Multi-Agent Based Intellectual Humility Intervention

Dhanush Bhargav Guruksha Gurnani
George Washington University

***Abstract*—Intellectual humility (IH), the recognition of the limits of one's knowledge and openness to new perspectives, is a crucial trait for constructive discourse. With increasing levels of polarization between individuals, communities, regions, and demographic groups, IH may be more important than ever. This project explores a novel web-based IH intervention that is rooted in established psychological theory and leverages artificial intelligence for scalability and interaction. The intervention protocol is designed to make users consider relevant pieces information about an issue and think about their implications. Multiple large language model (LLM) agents are used to guide each step of the intervention. Post-test IH was significantly higher than IH before the intervention. This work contributes to the young, rapidly-growing fields of intellectual humility and AI-driven interventions, and demonstrates the potential for multi-agent systems to promote cognitive and behavioral change.**

I. INTRODUCTION

The advent of internet and especially social media applications has made information more accessible and provided a platform for people to disseminate their views. However, weaponization of information has lead to increased polarization between ideologies, individuals, and communities. This process appears to exacerbate people's tendency to focus on and process content that confirms their initial perceptions and opinions, leading to attitudinal ”echo chambers”. As a result, open discussion and debate often does not foster broader perspective-taking, but rather increases entrenchment, and in turn, polarization.

Intellectual humility (IH) is the recognition of the limits of one's knowledge and openness to new perspectives, and is essential for good judgment and constructive engagement. The intervention protocol explored in this project is designed to provide training in techniques to improve IH by enhancing the awareness and consideration of information relevant to an issue, and generating more-informed responses. This intervention aims to reduce confirmation bias by having participants generate arguments from different sides of various issues, and consider the implications of these arguments for each issue.

We utilized large language models (LLMs) to collect, validate, and present arguments as well as prompt participants to consider implications of these arguments. Using LLM agents obviates the need for human mediation and enables greater scalability of the process. This web-based intervention also enables participants to experience and consider others' perspectives while remaining anonymous. The application interface was developed to reflect this goal, offering a structured and intuitive user experience that translates the theoretical steps of the intervention into a smooth, interactive process.

II. RELATED WORK

Prior research has established intellectual humility as a critical meta cognitive capacity for recognizing the limits of one's knowledge and fallibility, with implications spanning leadership, education, and social cohesion. Porter et al. (1) offers one of the most comprehensive syntheses, arguing that intellectual humility is distinct from related constructs like open-mindedness or modesty because it specifically concerns epistemic limitations. Unlike prior work that often conceptualized humility either as a stable trait or as a fluctuating state, Porter et al. (1) advocates for a dynamic, context-sensitive view that incorporates both situational variability and cultural influences. However while Porter et al's. (1) review extensively catalogs predictors and consequences of intellectual humility, it remains primarily descriptive. Their emphasis lies in mapping the landscape rather than designing mechanisms to foster humility in applied settings. In contrast, our work moves beyond characterization to active intervention: we propose a system designed to operationalize intellectual humility in real-time interactions, adapting to users' conversational cues. Where Porter et al. underscores the challenges posed by personal threat, group dynamics, and cultural contexts in sustaining humility, our approach explicitly addresses these barriers by integrating adaptive dialogue strategies that can reinforce epistemic openness even under conditions of perceived threat or disagreement. While prior work provides foundational insights into what intellectual humility is and why it matters, our project builds upon these insights to explore *how* it can be practically supported and sustained through technological mediation.

III. SOLUTION AND METHODOLOGY

A. IH INTERVENTION PROTOCOL

1) Participants were presented with a central question which can be answered yes/no. The question generally proposed an action or policy decision (e.g., ”Should the electoral college be abolished?”).
2) Participants were then asked to choose a stance 'yes' or 'no', followed by how strongly they felt about their stance on a scale of 1 (not at all strongly) to 10 (extremely strongly).
3) Participants were then asked to consider both sides of the question, and provide as many arguments as they could—with a minimum of one argument for each side.
4) Participants were then exposed to a larger collection of arguments sourced from other participants, after which they were asked to consider categories into which these arguments could be placed. This step was designed to expose participants to different perspectives.

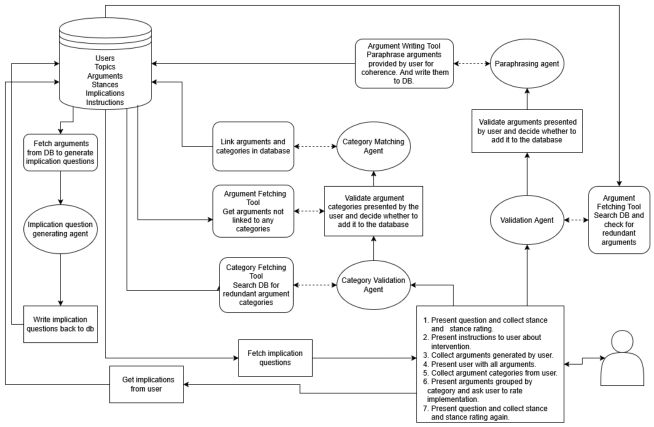

Users
Topics
Arguments
Stances
Implications
Instructions
Argument Writing Tool
Paraphrase arguments provided by user for coherence. And write them to DB.
Paraphrasing agent
Validate arguments presented by user and decide whether to add it to the database
Fetch arguments from DB to generate implication questions
Link arguments and categories in database
Category Matching Agent
Argument Fetching Tool
Get arguments not linked to any categories
Validate argument categories presented by the user and decide whether to add it to the database
Validation Agent
Argument Fetching Tool
Search DB and check for redundant arguments
Implication question generating agent
Category Fetching Tool
Search DB for redundant argument categories
Category Validation Agent
Write implication questions back to db
Fetch implication questions
Get implications from user
1. Present question and collect stance and stance rating.
2. Present instructions to user about intervention.
3. Collect arguments generated by user.
4. Present user with all arguments.
5. Collect argument categories from user.
6. Present arguments grouped by category and ask user to rate implementation.
7. Present question and collect stance and stance rating again.

Fig. 1: Application architecture showcasing database, AI agents, tools, and user interface

5) Next, participants were asked to rate the overall implications of each argument on society as positive, neutral, and negative. Participants were also asked to assess whether—all things considered—the implications of the argument supported the position represented by the question. This step was designed to improve participants' understanding of the extent to which the answer to the central yes/no question makes the outcome represented by each argument better, worse, or neither.
6) Finally, participants were presented with the central question again, and again asked to indicate their stance and how strongly they felt about it.

B. WEB APPLICATION

We developed a web application to provide the intervention protocol to participants and deployed it to the cloud. The application is built around a network of AI-agents, each equipped with tools and instructions to perform specific tasks. The backend is implemented on Python using the CrewAI library for the multi-agent network and hosted in a Flask server.

1) BACKEND ARCHITECTURE: Figure 1 shows the backend architecture of the web application including the database, AI agents, tools and tasks performed by each agent. The key components of the backend are:

1) **Database**: the application used an SQL database with tables to store the users, central questions, arguments, argument categories, implications, along with other data required for the application.
2) The **Argument Ingestion Crew** consisted of the **Validation Agent** and the **Paraphrasing Agent**. The former was used to validate arguments input by the user for redundancy and relevance while the latter was used to paraphrase the arguments for grammatical correctness and coherence (in case the user's input requires it).
3) The **Category Ingestion Crew** consisted of the **Category Validation Agent** and **Category Matching Agent**. The former was responsible for validating argument categories provided by the user for redundancy and relevance, while the latter was used to match the arguments to the most appropriate argument categories and to write theese links to the database.
4) The **Implication Question Generating Agent** was used to generate implication questions based on each argument and the central topic, which were then saved to the database, and presented to participants.
5) In addition to the above components, the backend consisted of application programming interfaces (APIs) for logging in, submitting stances, rating implications, and submitting answers to questionnaires.

2) USER INTERFACE: The user interface (UI) was developed to operationalize the multistage intervention protocol described in III-A. Built using React.js and styled with Material UI, the interface was designed to be responsive, accessible, and easy to navigate, allowing participants to engage meaningfully with each step of the intervention.

After logging in, participants began by completing a series of pre-assessment questionnaires measuring intellectual humility and social desirability. The IH questionnaire consists of 8 questions from (2), and the social desirability questionnaire consists of 11 questions taken from (3).

1) Participants were then introduced to a central policy question, asked to indicate their position (yes/no), and rated how strongly they felt about their stance on a scale of 1 (not at all strongly) to 10 (extremely strongly).
2) Participants were then informed about the purpose of the exercise (to help them make more-informed decisions), admonished to try—to the best of their ability—to put aside their opinions about the issue for the duration of the exercise. They were then instructed how to navigate each step of the exercise.
3) Participants were then instructed to generate as many arguments as they could for both sides of the question, with a minimum of one argument for each side.
4) These responses were then processed by large language model (LLM) agents that grouped the arguments into thematic categories, while also giving participants the opportunity to suggest additional categories.
5) Once categorized, participants assessed each argument's impact (positive, neutral, or negative) and evaluated its likelihood if the proposed action were implemented (more likely, less likely, or neutral).
6) Finally, the central question was again presented to the participants, who were again asked to indicate their stance and the strength of their position. This enabled a test of the intervention's impact on participants' positions vis-à-vis the central question.
7) Participants then repeated steps 2 through 7 for two more central policy questions.
8) At the end of the intervention, participants completed the IH assessment a second time to assess the intervention's impact on participants' intellectual humility.

3) CONTINUOUS DEVELOPMENT: The web application was continuously improved and development was undertaken to ensure it met usability standards while also implementing the IH protocol accurately. Feedback from stakeholders was sought at every major stage of development to ensure it aligned with their needs and expectations.

The backend went through multiple rounds of development to ensure speed and accuracy, while preventing hallucinations on part of the LLM agents. The UI also underwent many weeks of development to ensure it communicated correctly with the backend, was usable, and easy for users to navigate.

C. DATA COLLECTION

The three central questions selected for this experiment are shown in table I. These topics were selected due to their relevance in current times and likelihood of generating diverse perspectives and arguments on both sides (yes/no) of the issue.

Central Questions	
S.No.	Question
1	Should nuclear power be considered a key part of green energy solutions?
2	Should universal basic income (UBI) be adopted?
3	Should teachers be allowed to carry firearms in schools?

TABLE I: List of central questions presented to participants.

1) INITIAL DATA POPULATION: To populate the database before opening it up to human participants, we asked AI to generate arguments and argument categories for each of the three questions. We then fed the arguments into the application by following the steps outlined in III-B2.

2) HUMAN PARTICIPATION: Junior and senior university students were selected to participate in this intervention.

IV. RESULTS AND DISCUSSION

The data recorded for 48 participants was analyzed to measure the impact of the IH intervention protocol delivered through the web application. The analyses are reported below.

A. CHANGE IN IH SCORE

Because each question in the IH questionnaire can be answered on a scale of 1(strongly disagree) to 5(strongly agree), the average score over 8 questions was calculated for each participant to assign a single IH score. We then calculate the net change in IH score from the pre-assessment to post-assessment (Δ_{IH}) and performed a t-test on the difference from zero of the mean of net changes ($\mu_{\Delta_{\text{IH}}}$) to test for statistical significance.

One Sample T-test hypotheses	
Hypothesis	Statement
Null (H_0)	Mean of net changes in IH score ($\mu_{\Delta_{\text{IH}}}$) is not significantly different from zero.
Alternate (H_a)	Mean of net changes in IH score ($\mu_{\Delta_{\text{IH}}}$) is significantly different from zero.

TABLE II: Null and Alternate Hypotheses for the one-sample t-test on $\mu_{\Delta_{\text{IH}}}$.

The hypotheses used to perform the one-sample t-test is shown in table II and the results of the test are tabulated in table III.

The calculated mean of net changes in IH was 0.191 ($p <$ .0001) which is statistically significant with an alpha level of 0.01, two-tailed test. Hence, the null hypothesis stated in table II is rejected. The positive value of $\mu_{\Delta_{\text{IH}}}$ further indicates that the intervention, on average, led to an increase in IH scores for the participants.

One Sample T-test results	
Statistic	Value
$\mu_{\Delta_{\text{IH}}}$	0.191
t-statistic	4.557
p-value	$<$ .0001

TABLE III: Results for the one-sample t-test on $\mu_{\Delta_{\text{IH}}}$.

B. STANCE CHANGES REGARDING CENTRAL QUESTIONS

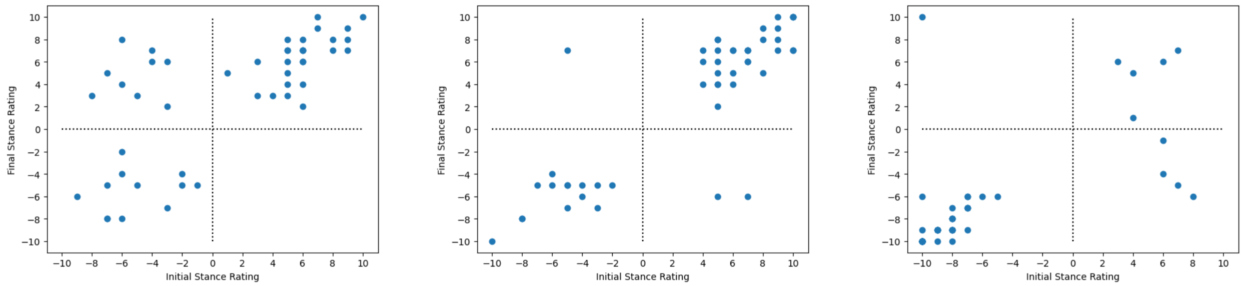
To assess stance changes regarding the three central questions, strength ratings (1-10) for all ”No” stances were given a negative valence, creating a scale of possible positions from -10 (maximally opposed) to $+10$ (maximally supportive). Scatter plots of initial and final stance rating towards the questions in table I are shown in Figure 2. The vertical and horizontal dotted lines divide the graph into 4 quadrants. The points right of the vertical dotted line represent initial ”Yes” stances and points to the left represent initial ”No” stances. The points above the horizontal dotted line represent final ”Yes” stances, and points below it represent final ”No” stances. The top right and bottom left quadrants represent participants whose stances remained the same, and the other two quadrants represent participants whose stances changed.

As shown in Figure 2a about half of the participants initially opposed to making nuclear power a key energy policy changed their stances after considering the relevant arguments. However, all of those initially in favor of nuclear power remained supportive. About twice as many participants supported as opposed universal income, with only three participants changing their stances (see Figure 2b). A majority of participants expressed strong opposition—both before and after the intervention—to arming teachers, only one of whom changed their stance, although this change was from maximally negative to maximally positive (see Figure 2c). About half of participants initially in favor of arming teachers ended up opposing this proposition.

C. IMPACT OF SOCIAL DESIRABILITY

The Social Desirability Scale is designed to assess the importance to participants of being seen by others in a positive light (3). It consists of 11 statements (answerable as ”yes”(1) or ”no”(0)) that reflect either reasonable critical (”I like to gossip at times”) or unreasonably positive (”I have never intensely disliked someone”) statements about one’s self. Responses to the critical items were reverse-coded, and these values were then combined to create a social desirability score between 0 and 11 for each participant, with higher scores reflecting more social desirability.

Since the central topics are designed such that one stance may be perceived as more socially desirable as the other, we need to study the impact of social desirability scores on initial and final stances for each central question.

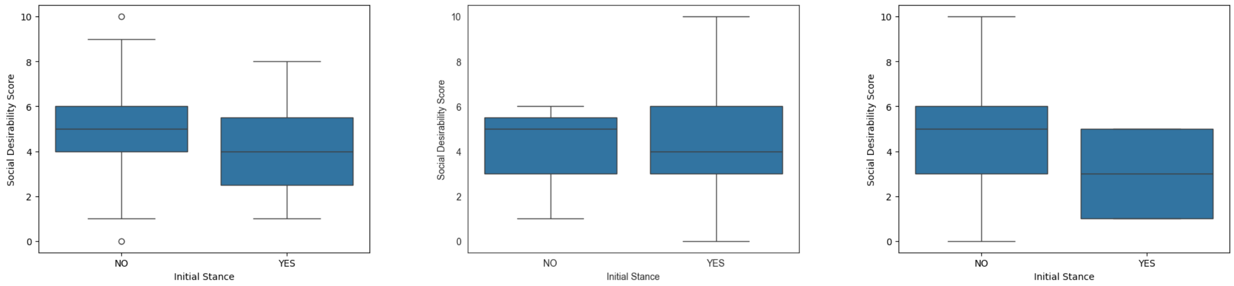


(a) Stance changes for question 1.

(b) Stance changes for question 2.

(c) Stance changes for question 3.

Fig. 2: Stance changes for each of the three central questions.



(a) Boxplot for question 1.

(b) Boxplot for question 2.

(c) Boxplot for question 3.

Fig. 3: Box plots of social desirability scores for initial 'yes' and 'no' stances for each central question.

Figure 3 shows the box-plots of social desirability scores for initial "yes" and "no" responses to all three central topics. We see that for all three central questions, participants who answered "yes" had a lower median social desirability score compared to those who answered "no".

Two Sample T-test results	
Topic No.	p-value
1	0.304
2	0.880
3	0.026

TABLE IV: Results for the two-sample t-test, comparing means of social desirability scores for initial "yes" and "no" stances for each central question.

Table IV shows the p-values for the two-sample t-tests that check if the mean social desirability scores are significantly different between participants who took the 'yes' and 'no' stances for each of the three central topics. The p-values for all three tests are higher than the significance level of 0.01 which indicates that there's no significant difference between mean social desirability scores.

Figure 4 shows the box-plots of social desirability scores for final "yes" and "no" responses to all three central topics. We see that for all three central questions, participants who answered "yes" had a lower median social desirability score compared to those who answered "no". However, we perform two-sample t-tests to check if the differences in means are statistically significant.

Table V shows the p-values for the two-sample t-tests that check if the mean social desirability scores are significantly different between participants who took the 'yes' and 'no' stances for each of the three central topics. The p-values for all three tests are higher than the significance level of 0.01 which indicates that there's no significant difference between mean social desirability scores.

Two Sample T-test results	
Topic No.	p-value
1	0.718
2	0.255
3	0.225

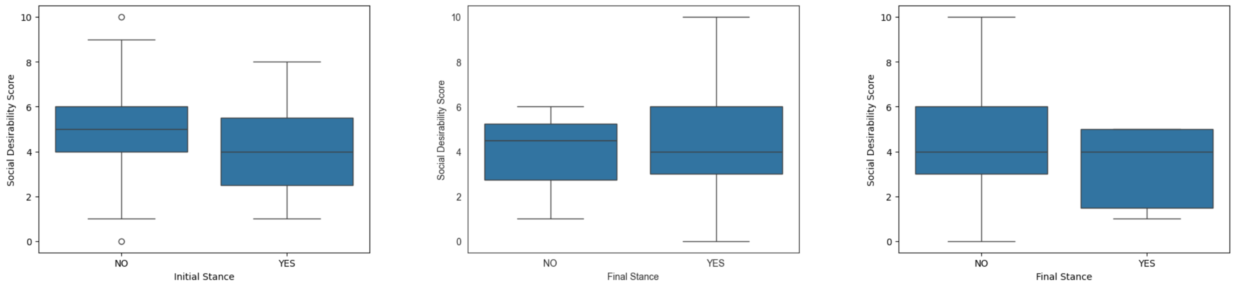
TABLE V: Results for the two-sample t-test, comparing means of social desirability scores for final "yes" and "no" stances for each central question.

From the two-sample t-tests, we can observe that the social desirability scores of participants does not have a significant impact on their selected stances to the central topics.

V. Conclusion

This intervention resulted in a significant increase in intellectual humility scores. Because this was a pilot study to check the feasibility and effectiveness of an AI-agent based approach to increase IH, we can conclude that this technique shows promise and warrants further exploration. A/B testing to compare this protocol versus naive methods, and subsequent improvements to the protocol and web application may provide more concrete results about the effectiveness of the approach discussed in this project.

Looking at the initial and final stance and stance ratings, we observe that participants did not typically change stances. However, the aim of this project is not to influence participants' stances on the central topics, but rather to increase the number



(a) Boxplot for question 1.

(b) Boxplot for question 2.

(c) Boxplot for question 3.

Fig. 4: Box plots of social desirability scores for final 'yes' and 'no' stances for each central question.

and implications of arguments they consider then making their judgments. The significant increase in IH suggests that the intervention may well have succeeded in this regard, and that this effect was not due to participants' desire to be viewed in a positive light by others.

References

[1] T. Porter, A. Elnakouri, E. A. Meyers, T. Shibayama, E. Jayawickreme, and I. Grossmann, "Predictors and consequences of intellectual humility," *Nature Reviews Psychology*, vol. 1, no. 9, pp. 524–536, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9244574/

[2] M. R. Leary, K. J. Diebels, E. K. Davisson, K. P. Jongman-Sereno, J. C. Isherwood, K. T. Raimi, S. A. Deffler, and R. H. Hoyle, "Cognitive and interpersonal features of intellectual humility," *Personality and Social Psychology Bulletin*, vol. 43, no. 6, pp. 793–813, 2017, pMID: 28903672. [Online]. Available: https://doi.org/10.1177/0146167217697695

[3] W. Reynolds, "Development of reliable and valid short forms of the marlowe-crowne social desirability scale," *Journal of Clinical Psychology*, vol. 38, pp. 119–125, 01 1982.