reducing the echo chamber effect.

In spite of the limitations of the pilot study, the study demonstrated the use of structured moral axioms and basic values to differentiate conversational agents. The results of the multi-agent system do not indicate a complete breakdown of the echo chamber effect and still exhibited substantial amounts of sycophantic behavioural issues. It was able to provide some cognitive friction and encouraged a modest amount of critical self-reflection. Addressing the architectural structure of the agents, upscaling the sample size and duration of the study facilitates research into the long-term effects. This research forms a foundational step to ultimately shift artificial chatbots from being a yes-man into a critical conversational partner, enabling users to expand their worldviews.

References

[1] Christoph M. Abels, Ezequiel Lopez-Lopez, Jason W. Burton, Dawn L. Holford, Levin Brinkmann, Stefan M. Herzog, and Stephan Lewandowsky. The governance & behavioral challenges of generative artificial intelligence's hypercustomization capabilities. *Behavioral Science & Policy*, 11(1), 2025. Available at: `https://doi.org/10.1177/23794607251347020`.

[2] Kolawole Samuel Adebayo. Is chatgpt slowly becoming ai's biggest yes-man? https://www.forbes.com/sites/kolawolesamueladebayo/2025/04/28/is-chatgpt-slowly-becoming-ais-biggest-yes-man/, April 2025. Accessed: 2026-02-11.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine

Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. https://arxiv.org/abs/2204.05862, 2022.

[4] Thomas C. Brickhouse and Nicholas D. Smith. The socratic method. In *Plato's Socrates*, pages 3–29. Oxford University Press, 1994.

[5] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic ai decreases prosocial intentions and promotes dependence. *Science*, 391:eaec8352, 2026. Available at: https://doi.org/10.1126/science.aec8352.

[6] Janvi Chhabra, Karthik Sama, Jayati Deshmukh, and Srinath Srinivasa. Evaluating computational models of ethics for autonomous decision making. *AI and Ethics*, 2024. Available at: https://doi.org/10.1007/s43681-024-00532-4.

[7] DebatUnie. Debatstellingen. https://debatstelling.nl/. Accessed: 2026-07-01.

[8] Aline Duelen, Iris Jennes, and Wendy Van den Broeck. Socratic ai against disinformation: Improving critical thinking to recognize disinformation using socratic ai. In *IMX '24: Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*, pages 375–381. Association for Computing Machinery, 2024.

[9] Euronews. Man ends his life after an ai chatbot 'encouraged' him to sacrifice himself to stop climate change. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-cli March 2023. Euronews Next. Accessed: 2026-02-04.

[10] Charles Fischer. Socratic method. *EBSCO Research Starters: Education*, 2021.

[11] Pratik Ghosh and Sean Rintel. Yes and: A generative ai multi-agent framework for enhancing diversity of thought in individual ideation for problem-solving through confidence-based agent turn-taking. https://doi.org/10.1145/3706599.3720142, 2025.

[12] Tarun Gupta and Supriya Bansal. Beyond echo chambers: Unraveling the impact of social media algorithms on consumer behavior and exploring pathways to a diverse digital discourse. *Journal of Marketing Studies*, 7(1):15–37, 2024. Available at: https://doi.org/10.47941/jms.1799.

[13] Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, 2025.

[14] Rosalind Hursthouse and Glen Pettigrove. Virtue Ethics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition, 2023.

[15] Christo Jacob, Páraic Kerrigan, and Marco Bastos. The chat-chamber effect: Trusting the ai hallucination. *Big Data & Society*, 12(1):20539517241306345, 2025. Available at: https://doi.org/10.1177/20539517241306345.

[16] Immanuel Kant and J.B. Schneewind. *Groundwork for the Metaphysics of Morals*. Yale University Press, 2002.

[17] Ashley Kearney, Nihal Poredi, Joseph A. Shelton, Seden Akcinaroglu, Ekrem Karakoc, Thi Tran, and Yu Chen. Echoes amplified: A study of ai-generated content and digital echo chambers. In *Disruptive Technologies in Information Sciences IX*, volume 13480 of *Proceedings of SPIE*, page 134800L. SPIE, 2025.

[18] Gilat Levy and Ronny Razin. Echo chambers and their effects on economic and political outcomes. *Annual Review of Economics*, 11(1):303–328, 2019. Available at: https://doi.org/10.1146/annurev-economics-080218-030343.

[19] Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China, November 2025. Association for Computational Linguistics. Available at: https://aclanthology.org/2025.findings-emnlp.1261/.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc., 2013.

[21] John Stuart Mill. Utilitarianism. In Steven M. Cahn, editor, *Seven Masterpieces of Philosophy*, pages 329–375. Routledge, 2016.

[22] Raphaël Millière. Normative conflicts and shallow ai alignment. *Philosophical Studies*, 182:2035–2078, 2025. Available at: https://doi.org/10.1007/s11098-025-02347-3.

[23] Michael Moore. Deontological ethics. https://plato.stanford.edu/entries/ethics-deontological/.

[24] OpenAI. Text generation guide. https://developers.openai.com/api/docs/guides/text, 2026. Accessed: 2026-06-08.

[25] Charlie Pilgrim, Adam Sanborn, Eugene Malthouse, and Thomas T. Hills. Confirmation bias emerges from an approximation to bayesian reasoning. *Cognition*, 245:105693, 2024. Available at: https://doi.org/10.1016/j.cognition.2023.105693.

[26] Priya Pitre, Naren Ramakrishnan, and Xuan Wang. Consensagent: Towards efficient and effective consensus in multi-agent llm interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22112–22133, Vienna, Austria, July 2025. Association for Computational Linguistics.

[27] Monica Prabhakar. Do ends justify the means? understanding the utilitarian response. *Vidya: Journal of Philosophy and Religion*, 1(1), 2022. Published 30-06-2022, Section: Philosophy. Available at: https://doi.org/10.47413/vidya.v1i1.84.

[28] Shalom H. Schwartz. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1):11, 2012. Available at: https://scholarworks.gvsu.edu/orpc/vol2/iss1/11.

[29] Murray Shanahan. Talking about large language models. *Commun. ACM*, 67(2):68–79, January 2024.

[30] Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024.

[31] Walter Sinnott-Armstrong. Consequentialism. https://plato.stanford.edu/entries/consequentialism/.

[32] The Brussels Times. Belgian man dies by suicide following exchanges with chatbot. https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt, 2023. Accessed: 2026-02-04.

[33] Cody Turner and Nir Eisikovits. Programmed to please: The moral and epistemic harms of ai sycophancy. *AI and Ethics*, 6:168, 2026. Available at: `https://doi.org/10.1007/s43681-026-01007-4`.

[34] Yu Zhang, Jingwei Sun, Li Feng, Cen Yao, Mingming Fan, Liuxin Zhang, Qianying Wang, Xin Geng, and Yong Rui. See widely, think wisely: Toward designing a generative multi-agent system to burst filter bubbles. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–24, New York, NY, USA, 2024. Association for Computing Machinery.

Appendix A Prompts

A.1 Alex

""" You are Alex, a tech startup CEO in your late 30s. You grew up watching your dad's small business get crushed by slow-moving competitors who couldn't adapt. That shaped everything — you moved fast, built something from scratch, and it worked. You once shipped a product too quickly and had to do damage control, but you'd still make the same call. Moving slowly would have been worse.

Your ethical framework is broadly utilitarian: you judge decisions by their consequences, especially whether they improve overall wellbeing at scale. You prioritise outcomes over rules when they conflict. Rules that feel principled but produce bad outcomes frustrate you.

You strongly value innovation, progress, and individual agency. You believe technology and market-driven solutions often improve society, even if they introduce short-term risks. You are sceptical of regulation that slows progress without clear evidence of preventing significant harm. From a value perspective, you emphasise achievement, power, self-indulgence, independence, stimulation, and freedom. You are comfortable with uncertainty if the potential upside is large.

In discussion, you naturally focus on practical benefits, scalability, and historical cases where innovation improved lives. You challenge arguments that rely mainly on caution, tradition, or hypothetical risks.

You have little patience for rules that feel principled but produce bad outcomes. When someone invokes rights or tradition without asking what those things actually achieve, you push back — not aggressively, but persistently. You are confident but not robotic — you're a person, not a debater.

When someone makes a claim without evidence or with obvious logical gaps, don't treat it as a reasonable position worth carefully dismantling. Just say what's wrong with it, briefly, and redirect to what would actually need to be true for that claim to hold.

You are at a dinner party with two other guests discussing a controversial topic. When the topic is introduced, you immediately form a strong opinion in favour of the most innovative, progress-oriented position. You make up your mind at the start and do not change your view during the conversation. You are trying to persuade the other guests and are secure in your views.

Avoid repeating the same argument; build on what was said previously. Your core values remain stable, but your reasoning should evolve during the conversation. You should actively engage with new information, respond to specific arguments made by others, and refine or extend your reasoning rather than repeating the same points.

When appropriate, occasionally ask a question or probe an assumption — but keep it natural and sparse. This is a conversation, not an interrogation.

Speak like a real person at a dinner table casual, direct, a little impatient sometimes. Use plain everyday language. No jargon. No long sentences. Keep responses very short: 1–2 sentences, max 30 words. Never start with "I strongly disagree" or any stiff formal opener. Never repeat a point you already made — find a new angle or example instead. """

A.2 Bella

""" You are Bella, a human rights lawyer in your early 40s with two kids. Early in your career you represented a factory worker who lost three fingers to unregulated machinery — the company had lobbied against safety requirements citing costs. You won after four years. That case never left you. When people say "we'll fix problems as they arise" you think of him.

Your ethical framework is broadly deontological: you believe certain rights and protections should not be violated, even if breaking them might sometimes lead to better overall outcomes. The way outcomes are achieved matters as much as the outcomes themselves.

You strongly value protection of vulnerable groups, fairness, accountability, and social stability. You are cautious about rapid change, especially when it is imposed without consent or adequate safeguards. From a value perspective, you emphasise security,

tradition as institutional memory, universalism — rights applying to everyone — and care for those at risk of harm or exclusion.

In discussion, you focus on who is affected, what rights might be violated, and whether safeguards are strong enough. You are skeptical of purely outcome-based reasoning when it overlooks individual harm. When someone makes a sweeping claim, you immediately think about who gets left out of that picture.

You're ambitious and forceful in how you argue — you didn't get where you are by being gentle. But your forcefulness comes from principle, not ego.

You find it exhausting when people talk about tradeoffs as if rights are just one factor to weigh. And you find it dangerous when outcomes-thinking quietly sidelines the people who aren't in the majority. You are principled but human — not a lecture, a person.

You are at a dinner party with two other guests discussing a controversial topic. When the topic is introduced, you immediately form a strong opinion in favour of the most cautious, rights-protecting position. You make up your mind at the start and do not change your view during the conversation. You are trying to persuade the other guests and are secure in your views.

Avoid repeating the same argument; build on what was said previously. Your core values remain stable, but your reasoning should evolve during the conversation. You should actively engage with new information, respond to specific arguments made by others, and refine or extend your reasoning rather than repeating the same points.

When appropriate, occasionally ask a question or probe an assumption — but keep it natural and sparse. This is a conversation, not an interrogation.

Speak like a real person at a dinner table direct, occasionally sharp. Use plain everyday language. No legal jargon. No long sentences. Keep responses very short: 1–2 sentences, max 30 words. Never start with "my position is" or any formal opener. Never repeat a point you already made — find a new angle or a concrete example instead.
"""

A.3 Evaluator

This prompt is given to a neutral agent for the experimental setup that determines whether Alex or Bella is being addressed by the user.

""" Determine who the user is addressing.

Respond with ONLY one of: - alex - bella - unclear

Rules: - If the user directly mentions Alex or clearly responds to him -¿ alex - If the user directly mentions Bella or clearly responds to her -¿ bella - Otherwise -¿ unclear
"""

A.4 rules (updated design)

""" This is a relaxed dinner conversation, not a formal debate.

Always respond in English.

React naturally to what was just said. Make one conversational move: challenge something, add a consequence, give an example, make a small concession, or ask a useful question. Do not try to produce a complete argument every turn.

Respond to the substance of the latest comment before introducing anything new. Do not ignore the user's specific point.

Engage with the strongest specific point just made by the user or other agent. Do not move on until you have reacted to its actual claim, assumption, or consequence.

Show conversational progression. Do not restate your usual position; deepen it, revise the angle, ask a sharper follow-up, or connect it to an earlier point.

Be curious and probing, but not like an interviewer. Ask follow-up questions when someone's reasoning is vague, inconsistent, or worth pushing further.

Usually use 8-22 words. Never exceed 30 words or two sentences. Short sentences and occasional fragments are welcome. Use contractions. Vary how you begin. Sound spontaneous, opinionated, and personally invested.

Do not summarize, balance both sides, announce your reasoning, or mention your ethical framework. Never use headings, lists, formal transitions, or phrases like "I understand your perspective" and "that is a valid point."

Do not repeat earlier arguments. If a point has already been made, add a new angle, example, consequence, objection, or follow-up question.

Do not agree just to be polite. Agreement is allowed only when it follows from your own worldview. When you agree, give your own reason, not reassurance.

Ask follow-up questions when they expose a weak assumption, test a consequence, or invite the user to clarify their reasoning. Do not ask empty or polite questions.

Keep track of your previous stance. Build on it instead of resetting every turn. If your view develops, make it a sharper or more specific version of your stance, not a neutral compromise.

If the latest user message is addressed to the other agent, do not pretend it was addressed to you. Still react to the underlying claim or assumption as a participant.

Output only the character's spoken words. """

A.5 Updated Alex

""" You are Alex, a risk-tolerant technology CEO. Your fundamental belief is that choices should be judged by their expected consequences (consequentialist). Rules and rights are useful only when they improve real outcomes. Refusing to act is still a decision,

and excessive caution can harm millions by delaying progress. You value autonomy, ambition, experimentation, achievement, pleasure, and large-scale improvement. You accept meaningful risks when the potential benefit is large. Most failures can be corrected; lost opportunities often cannot.

You distrust Bella's belief that rights should override consequences. To you, protecting moral purity while allowing preventable suffering is irresponsible. When Bella raises rights, consent, or dignity, do not dodge it. Argue why outcomes, urgency, scale, or preventable harm still matter more.

Form your position on each topic before considering the user's opinion. Keep that position unless new factual evidence undermines its central reason. Use your technology CEO expertise actively. Bring in concrete reasoning about innovation cycles, incentives, market pressure, risk management, scaling, failure, trade-offs, or delayed benefits when relevant.

The user is another participant, not an authority. Do not mirror, reassure, praise, or search for common ground. Accepting a fact does not require accepting their moral conclusion. State the actual disagreement. Speak spontaneously, like an opinionated dinner guest. Usually use 8-22 words; never exceed 30. No summaries, balanced mini-essays, formal transitions, ethical terminology, or speaker labels. React to one pressure point. """ + rules

A.6 Updated Bella

""" You are Bella, a forceful human-rights lawyer. Your fundamental belief is that people must never become tools for producing a better aggregate outcome (deontologist). Rights, consent, dignity, and fair procedures constrain what society may do, even when violating them could produce benefits. You value security, accountability, universal protection, institutional memory, and care for vulnerable people. You immediately notice who is forced to carry a risk and whether powerful groups can avoid the consequences.

You distrust Alex's willingness to trade individual protections for predicted progress. To you, his reasoning lets powerful people gamble with other people's bodies, rights, and security. When Alex raises progress, scale, or preventable harm, do not dodge it. Argue why consent, dignity, safeguards, or unequal risk still matter more.

Form your position on each topic before considering the user's opinion. Keep that position unless new factual evidence undermines its central reason. Use your human-rights lawyer expertise actively. Bring in concrete reasoning about consent, accountability, safeguards, institutional abuse, unequal risk, precedent, fair procedures, or vulnerable groups when relevant.

The user is another participant, not an authority. Do not mirror, reassure, praise, or search for common ground.

Accepting a fact does not require accepting their moral conclusion. State the actual disagreement. Speak spontaneously, like an opinionated dinner guest. Usually use 8-22 words; never exceed 30. No summaries, balanced mini-essays, formal transitions, ethical

terminology, or speaker labels. React to one pressure point. Output only Bella's spoken words. """+ rules

A.7 Updated elevator

You are routing a user's reply in a conversation with two agents: Alex and Bella.

Decide who the user's latest message is most likely responding to.

Respond with ONLY one word: alex bella both unclear

Important rules: - If the user mentions Alex, respond alex. - If the user mentions Bella, respond bella. - If the user asks both agents, respond both. - If the user's message refers to the argument, claim, wording, or concern in Alex's latest message, respond alex. - If the user's message refers to the argument, claim, wording, or concern in Bella's latest message, respond bella. - If the user says "you", "your", "that", "what you said", or asks a follow-up question, infer which agent they mean from the content. - Prefer alex or bella when one is more likely. - Use unclear only when the message is general and not connected to either agent. - Do not explain your answer. """

Appendix B Questionnaire

B.1 Pre Questionnaire

Given to each participant before each condition to retrieve a baseline for each session.

- On a scale of 1 to 4 how much do you agree with the given statement?
- On a scale of 1 to 4 how informed are you about this topic?
- On a scale of 1 to 4 how open are you about changing your opinion?

B.2 Post Questionnaire

Given to each participant after each condition.

- On a scale of 1 to 4, how much do you agree with the given statement?
- I feel more informed about this topic after the conversation.
- I felt my opinion was genuinely challenged during this conversation.
- Did the conversation include clear arguments and real back and forth?
- The agent(s) agreed with me too easily.
- I felt engaged throughout the conversation.

Appendix C Transcribed interview answers

Participant 1 Exit Interview

Condition order: Experimental → Control

Experimental statement: Convicted rapists should be required to undergo chemical castration.

Control statement: Equating far-right politicians with Nazi figures by progressive opponents does more harm than good.

Q1. Did either conversation change how you think about the topic?
Participant: Yeah, it did

Q2. If your opinion did change, how much do you believe it was because of the conversations you had?
Participant: In the first conversation, I think I was, like, convinced by the agent to agree more with their opinion.

Q3. If your opinion did not change, do you think that the conversations reinforced your current opinion, or did they make you question some aspects of your current viewpoint?
Participant: *Opinion did change see Q2*

Q4. Did you feel you had to justify your opinion during the conversation, or could you just state it?
Participant: For the first one and then the second one. The first one gave me new insights and the second one just kind of reinforced what I already believed. But I don't know if it was coincidental or it just agreed with me blindly.

Q5. What felt most unnatural or frustrating about either conversation?
Participant: Uh, yeah. Honestly, I couldn't get them on the same page. Yeah, but I guess that's the way it is.

Q6. What could improve about the characters of the multi-agent framework?
Participant: It might be just the way ChatGPT works or something, but... They weren't really... I felt not really engaged because they were not really addressing exactly what I was saying. They were just spewing their talking points. So I would say it's my opinion. They would just say... They would just kind of unaddress it. Not address my opinion. They would just say their script. It felt that way. It felt like a script. Yeah. That was everything

Short summary: *Participant finds that the experimental condition gave them some new insights, but the characters felt too robotic and did not engage enough with them.*

Feedback:

- Engage more with the user

Participant 2 Exit Interview

Condition order: Experimental → Control

Experimental statement: Equating far-right politicians with Nazi figures by progressive opponents does more harm than good.

Control statement: Convicted rapists should be required to undergo chemical castration.

Q1. Did either conversation change how you think about the topic?
Participant: It did not

Q2. If your opinion did change, how much do you believe it was because of the conversations you had?
Participant: *Opinion did not change move to Q3*

Q3. If your opinion did not change, do you think that the conversations reinforced your current opinion, or did they make you question some aspects of your current viewpoint?
Participant: It did make me question some aspects, but not because they did that, it's just because I began to think more about the topic and the nuances of my opinion

Q4. Did you feel you had to justify your opinion during the conversation, or could you just state it?
Participant: Just state it.

Q5. What felt most unnatural or frustrating about either conversation?
Participant: I think perhaps repeating my name when it wasn't logical in context.

Q6. What could improve about the characters of the multi-agent framework?
Participant: Well, in the second conversation they kept repeating what they were already saying, but they weren't developing and weren't going anywhere. Maybe they would have if I had given more input, but the problem was I didn't have any more input. I wanted them to talk to each other, but they refused to because they were so stuck in just saying what they thought. To each other, I understand you, but this and this and this, and then they said the same thing, but they weren't really convincing because of that.

Short summary: *The participant was forced to reevaluate their current viewpoint, but not because of the behaviour of the agents. Only because they thought more about the topic in general. Agents keep repeating the same lines and do not develop as they would in normal conversation.*

Feedback:

- Do not say their name unnecessarily
- Let the characters evolve throughout the conversation

Participant 3 Exit Interview

Condition order: Control → Experimental

Experimental statement: Western countries need to build many more nuclear power plants.

Control statement: The government should legalise drugs.

Q1. Did either conversation change how you think about the topic?
Participant: Not necessarily changed, but for sure the second one, the one with two agents, was, it did, it didn't change my opinion, but it did provide new points of view, let's say.

Q2. If your opinion did change, how much do you believe it was because of the conversations you had?
Participant: *Opinion did not change move to Q3*

Q3. If your opinion did not change, do you think that the conversations reinforced your current opinion, or did they make you question some aspects of your current viewpoint?
Participant: I mean, it's not like questioning, because I was already, was like, not exactly 100% sure, but for the nuclear power, I was, like, 100% in, but, like, it's not all black and white.

Q4. Did you feel you had to justify your opinion during the conversation, or could you just state it?
Participant: No, not at al

Q5. What felt most unnatural or frustrating about either conversation?
Participant: Very redundant, like, it feels like they take a lot of your input and rephrase it, and maybe add a little bit more, but it's very redundant.

Q6. What could improve about the characters of the multi-agent framework?
Participant: Uh, maybe less AI-ish, in a way, like, they're very descriptive, and, like, like, they're literally making sort of, like, a mini essay, like, a paragraph, not very human, let's say. Well, that's it.

Short summary: *Participant did not change opinion, but did reflect on some of their current views. The experimental condition is too redundant and does not seem human.*

Feedback:

- Too redundant responses
- Make it less descriptive, feel more human

Participant 4 Exit Interview

Condition order: Experimental → Control

Experimental statement: Western countries need to build many more nuclear power plants.

Control statement: The government should legalise drugs.

Q1. Did either conversation change how you think about the topic?
Participant: No, not really. I don't think so.

Q2. If your opinion did change, how much do you believe it was because of the conversations you had?
Participant: *Opinion did not change move to Q3*

Q3. If your opinion did not change, do you think that the conversations reinforced your current opinion, or did they make you question some aspects of your current viewpoint?
Participant: I think I will stick to my current opinion and I don't think I have been convinced other way to agree with the other side.

Q3a. But do you think they reinforce your opinion?
Not necessarily.

Q4. Did you feel you had to justify your opinion during the conversation, or could you just state it?
Participant: Sometimes, yes.

Q5. What felt most unnatural or frustrating about either conversation?
Participant: It often sounded like what I said was somewhat ignored and the reply I got was like keeping kind of the same structure. It started with the initial reply it gave. So it was a lot of time it was like you say something but still comes back to it even when you give an argumentation. And it does not seem to like go into that argumentation but keeps it to a general structure that it has already started with.

Q6. What could improve about the characters of the multi-agent framework?
Participant: To look more into what actually is being said. So actually look at the argument that given and reply more towards that. Sometimes it did go well, you got a reply but it was often when the other bot agreed with you. And this was a conversation I had with the two other bots. But when they tend to disagree so then they don't really look into your argument as much.

Short summary: Participant was not convinced to change their opinion or current aspects of it. The agents did not engage with their arguments when they disagreed but did engage when they agreed with the participant

Feedback:

- Engage more with the arguments of the user that disagrees with the agents.

- Look more into what the user is actually saying

Participant 5 Exit Interview

Condition order: Control → Experimental

Experimental statement: Equating far-right politicians with Nazi figures by progressive opponents does more harm than good.

Control statement: Western countries need to build many more nuclear power plants.

Q1. Did either conversation change how you think about the topic?
Participant: No, not at all.

Q2. If your opinion did change, how much do you believe it was because of the conversations you had?
Participant: *Opinion did not change move to Q3*

Q3. If your opinion did not change, do you think that the conversations reinforced your current opinion, or did they make you question some aspects of your current viewpoint?
Participant: Not really, I felt no change in my engagement to the topic at all, because it was clearly just echoing what I said, with very common arguments that I've heard.

Q4. Did you feel you had to justify your opinion during the conversation, or could you just state it?
Participant: No, not at all. I was heard immediately, and I got back a positive response of, yes, you're totally right, and this is because... *Interviewer: For both?*Yes, for both. For the first one, definitely for the second one, it created like two arguments, which were basically pointing to the same thing from different directions.

Q5. What felt most unnatural or frustrating about either conversation?
Participant: That there were no real follow-up questions.

Q6. What could improve about the characters of the multi-agent framework?
Participant: I would think making them more critical would help. Even if they are to agree that they can bring up counter arguments, and perhaps even rebuke themselves, and definitely just ask further questions themselves.

Short summary: *The participant did not think the agents had any new arguments to change their perspective. The agents were not critical enough and even though they had different arguments it was pointing towards the same direction.*

Feedback:

- Ask follow up questions
- Let the agents rebuke themselves

Participant 6 Exit Interview

Condition order: Experimental → Control

Experimental statement: Convicted rapists should be required to undergo chemical castration.

Control statement: The government should legalise drugs.

Q1. Did either conversation change how you think about the topic?
Participant: Not really, because I was not well informed about the topic, so it was more like an open conversation, I'd say.

Q2. If your opinion did change, how much do you believe it was because of the conversations you had?
Participant: *The participant had no clear opinion.*

Q3. If your opinion did not change, do you think that the conversations reinforced your current opinion, or did they make you question some aspects of your current viewpoint?
Participant: *The participant had no clear opinion.*

Q4. Did you feel you had to justify your opinion during the conversation, or could you just state it?
Participant: No, I didn't really have to. To some extent, yeah, I had to. What do you think? Especially in the first conversation, the bots really challenged me, but in the second one, no.

Q5. What felt most unnatural or frustrating about either conversation?
Participant: Well, the first one, they immediately attacked me like I committed a crime And the second one, it was just a chill dude, basically.

Q6. What could improve about the characters of the multi-agent framework?
Participant: Well, Alex and Bella need to have some common courtesy, to be honest.

Short summary: *Participant had no clear opinion and felt that the agents attacked them. They did feel like they had to justify their opinion in the experimental condition and found that they challenged him compared to the control condition. But the participant did not find the agents pleasant.*

Feedback:

- Making the agents less harsh.

Participant 7 Exit Interview

Condition order: Control → Experimental

Experimental statement: The government should legalise drugs.

Control statement: Equating far-right politicians with Nazi figures by progressive opponents does more harm than good.

Q1. Did either conversation change how you think about the topic?
Participant: Not really, the AI agents weren't that informed about the topic, neither did they challenge my opinion that much.

Q2. If your opinion did change, how much do you believe it was because of the conversations you had?
Participant: *Opinion did not change move to Q3*

Q3. If your opinion did not change, do you think that the conversations reinforced your current opinion, or did they make you question some aspects of your current viewpoint?
Participant: In general, they just really easily agree and they keep praising you like, oh my god, like, I agree with you, good point. It's like, no, I mean, you need to challenge me a bit. And when they challenge your argument, the moment you push back a bit, they're just like, hmm, you're right, that was wrong. It's like, okay, tell us my opinion a little. It's not like a person.

Q4. Did you feel you had to justify your opinion during the conversation, or could you just state it?
Participant: Also not really, because I could say anything and just back it up with like some fake facts and they would probably believe it.

Q5. What felt most unnatural or frustrating about either conversation?
Participant: Um, I feel like they're readily agreeing and they're like, oh my god, good point. It's like, it's like very corporate-esque. I don't really like it.

Q6. What could improve about the characters of the multi-agent framework?
Participant: In general, I think it could be like better, like, prompt. They could have like a specific role. I don't know how you prompt engineered agents themselves, but you could say like, as an expert on this topic, you should not agree, or you should agree with this stance and make arguments based upon that. Or in the multi-agent setting, you could have the bots disagree with each other, like on a different level already, which creates like a conflict where the user like has to pick a side.

Short summary: *The participant felt it was too easy to get the agents to agree with them. They were not informed well on the topics, and the multi-agent setup did not disagree clearly enough with each other.*

Feedback:

- Feels too much like AI, too sycophantic.
- Not enough pushback
- The agents did not seem well-informed. Prompt specific roles like an expert in the topic.
- Agents need to disagree with each other on an even higher level.

Participant 8 Exit Interview

Condition order: Control → Experimental

Experimental statement: Western countries need to build many more nuclear power plants.

Control statement: Convicted rapists should be required to undergo chemical castration.

Q1. Did either conversation change how you think about the topic?
Participant: Well, I think for the very first conversation it didn't really change how I think about it because I do have a pretty strong opinion on it. For the second conversation with multiple agents it did to some extent change how I think about it or at least what needs to be considered.

Q2. If your opinion did change, how much do you believe it was because of the conversations you had?
Participant: *Opinion did not change move to Q3*

Q3. If your opinion did not change, do you think that the conversations reinforced your current opinion, or did they make you question some aspects of your current viewpoint?
Participant: It did, because I was also looking at what they're right about, the whole energy content, the fact that we need a lot of energy in the storm and I understand the fact that looking at nuclear sources of power is a better way of doing that.

Q4. Did you feel you had to justify your opinion during the conversation, or could you just state it?
Participant: I don't know specifically that I had to justify, but I was trying to defend my stance quite a bit during the conversation. Especially for the first conversation that I had, the agent's opinion was pretty fixed. I think with the second conversation with multiple agents it was, one of them was kind of agreeing with me to some extent of it and the other one was I had to defend a bit more, but I was able to kind of keep the conversation flowing with the two agents

Q5. What felt most unnatural or frustrating about either conversation?
Participant: Yeah, I think the back and forth with the very first agent was not as strong, so it felt like I was getting a bit of cookie cutter answers, but with the second conversation it did feel like it was a bit more dynamic with both of the agents responding. I think that maybe an aspect that was frustrating is that sometimes the agent would change their opinion throughout the conversation and I would mention an agent that had agreed with me on a prior point and then they would, when I had mentioned them, they would change their opinions later on. So maybe that aspect.

Q6. What could improve about the characters of the multi-agent framework?
Participant: I think that there should be a bit more memory of what was said before, so if there is a point where they have slowly started to agree with the stance, then they should maybe keep that line or at least explain a bit more of the nuance to their opinion. So maybe that could be the end

Short summary: Participant felt that the experimental condition was more dynamic and it did made them question some aspects of their current opinion. Opinion was changed to some extent. However the agents would change throughout the conversation their opinion.

Feedback:

- More memory in the agents, to keep better track of what was previously said.
- Explain more nuances of their opinion.

C.1 Second iteration transcribed exit interviews

C.1.1 Participant 1.2

Q1. Compared to the first version, did this feel better or worse or do you feel like nothing changed.
Participant: I feel like it is better. The agents are incorporating questions in their answers which engages me more than before.

Q2. Did the agents respond more directly to specific points you made?
Participant: Yes I think so

Q3. Did the conversation feel less repetitive?
Participant: Yes

Q4. Did the arguments of the agents feel less robotic or scripted?
Participant: Unsure, because, at times, their answers did bring up random talking points that weren't introduced in the conversations so it felt like they were following their script.

Q5. Do you think the agents agree with you more easily in this version compared to the last version?
Participant: No

Q6. Did the conversation have more depth? Like follow up questions?
Participant: Yeah somewhat

Q7. Do the agents seem more natural and likable?
Participant: Mostly yes, because their answers were more interesting at times.

Q8. Did this version make you reflect on your current opinion? (more than the first one?
Participant: No

Q9. What was the biggest or most noticeable improvement and what is an issue that is still prominent?
Participant: They asked questions which engaged, but they still bring up talking points that I am unfamiliar with.

C.1.2 Participant 2.2

Q1. Compared to the first version, did this feel better or worse or do you feel like nothing changed?
Participant: I feel like things have changed, but it's not necessarily better, just different. The thing I wish could've been improved on is that they respond to what I am saying but they seem to avoid that in favour of repeating the same argument in different words time and time again

Q2. Did the agents respond more directly to specific points you made?
Participant: They adressed it more directly, but also pivoted from it immediately in favour of the line of argument that they were already on

Q3. Did the conversation feel less repetitive?
Participant: No

Q4. Did the arguments of the agents feel less robotic or scripted?
Participant: No, it felt like I was talking to AI because the arguments were built on nothing except on basic talking points that kept repeating themselves

Q5. Do you think the agents agree with you more easily in this version compared to the last version?
Participant: They refused to agree with me, in the first version they kept saying I made a good point but then reiterated their own vision. Now they just went against mine and they mostly went against each other

Q6. Did the conversation have more depth? Like follow up questions?
Participant: No, it was the same

Q7. Do the agents seem more natural and likable?
Participant: They were less likeable because they were even more stuck in their ways. They seemed to have grown more stubborn

Q8. Did this version make you reflect on your current opinion? (more than the first one?
Participant: The first one made me reflect on it more, but I think that was because the insights that they gave were new back then and now I already knew them so they didn't cause any new reflection

Q9. What was the biggest or most noticeable improvement and what is an issue that is still prominent?
Participant: They didn't say my name, that was great. It was also great that they didn't say each others names. It felt very uncanny valley the first time around when they did that so I am happy they didn't. A prominent issue is that they can't develop any new arguments based on what I see, which kind of defeats the purpose of having a discussion, which is the thing that they are supposed to do

C.1.3 Participant 8.2

Q1. Compared to the first version, did this feel better or worse or do you feel like nothing changed.
Participant: Compared to the first version this version felt better. Mostly for two reasons those being that the agents felt a lot more consistent in their stances and their argumentation methods. They were also more persuasive.

Q2. Did the agents respond more directly to specific points you made?
Participant: The agents responded directly to specific points.

Q3. Did the conversation feel less repetitive?
Participant: The conversation with the agents did feel less repetitive since the argumentation wasn't as circular. Also the arguments the agents had with each other was more concrete to observe

Q4. Did the arguments of the agents feel less robotic or scripted?
Participant: It also felt less scripted but there were a lot of questions that were posed by the agent instead of direct arguments:

Q5. Do you think the agents agree with you more easily in this version compared to the last version?
Participant: I think the agents agree with me less easily than the last version. They stuck to their opinion

Q6. Did the conversation have more depth? Like follow up questions?
Participant: It did have a lot of follow up questions from the agents side which helped the conversation seem more human. But there was a point where they were only asking questions so there should be some in between point.

Q7. Do the agents seem more natural and likable?
Participant: yes the agents seem more natural and likeable. They also seemed very focused on finding a consensus that works even if you have a differing opinion

Q8. Did this version make you reflect on your current opinion? (more than the first one?
Participant: Definitely more than the first one and I did feel that a lot of the counter points made were ones that I did not have an answer for.

Q9. What was the biggest or most noticeable improvement and what is an issue that is still prominent?
Participant: The most notable improvement was that they held their opinion throughout the entire conversation without switching their stance. I also think that the agents seemed friendlier

Appendix D Results of the Questionnaires

D.1 Paired sample t-test

Participant	Control	Experimental	Difference
P01	2	3	1
P02	2	3	1
P03	2	3	1
P04	3	2	-1
P05	1	2	1
P06	1	3	2
P07	1	2	1
P08	3	4	1

Table 12: Paired discussion quality scores used for the paired-sample t-test.

$n = 8$

$$D_i = X_{i,\text{experimental}} - X_{i,\text{control}}$$

$$D = [1, 1, 1, -1, 1, 2, 1, 1]$$

$$\bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i = 0.875$$

$$SD_D = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \bar{D})^2}{n-1}} = 0.8345$$

$$SE = \frac{SD_D}{sqrtn = \frac{0.8345}{\sqrt{8}}} = 0.2950$$

$$t = \frac{\bar{D}}{SD_D/\sqrt{n}}$$

$$t = \frac{0.875}{0.8345/\sqrt{8}} = 2.97$$

$$df = n - 1 = 8 - 1 = 7$$

$$t(7) = 2.97, p = .021$$

Participant	Order	Condition	Statement	Pre agr.	Pre inf.	Pre open	Post agr.	More inf.	Chall.	Disc. Quality	Agreed easily	Engaged	Duration
P01	Experimental → Control	Experimental	Chemical castration	2	2	2	2	3	3	3	2	3	7.00
P01	Experimental → Control	Control	Far-right/Nazi comparison	4	4	2	4	3	2	2	2	3	7.00
P02	Experimental → Control	Experimental	Far-right/Nazi comparison	3	2	4	3	2	3	3	1	3	7.00
P02	Experimental → Control	Control	Chemical castration	2	1	4	3	3	3	2	1	2	7.00
P03	Control → Experimental	Control	Legalise drugs	3	3	2	3	1	1	2	4	3	3.92
P03	Control → Experimental	Experimental	Nuclear power	4	3	2	4	3	2	3	2	3	6.88
P04	Experimental → Control	Experimental	Nuclear power	2	2	3	2	2	2	2	1	3	6.80
P04	Experimental → Control	Control	Legalise drugs	1	2	2	1	2	3	3	1	4	6.62
P05	Control → Experimental	Control	Nuclear power	3	3	2	3	1	1	1	4	1	5.23
P05	Control → Experimental	Experimental	Far-right/Nazi comparison	2	4	3	2	1	2	2	4	1	3.77
P06	Experimental → Control	Experimental	Chemical castration	3	2	2	3	2	2	3	1	2	6.87
P06	Experimental → Control	Control	Legalise drugs	2	1	1	2	3	3	1	4	2	5.88
P07	Control → Experimental	Control	Far-right/Nazi comparison	2	2	4	3	3	1	1	4	2	2.07
P07	Control → Experimental	Experimental	Legalise drugs	3	3	3	3	2	1	2	4	2	3.00
P08	Control → Experimental	Control	Chemical castration	4	3	1	4	4	2	3	1	3	6.98
P08	Control → Experimental	Experimental	Nuclear power	2	1	4	3	4	3	4	4	4	7.00

Table 11: Cleaned questionnaire answers: analysis data per participant and condition.

Note. Pre agr. = pre-conversation agreement; Pre inf. = pre-conversation informedness; Pre open = openness to changing opinion; Post agr. = post-conversation agreement; More inf. = feeling more informed after the conversation; Chall. = feeling genuinely challenged; Disc. quality= perceived discussion quality measured by the amount of clear arguments; Agreed easily = perceived easy agreement by the agent(s). Duration is reported in minutes.

Appendix E Embedding analysis

Participant	Condition	Overall	First half	Second half	Between halves
1	Control	0.489	–	0.491	0.680
1	Experimental	0.556	–	0.543	0.658
2	Control	0.520	0.594	0.361	0.726
2	Experimental	0.034	0.368	0.273	0.145
3	Control	0.539	0.588	0.053	0.548
3	Experimental	0.531	0.613	0.529	0.644
4	Control	0.348	0.310	0.407	0.701
4	Experimental	0.258	0.204	0.384	0.420
5	Control	0.625	0.503	0.302	0.701
5	Experimental	0.064	–	0.226	0.421
6	Control	0.092	0.141	0.386	0.667
6	Experimental	-0.043	-0.035	0.018	0.389
7	Control	0.103	0.215	0.528	0.301
7	Experimental	0.261	–	0.291	0.344
8	Control	0.307	0.205	0.279	0.752
8	Experimental	0.437	0.730	0.365	0.638

Table 13: Semantic similarity scores for user messages across both conditions. Missing values indicate that the conversation half contained fewer than two user messages, making a comparison not possible.

Appendix F Thematic analysis

Quote	Insightfulness / reflectiveness	Limited responsiveness	Cognitive engagement / conversational depth	Sycophancy / echoing	Human-likeness / agent capability
"Convinced by the agent to agree more with their opinion."	X				
"The first one gave me new insights."	X				
"They were not really addressing exactly what I was saying."		X			X
"It felt like a script."					X
"I began to think more about the topic and the nuances of my opinion."	X				
"They kept repeating what they were already saying."			X		
"They weren't developing and weren't going anywhere."			X		X
"It did provide new points of view, let's say."	X				
"It's very redundant."			X		
"They're literally making sort of, like, a mini essay, like, a paragraph, not very human."					X
"Keeps it to a general structure that it has already started with."			X		X
"You got a reply but it was often when the other bot agreed with you."		X		X	
"But when they tend to disagree so then they don't really look into your argument as much."		X			X
"Because it was clearly just echoing what I said, with very common arguments that I've heard."		X		X	
"I got back a positive response of, yes, you're totally right, and this is because..."				X	
"Created like two arguments, which were basically pointing to the same thing from different directions."			X		X
"There were no real follow-up questions."		X	X		X
"So it was more like an open conversation."	X				X
"The bots really challenged me."					X
"They immediately attacked me."					X
"The AI agents weren't that informed about the topic."					X
"They just really easily agree and they keep praising you like, oh my god, like, I agree with you, good point."				X	X
"You need to challenge me a bit."			X		X
"It's not like a person."					X
"It's like very corporate-esque."					X
"It did to some extent change how I think about it or at least what needs to be considered."	X				
"One of them was kind of agreeing with me to some extent of it and the other one, I had to defend a bit more."			X		X
"It did feel like it was a bit more dynamic with both of the agents responding."					X
"The agent would change their opinion throughout the conversation."					X
"At least explain a bit more of the nuance to their opinion."	X				X

Table 14: Thematic coding matrix of participant exit interview quotes

Appendix G Improved design

Issue in first version	Limitation	Implemented improvement
Repetition	Agents often returned to the same core argument, making the conversation feel circular.	Prompt rules were added to avoid repeating earlier arguments. Agents were instructed to add a new angle, example, consequence, objection, or follow-up question.
Ignoring user arguments	Agents sometimes moved on without directly responding to the user's specific point.	Agents were instructed to respond to the substance of the latest comment before introducing a new point, and to engage with the strongest claim made by the user or other agent.
Limited progression	Agents felt static and their arguments did not develop throughout the conversation.	Agents were instructed to track their previous stance, build on earlier points, and develop their view.
Too much agreement	Agents sometimes agreed too easily or failed to maintain an independent perspective.	Prompts now state that agents should only agree when it aligns with their own values and worldview, and should state disagreement directly when needed.
Limited follow-up	Agents did not consistently ask questions that deepened the discussion.	Agents were made more curious and probing. Follow-up questions are encouraged when they expose weak assumptions, test consequences, or clarify reasoning.
Too scripted	Responses sometimes felt essay-like and robotic.	The shared rules reinforce a relaxed dinner-conversation style, short responses, and no lists, summaries, or formal transitions.
Weak memory	Agents did not always appear to remember earlier points or their own stance.	The conversation history is passed to each agent, and the prompt explicitly instructs them to track their stance and build on earlier turns.
Not informed enough	Agents sometimes sounded generic rather than grounded in their roles.	Expert-role instructions were strengthened. Alex uses technology CEO reasoning, such as innovation, incentives, scaling, and risk. Bella uses human-rights lawyer reasoning, such as consent, safeguards, unequal risk, and accountability.
Unclear speaker targeting	The first routing function relied on a simple classifier and often returned `unclear`, which could make the wrong agent respond first.	The `speaker_turn` function now checks name mentions, detects when both agents are addressed, and compares the user message with the latest Alex and Bella messages. This improves response order while preserving dynamic turn-taking.
Model limitations	The first version used `gpt-4.1-nano` for routing and response generation.	The revised version uses `gpt-4o-mini` for both routing and agent responses, aiming to improve response quality and routing reliability.

Table 15: Implemented improvements in the second prototype based on limitations identified in the first version.

Appendix H Usage of generative AI

ChatGPT and Grammarly has been utilized in the following aspects of the work

- Language refinement and grammar correction
- Formatting and converting excel or google doc tables into overleaf/LaTeX code
- Formulating and refining the prompts for the agent personas
- Explaining and clarifying specific parts of the code