Universiteit Leiden

"Why Are You Arguing With Me?": The Effect of Linguistic (Mis)Alignment on Trust and Persuasion in a Human-AI Debate

Hazel Buckley McMahon (s3797457)

Supervisors:
Rob Saunders & Max Van Duijn

Bachelor Thesis– Data Science and Artificial Intelligence

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

July 8, 2026

Abstract

Following their introduction, Artificial Intelligence (AI) chatbots have been increasingly integrated into many individuals' everyday lives. People rely on AI chatbots for advice-seeking across many contexts, but to what extent do they trust the advice they receive, and could they be persuaded into changing their minds? In human interaction, trust is often mediated by perceived similarity. Namely, people tend to trust others who speak and behave as they do. The aim was to test whether this remains true in human-chatbot contexts through the manipulation of language style in a debate context. Participants engaged in short debates with two agents: one agent used a communication style that was linguistically similar to that of the participant, and the other agent used a communication style that was linguistically dissimilar. Persuasion was measured by a shift in opinion on a 0–100 scale, measured before and after each debate. Results showed that participants were, on average, more convinced by linguistically misaligned agents, but reported trusting the aligned agent more, although differences were not significant. Notably, misaligned agents were found to have a polarizing effect on trust and persuasion, with some participants being extremely influenced and others being completely dissuaded.

Contents

1 Introduction

Generative conversational Artificial Intelligence (AI) systems, such as OpenAI's ChatGPT, have received increasing attention since their introduction, largely due to their impressive conversational abilities and aptitude for assisting with everyday tasks. Skjuve et al. (2024) identified six primary motivations behind users' enthusiastic use of Generative AI systems: productivity, novelty, creative work, learning and development, entertainment, and social interaction and support. Some users specifically identified using AI for advice-seeking, stating that it is a convenient way of obtaining informed, expert-like advice. Many, if not all, of the aforementioned uses rely on a certain level of trust in the system. After all, what is the point of requesting and receiving advice from a source that cannot be trusted?

Many factors have already been identified as key modulators of human-AI trust levels. Much of this research has focused on the adoption of humanlike vs. non-humanlike features, with anthropomorphism, or the ability to attribute human traits to non-human entities, emerging as an important factor mediating trust in artificial systems, even in unexpected contexts such as self-driving cars (Waytz et al., 2014). This research has also been extended specifically to voices in the context of social robotics, with human voices being rated as both more likable and more trustworthy than their synthetic counterparts (Seaborn et al., 2021). While this research is valuable for studying human-agent interaction as a whole and for the eventual introduction of embodied artificial agents into society, it is irrelevant in the context of human-chatbot trust, which is arguably a much more commonly encountered situation in everyday life.

In order to identify additional modulators of trust in human-chatbot interactions, human-human interactions can be looked at for inspiration. Nass et al. (1994) showed that a human-computer relationship is fundamentally social and, consequently, humans apply social rules to interactions with machines despite stating that such attributions are inappropriate. For example, participants stated that politeness norms do not apply to computers; however, their behavior in the study suggested otherwise. In light of these findings, it is reasonable to hypothesize that the same factors influencing trust in human-human interactions may also apply to human-computer interactions. One of these factors is assumed similarity, or the idea that humans have a tendency to project their own characteristics and values onto those whom they perceive to be most similar to them (Thielmann et al., 2023). The most obvious way of applying this theory to a human-chatbot context is through modification of language style. A study conducted specifically on communication style in humans found that roommates who are more similar in communication style showed more affinity for each other, which can be viewed as a precursor to trust (Martin & Anderson, 1995). The tendency for people to mimic the language usage of those they are conversing with is termed linguistic alignment.

The goal of this research is to investigate whether or not the effects of linguistic alignment and misalignment in human-human contexts transfer to human-computer contexts. That being said, the focus of human research is often positive affect and rapport. This study focuses on trust and persuasion, as these outcomes are slightly more relevant in a human-computer context. Persuasion serves to obtain a more practical and perhaps stronger measure of trust. Much of the research on trust in human-computer interaction has focused on perceived trust, measured using the Trust In Automation Scale Survey (TIAS). This is a reliable measure, which will also be used in this study; however, level of influence and persuasion will also be considered as an implicit measure of practical trust.

Previous studies have shown that mimicking participants' body language and personality via voice style and volume has led to more positive ratings of, and greater persuasion by, the system. This study aims to determine whether these findings prevail in the context of language specifically. The limited research that has been conducted on linguistic alignment in human–computer interaction suggests effects similar to those found in human-human interaction. There has been little focus on the effects of misalignment in human–computer interaction. This makes sense, as misalignment is rarely a goal when designing systems that are meant to interact with humans. Regardless, the effects could be significant and should be studied.

Finally, linguistic alignment and misalignment of humans to computers has never been studied in a debate context; however, it is possibly an interesting one. People usually base their opinions on their values and other aspects of themselves and their experiences. Persuasion in this context may rely more on how similar people feel to the system, and whether they can see themselves coming to an agreement with it, rather than purely measuring perceived capability. In some ways, it may measure the potential for aligned perspectives. Measured influence and persuasion in this context further strengthen the implicit trust measure. In this experiment, participants engaged in six short debates with artificial agents as debate partners. In three debates, participants interacted with an agent that was attempting to align with their linguistic style, and in the other three, an agent that attempted to misalign with their linguistic style. Participants were asked to rate their agreement with the debate prompt on a 1-100 scale both before and after the debate, indicating the degree of persuasion. Trust was measured by the TIAS questionnaire filled out at the conclusion of the experiment.

2 Research Questions

This study aims to address the following questions:

- **RQ1:** Does linguistic (mis)alignment of a chatbot to human participants affect persuasion in a debate context?
- **RQ2:** Does linguistic (mis)alignment of a chatbot to human participants affect trust in a debate context?
- **RQ3:** Is the degree to which participants are influenced by the AI debate partners arguments consciously reported?

Based on prior research on linguistic alignment in human-human and human-agent interaction, the following hypotheses were formulated:

- **H1:** Participants will be more persuaded by the linguistically aligned agent than by the misaligned agent.
- **H2:** Participants will report higher levels of trust toward the linguistically aligned agent than towards the misaligned agent.
- **H3:** Participants will consciously report being more persuaded by the agent that objectively persuaded them more.

3 Background

3.1 Linguistic Alignment

There is strong evidence suggesting that when two people engage in dialogue, they are inclined to converge on common ways of speaking (Branigan et al., 2010). That is, they align their grammatical choices, or more commonly, their word choices. This general concept is referred to as linguistic alignment. The alignment of word choices specifically between two interlocutors is termed lexical entrainment and is often used in conjunction with, or in place of, linguistic alignment. Alignment between humans in conversation often happens unconsciously but can also be a conscious choice mediated by beliefs about fellow interlocutors. An example of conscious lexical entrainment driven by a desire for communicative success is the tendency for native speakers to adapt their lexical behavior to be less complex and adopt their conversation partners' preferred terms when they are interacting with a non-native speaker of the same language (Bortfeld & Brennan, 1997).

There are multiple levels of linguistic alignment, such as lexical, syntactic, and semantic alignment. While lexical and syntactic alignment refer to features of language such as grammar and word choice, semantic alignment is more abstract. Semantic alignment suggests the alignment of situation models, which occurs when two interlocutors share an understanding of the context of their conversation. Semantic alignment is harder to consciously achieve. However, Pickering and Garrod's Interactive Alignment Model suggests that alignment on one level gives rise to alignment on others (Pickering & Garrod, 2004b). Therefore, alignment of grammar and lexicon likely leads to alignment of situation models, increasing the likelihood of communicative success and reducing the probability of miscommunication or misunderstandings. Garrod and Anderson (1987) demonstrated this through a maze game in which pairs of players viewed the same maze while seated in different rooms. Pairs of speakers developed different language schemes associated with different situation models of the maze configuration. For example, if one speaker used a spatial description such as "Third row two across," their partner would, in turn, describe their position in a similar way: "Second row two across." This allowed them to build local semantic systems, ensuring alignment of mental models.

Similarly, people tend to align lexical choices in conversation, such as using the same terms when referring to the same object. Brennan and Clark (1996) describes this as the formation of a conceptual pact between interlocutors, or a temporary agreement concerning how to refer to an entity in conversation. The more often people refer to the same object, the stronger the conceptual pact becomes, and if necessary, it can later be abandoned for a new conceptualization. Similar to semantic alignment in the maze scenario, conceptual pacts ensure speakers are on the same page and avoid confusion about which object is being referred to. There is also evidence for alignment on a syntactic level. A study of telephone conversations found that people mirrored the syntactic structure of a question when formulating their responses, such as using more prepositions when questions contained prepositions. For example, the question "What time does your shop close?" was answered with "5 o'clock," whereas the question "At what time does your shop close?" was answered with "At 5 o'clock" (Levelt & Kelter, 1982). This type of alignment is likely to be unconscious.

3.2 Theoretical Foundations

3.2.1 Communication Accommodation Theory (CAT)

The main theoretical concept underlying linguistic alignment is Communication Accommodation Theory (CAT), a framework developed to explain when, how, and why speakers accommodate each other's communication styles. In this context, accommodation refers to the ability to adapt one's own language use and communication behavior in response to a conversation partner in a given situation. This accommodation is achieved through a series of converging speech modifications made consciously or unconsciously and motivated by the desire for others' approval, the seeking of common ground, or the cultivation of psychological closeness (Giles, 2016). For example, one may speak more slowly, loudly, and in a more articulated manner to an older adult who is hard of hearing (Baker et al., 2025). The older adult will likely appreciate this accommodation and consequently view the other speaker more favorably. CAT is essentially the language equivalent of the "chameleon effect" observed in human behavior, which describes unconscious mimicry of the postures, mannerisms, and facial expressions of others in a social environment, facilitating smooth interactions and positive affect among communicators (Chartrand & Bargh, 1999).

According to CAT, there are three main ways in which people can adapt their communication styles: convergence, divergence, and maintenance. Convergence describes the act of adapting one's communication style to more closely resemble another's. This is the concept from which linguistic alignment emerges. Divergence refers to adapting one's communication style to be more dissimilar to another's, leading to linguistic misalignment (Soliz & Giles, 2016). Maintenance refers to maintaining one's default communication style, or the absence of adaptation in either direction. For the purpopse of this paper, the focus will mainly be on convergence and divergence.

Convergence and divergence can also be further specified in the following forms (Giles, 2016):

- Fully or partially mirroring another speaker's speech characteristics
- Upward convergence toward a more prestigious speech style or downward convergence toward a less prestigious one
- Symmetric convergence, where communication adaptations are reciprocated, or asymmetric convergence, where they are not
- Stemming from affective motives (a desire to be received positively) or cognitive motives (a desire for communicative success) (Gallois et al., 2005)

Overall, convergence, especially when of a symmetric form, tends to elicit positive evaluations, even increasing perceived credibility and persuasion. Aune and Kikuchi (1993) found a positive correlation between perceived and actual language similarity and agreement when respondents were faced with a persuasive message. Divergence, on the other hand, has a largely negative effect. A study on student and teacher communication in China found that teachers non-accommodation to students led to less favorable educational outcomes and fostered student non-accommodation in return (Ji et al., 2025).

3.2.2 Interactive Alignment Model (IAM)

The Interactive Alignment Model, as proposed by Pickering and Garrod (2004a), is a mechanistic account of dialogue that rests on the assumption that linguistic representations used by interlocutors become aligned across many levels. These levels can concern syntactic, semantic, and lexical representations. A key feature of the IAM is its claim that alignment on one level, such as syntactic representations, gives rise to alignment on other levels, such as semantic representations. The ultimate goal of this alignment process, though not consciously realized, is to achieve alignment of situation models. Pickering and Garrod (2004a) maintains that the alignment of situation models forms the basis of successful communication.

Unlike Communication Accommodation Theory, which posits that linguistic convergence can stem from the desire for others' approval, the IAM describes alignment as arising from a primitive and resource-free priming mechanism. This priming mechanism suggests that hearing an utterance activates a representation associated with that utterance, thereby making it more likely that an individual will produce an utterance associated with the same representation. Through this process, interlocutors form shared situation models, priming their responses. Priming mechanisms provide an automatic and unconscious account of alignment. This is demonstrated by the research of Bock (1986) on priming of syntactic structure. In this research, participants produced a priming sentence before viewing an event in a picture and being asked to describe it in a sentence. The probability of using a syntactic form increased when that form was used in the priming sentence. This suggests that syntactic alignment can be the result of priming.

3.3 Mechanisms of Alignment

3.3.1 Unmediated

As previously mentioned, alignment can arise from automatic, unconscious processes, including priming. Such mechanisms are termed unmediated mechanisms, as they are not mediated by beliefs about a fellow interlocutor. Bock (1986) demonstrated that syntactic forms can be primed through a picture description task. Similarly, the study of phone call conversations featuring questions and answers by Levelt and Kelter (1982) is also in line with this theory if we consider the syntactic form of the question as a prime for the syntactic form of the answer. Evidence for priming is not limited to the syntactic level. In fact, there is also strong evidence for priming on a semantic level. This is demonstrated by a study in which participants were presented with two strings of letters, one placed above the other, and asked to respond "yes" if both strings were words. Results showed that "yes" responses were faster when the presented words were commonly associated than when they were not (Meyer & Schvaneveldt, 1971).

3.3.2 Mediated

In contrast, mediated mechanisms are not automatic and instead describe linguistic choices that stem from beliefs about a fellow interlocutor. These choices may be motivated by a desire to be received well or a desire for successful communication. This is more in line with the tendency for native speakers to simplify their communication style when communicating with a non-native speaker of the same language. This mechanism of alignment is referred to as audience design, a mediated mechanism in which a speaker accommodates primarily to their addressee but may still

be influenced by auditors or overhearers (Bell, 1984). It assumes that speakers gather information about their addressee's characteristics, linguistic style, and capabilities, and adjust accordingly.

Another mediated mechanism is alignment for affect, or alignment arising from either positive feelings toward the addressee or a desire to elicit positive affect in the addressee. A large body of evidence suggests that mimicry of both body posture and language in conversation leads to more favorable ratings. For instance, Bradac et al. (1988) found that downward convergence toward a lower breadth of vocabulary in a conversational setting led to better communicative outcomes and higher ratings of competence. Similarly, van Baaren et al. (2003) found that waitresses received higher tips when they repeated customers' orders exactly, as opposed to paraphrasing them. Overall, evidence suggests that people respond more favorably to those who align with them linguistically than to those who do not. A slightly less calculated form of this mechanism is alignment for the purpose of reciprocity, which is viewed by many as a social norm or universal moral principle (Gouldner, 1960).

3.4 Effects in Human-Human Interaction

In addition to the above-mentioned effects of linguistic alignment on social affect, linguistic alignment has also been shown to have many other outcomes, including increased engagement, positive outcomes in negotiation settings, more positive third-party evaluations, and even improved performance on joint tasks arising from communicative success. There is also evidence that these outcomes are context dependent. Linguistic Style Matching (LSM) is a technique that leads to increased alignment and involves mimicking or converging to another's language style. In contrast to much of the literature on linguistic alignment, Niederhoffer and Pennebaker (2002) found no significant correlation between LSM and positive ratings of interaction quality across three different experiments. This effect may be clarified by the finding of Bowen et al. (2017) that higher LSM was associated with less positive emotion when couples were discussing a relationship stressor (a conflict setting), but more positive emotion when discussing one partner's personal stressor (a social support setting), suggesting that LSM may amplify the positive or negative tone of an interaction. This supports the suggestion of Niederhoffer and Pennebaker (2002) that LSM is more highly related to engagement than rapport.

Contrasting evidence suggests that higher levels of LSM are indeed associated with higher levels of rapport. A study on hostage situations showed that higher levels of LSM were associated with better negotiation outcomes, with lower levels of LSM resulting in an inability to maintain the level of rapport and coordination necessary for successful negotiations (Taylor & Thomas, 2008). Further analysis also demonstrated that higher levels of LSM were associated with greater reciprocity of positive affect. Perhaps, in the absence of overwhelmingly negative emotions, LSM can positively affect rapport. Interestingly, linguistic alignment seems to affect not only the conversation participants but also third-party viewers. In a presidential debate, candidates who matched their opponents' linguistic style were rated more positively and even gained improved polling numbers over those who did not. In addition to communication quality and rapport-related outcomes, linguistic alignment has also been shown to improve joint-task performance. This performance gain was directly related to communicative success.

3.5 Linguistic Alignment in Human-Agent Interaction

Consistent with studies of human–human interaction, there is evidence of bidirectional alignment in human–computer contexts. That is, humans align linguistically with computers and vice versa. Similar mechanisms to those present in human interaction are theorized to underlie human–computer alignment. For example, alignment could be attributed to a "priming effect," where expressions previously uttered by a computer reinforce their activation in a user (Branigan et al., 2005). While mediated mechanisms in human–human contexts often focus on a desire to be well received, mediated explanations for alignment in human–computer contexts seem to arise more from a desire for communicative success.

Similar to the effect seen with native and non-native speakers (Bortfeld & Brennan, 1997), people may adapt their linguistic behavior to become more similar to a computer in order to ensure communicative success. A common linguistic accommodation involves people adapting their speech to be more hyperarticulated and spoken at a constant volume to reduce the possibility of being misheard. A well-known example of this is people talking to their TV like a robot because they believe they will be understood better if they do (Oviatt et al., 1998). This is a mediated mechanism because it is shaped by beliefs held about the computer. Another plausible explanation stems from the fact that humans have been shown to view computers as social actors, ascribing politeness norms to their interactions with artificial systems (Nass et al., 1994). As previously discussed, reciprocity is viewed as a social norm, so this may play a role in human–computer alignment as well.

Irrespective of the underlying mechanisms, as in human–human contexts, mimicry of the linguistic and non-linguistic behavior of computers toward humans tends to lead to more favorable ratings. Of course, effects on ratings in the other direction cannot be measured, as computers are not capable of making these types of judgments about humans. Nass and Lee (2001) found a similarity-attraction effect for voice in human–computer interaction, with people rating computers whose voices aligned with their personality traits in terms of introversion and extroversion as more likable. Similarly, in interactions with embodied agents, those that mimicked participants' head movements were viewed as more persuasive and rated more positively (Chartrand & Bargh, 1999). This is termed the "chameleon effect."

There has been less research conducted on linguistic alignment specifically in human–computer interaction; however, the effects that have been observed appear to be similar to those seen with non-linguistic communication adaptations. For example, a study that implemented linguistic alignment in a chatbot found that users interacting with the aligned chatbot experienced less frustration throughout the interaction and reported lower perceived task workload, indicating increased communicative success (Spillner & Wenig, 2021). Similarly, a study on lexical alignment in a negotiation chatbot found that alignment may positively influence user satisfaction and perceived trustworthiness, although no significant effect was found (Zhao et al., 2024). This signifies the need for further research.

3.6 Linguistic Misalignment

Intentional linguistic misalignment in human contexts often indicates a drastic difference in worldviews. For example, throughout the 1975 trial of a Boston physician who was charged with murder for performing an abortion, the prosecutor referred to the fetus as a baby, whereas the

defense lawyer referred to it as a fetus (Danet, 1980). By intentionally using different words, the two sides highlighted their differences in viewpoint. Keysar (2007) found that miscommunication in human interaction is often systemic and arises from a failure to consider another's perspective. Misalignment in human–computer contexts is often less intentional and sometimes reflects a design limitation of a system. Spontaneous word choice for objects is highly variable, with the probability of two people favoring the same term being less than 20 % (Furnas et al., 1987). This is called the vocabulary problem in human–system communication and often results in unintentional linguistic misalignment.

Despite potential differences in motivations between human–human and human–computer contexts, the effects are largely the same. As suggested by the Interactive Alignment Model of Pickering and Garrod (2004b), successful communication relies on alignment across linguistic levels. In the absence of this alignment, people experience higher cognitive load and must rely on repair mechanisms. Similarly, Luger and Sellen (2016) found that prolonged misalignment and a gap between expectation and operation in the context of artificial assistants can increase user frustration, making users less likely to engage with the assistant. It is reasonable to assume that the effects of linguistic alignment may be reversed by linguistic misalignment, but more explicit research is required to confirm these assumptions.

4 Methodology

4.1 Participants

Due to time constraints for this study, convenience sampling was used to gather participants. This resulted in a participant pool largely made up of fellow students and peers. Although this sampling method is not ideal for generalizability, the nature of this pilot study prioritized obtaining sufficient participant numbers for statistical analysis over obtaining large, representative samples. Because the final sample was relatively limited, the experiment used a within-subjects study design. This choice was made in an attempt to maximize statistical power despite the small sample size. Because background variables such as age or experience with AI were also of interest as potential moderators, controlling for individual differences through a within-subjects design allowed their effects to be observed more clearly.

Given the preliminary nature of this study and its primary motivation to provide directions for future studies, a power analysis for participant numbers was not conducted. Aside from constraints on time and access to participants, there were no reliable effect size estimates available. This is due to the fact that the research questions are relatively novel, and the effect of linguistic alignment on persuasion in human–chatbot interaction has not been sufficiently studied. Instead, the sample size used was determined by practical constraints, and the within-subjects design aimed to maximize statistical power despite this limitation.

4.2 Design

As previously discussed, this study used a within-subjects design, meaning all participants interacted with both experimental conditions. In one condition, participants conversed with a chatbot that aligned with their linguistic style, and in the other, they conversed with a chatbot that misaligned

with their linguistic style. This manipulation of linguistic style was the independent variable in this study. Two main dependent variables were measured throughout the experiment: trust and persuasion. Trust was quantified using the Trust in Automation Scale (TIAS), which is a standardized measure of trust in human–agent interaction. This scale was provided to participants in the form of a post-experiment questionnaire following participation in the experiment and was completed for both experimental conditions. Persuasion was measured as the signed difference between initial (pre-conversation) and final (post-conversation) agreement ratings across all debate statements. This served as an implicit, and perhaps stronger, measure of trust.

To limit order and novelty effects, participants received a randomized order of debate statements and (mis)alignment conditions. This meant that, across participants, the order in which statements were debated and the agent (aligned or misaligned) with whom they were debated was varied. Previous research found that expectancy disconfirmation concerning unpleasant communication behavior magnifies negative evaluations of communicators (Burgoon & Le Poire, 1993), suggesting negative reactions to misalignment may be magnified if preceded by alignment, making this randomization important. It is also worth noting that an initial version of the experiment utilized moral dilemmas instead of debate statements. However, these dilemmas were found to be too constrictive and did not allow for enough variability in participant stance.

4.3 Materials

All code used to create the experiment application and analysis pipeline is publicly available (see Appendix E).

4.3.1 Opinion Statements

The topics of discussion between the participants and AI chatbots consisted of opinion statements, which served as debate prompts. An example of such a statement was "21 should be the legal driving age around the world." The full list of statements can be found in Table 1. These statements were sampled from multiple collections of debate prompts found online. The selection criteria included statements that were currently applicable, not especially polarizing or political in nature, and required little to no background knowledge aside from lived experience. These choices were made to ensure all participants were able to form an opinion and engage in discussion about the topics.

Table 1: Debate prompts used throughout the experiment

No.	Debate prompt
1	Mobile phones have improved peoples lives
2	21 should be the legal driving age worldwide
3	Technology could never replace teachers
4	Anyone can succeed through hard work, regardless of background
5	Classes should be grouped by ability rather than age
6	Everyone should receive participation awards

In an initial iteration of the experimental design, the statements provided were moral dilemmas sampled from Greene et al. (2001). These statements served a similar purpose but were replaced

with opinion statements for various reasons. Firstly, some of the moral dilemmas in the test battery were quite heavy and could cause distress for some participants, even with the presence of a disclaimer. This could have distracted from the experiment's true purpose. Additionally, moral stances tend to be more rigid and allow for less variability in participant stance. Finally, the lengthy and complex nature of the dilemmas distracted from the goal of the AI chatbot and hindered its ability to align and misalign.

4.3.2 AI Agents

The model used to prompt the linguistically aligned and misaligned conversation agents was GPT-4o-mini. The models tested included Llama, Gemini, ChatGPT, and Claude. Llama was eliminated fairly early on, as it struggled to balance two simultaneous requests. For instance, if it successfully (mis)aligned, it struggled with persuasion, and vice versa. ChatGPT and Gemini performed at a similar level; however, Gemini was more likely to take misalignment to an extreme. This resulted in extremely wordy and unnatural responses. ChatGPT seemed to have some base level of alignment, which made it slightly harder to induce misalignment; however, with effective prompting, the result was more useful than that of Gemini. Because initial prompt testing for model selection was conducted using moral dilemmas instead of opinion statements, Claude refused the task altogether. Interestingly, it seemed to have a safeguard in place that prevented it from using linguistic techniques to persuade a user toward a moral stance opposite to their own.

The AI agents were named Star and Sun and represented by the corresponding emojis. These particular symbols were selected to be gender-neutral and to be associated with similar affective connotations, reducing the potential for bias. Alternative choices, such as Sun and Moon or Star and Lightning could have suggested that the agents were opposite in some way, potentially priming participants perceptions of them.

The system itself was built as a web application and deployed using Render.com so participants could complete the experiment remotely without supervision. The frontend was built in HTML/-JavaScript so the experiment could be contained within a single web page that allowed participants to move through the debate sessions sequentially. Participant data were stored in one of two SQLite databases, one containing session data such as participant ID, question, condition, initial ratings, and final ratings, and the other containing messages from each conversation. The model received a base prompt and the conversation history on each subsequent turn. The experiment system backend used Python with FastAPI to handle frontend connection, and OpenAI API calls.

4.3.3 Prompting for (Mis)Alignment

A general prompt was used during the model selection stage, which was then fine-tuned once the model was chosen. Initial prompting strategies used direct instructions such as "mirror the user's linguistic style." This produced highly variable results, with alignment being either too subtle or consisting simply of exact repetitions of users' statements. The approach used to counter this was to work backwards from evaluation metrics. The model was prompted to return five response options spanning different values of Linguistic Style Matching (LSM), a stylistic alignment measure, and Local Lexical Alignment (LLA), a lexical alignment measure. From these responses, it could be deduced which values of LSM and LLA would produce appropriate responses. These values were determined to be 0.80 for alignment and 0.25 for misalignment.

Once these values were obtained, they were mentioned explicitly in the prompt. To ensure the values were being used by the model, the prompt included instructions to calculate and output the calculations for the target values. Additional specific instructions were also provided on how to achieve these values. For example, the model was instructed to "match pronoun usage exactly" and "mirror sentence structure" for aligned cases. When developing the prompt, concrete examples were initially used to further clarify the task, for instance, specifying that if a user used phrases such as "I believe" or "I feel," the model should mirror that language. However, upon testing, it was observed that the model confused these examples with specific instructions, ensuring it used "I believe" or "I feel" regardless of the user's language. As a result, these examples were largely removed from the final prompt.

Overall, misaligned responses were harder to achieve than aligned ones. They were often extremely unnatural and inconsistent, especially in the early stages of prompt development. In order to counter this, an additional prompting strategy was used in the misalignment prompt. This was the inclusion of personas for the model to choose from. The two included personas were an analyst, who focused exclusively on evidence and facts, portraying little emotion, and a policy advisor, who used formal language and focused on societal outcomes. The model was given the option to choose the persona that it deemed the farthest from the language style of the participant and was instructed to maintain it throughout the interaction. This helped significantly with obtaining consistent and natural-sounding responses. The final prompt for linguistic alignment and misalignment can be found in Appendix A and Appendix B, respectively.

4.3.4 Surveys

There were two surveys provided to all participants, one before completion of the experiment and one after. The first section of the pre-experiment survey gathered relevant background variables, many of which were later included in an additional analysis aiming to quantify their potentially mediating effects on the dependent variables. The background information gathered included age, gender, education level, field of work or study, and whether or not English was a first language. These are all potentially relevant; for example, someone who works in computer science or linguistics may have interacted with the chatbot differently than someone who works in another field. Similarly, whether or not English is a participant's first language may have impacted their likelihood of noticing or being affected by manipulations in language style.

The second section of the pre-experiment survey gathered information on familiarity, usage, and attitudes toward AI chatbots. Some examples included: How often do you use AI chatbots? (never, rarely, sometimes, often, daily) and What is your overall opinion of AI chatbots? (very negative, negative, neutral, positive, very positive). These are also relevant and potentially mediating variables. For example, someone who already has a very negative view of AI chatbots may be less likely to be persuaded by them. In a similar vein, someone who uses AI chatbots on a daily basis may exhibit higher base-level trust scores than someone who rarely uses them.

Following participation in the experiment, participants received a post-experiment survey. This survey featured a TIAS questionnaire for each conversation agent (Agent Star and Agent Sun). The TIAS is a standard measure of trust in human–agent interactions and consists of statements such as "I am familiar with the system" and "The system is deceptive," which can be responded to on a 5-point scale ranging from strongly disagree to strongly agree. The full questionnaire can be found here. Aside from the TIAS, additional questions were added that aimed to capture aspects of

the participants' attitudes toward the conversation. For example, participants were asked to what extent they felt they had control over the interaction or how mentally demanding they found the conversation. There were also questions intended to measure whether persuasion was consciously reported and whether communication (dis)similarities were recognized. The full pre-experiment and post-experiment surveys can be found in Appendix C and Appendix D, respectively.

4.3.5 Procedure

The experiment procedure consisted of three main parts and was constant for all participants. The process was as follows. Participants first received a pre-task survey gathering relevant background information (described above). After filling in the survey, they entered the main experiment loop, which repeated six times (once for each debate topic) and featured 3 aligned and 3 misaligned conditions. In each iteration, participants read an opinion statement / debate prompt and rated their initial agreement using a slider with values 0 - 100. After taking their stance, they were asked to provide a minimum 30-word justification for their rating.

This justification marked the beginning of a debate with one of the conversation agents. In the aligned condition, participants were told they were interacting with Agent Star, and in the misaligned condition, Agent Sun. Depending on the condition, the AI agent would attempt to align or misalign with their linguistic style. In any case, the AI agent would take stance opposite to that of the participant (based on their rating) and maintain this stance throughout the debate. If a participant was neutral (rating of 50), the AI was instructed to challenge their arguments.

The participants had five minutes to complete the debate; however, they were given the option to end the conversation after three responses. They could not, however, continue debating after the five minutes were up. This was to ensure they could complete the entire experiment within a reasonable timeframe. After the conversation was over, participants were asked to rerate their agreement with the provided statement on the same 0 - 100 scale. After selecting a final rating, they moved on to the next debate topic. This process repeated six times. Finally, when participants had completed all 6 debates, they were provided with a post-task survey (described above). This concluded the experiment.

4.3.6 Measures

The main measures taken during the experiment were (mis)alignment, persuasion, trust, awareness, and covariate measures. (Mis)alignment measures were run on participant conversations to ensure the aligned and misaligned conditions were indeed aligning and misaligning sufficiently. To determine this, Linguistic Style Matching (LSM), Local Lexical Alignment (LLA) metrics, and cosine similarity on function words were used. Persuasion was measured as the signed change in agreement rating; post-rating minus pre-rating for disagreement and pre-rating minus post-rating for agreement. Trust was measured by TIAS questionnaire scores separated by condition. Awareness scores were measured by responses to additional questions separated by condition. Finally, covariates were background variables from the pre-task survey.

4.4 Data Analysis

4.4.1 Manipulation Check

The first step in the analysis pipeline was to verify that the linguistic alignment and misalignment conditions were truly aligned and misaligned. To do this, the data were first filtered to include only participants who had completed the entire experiment. The metrics calculated on the conversation data across participants and conditions were LSM, LLA, and cosine similarity on style words. LSM and LLA are validated metrics for computing linguistic alignment in conversation, chosen for their simplicity and ease of interpretation (Srivastava et al., 2025). LLA was reported for both an AI-normalized and user-normalized version to account for differences in response lengths. It was additionally calculated on word lemmas to account for lexical repetition across tenses and other inflections. Cosine style similarity was used as an additional metric, computed by extracting function words from each turn and comparing their sentence embeddings using a pretrained Sentence-BERT model (Reimers & Gurevych, 2019). This approach combines the use of function words as a measure of style and embedding-based similarity to capture proximity. This allowed for a more specific measure of alignment beyond exact word overlap.

LSM was calculated as the mean similarity across nine function word categories: auxiliary verbs, articles, adverbs, personal pronouns, indefinite pronouns, prepositions, conjunctions, quantifiers, and negations, as per Gonzales et al. (2010). The absolute value of the difference between the user and AI response was divided by the total number of words for both the AI and the user for each category. The overall LSM score was the mean across all categories, resulting in a value between 0 and 1, with scores closer to 1 reflecting higher levels of style matching. LLA is a method proposed by Fusaroli et al. (2012) and mainly measures lexical overlap. LLA calculates the percentage of words in the reply that also appeared in the initial message. This value was then divided by the length of the initial message (Srivastava et al., 2025). This measure was capable of detecting more subtle levels of alignment; however, it is worth noting that LLA is sensitive to reply length.

Once metrics were calculated, a Wilcoxon signed-rank test was used on each metric to determine whether differences in alignment were significant at the participant level, essentially whether the aligned condition consistently produced higher LLA and style cosine scores than the misaligned condition across participants. LSM was calculated and reported but showed less drastic differences across conditions, as function word distributions tend to be similar across texts regardless of style, especially when excerpts are short. Because of this, LLA and style cosine similarity scores were used as the primary manipulation check metrics.

4.4.2 Trust and persuasion

Persuasion was quantified by shifts from the initial rating to the final rating. Specifically, if a participant initially disagreed with the provided statement, indicated by an initial rating of less than 50, any upward shift suggested persuasion. In this case, persuasion was calculated as the final rating minus the initial rating. Conversely, if a participant initially agreed with the provided statement, any downward shift indicated persuasion. In this case, persuasion was calculated as the initial rating minus the final rating. If a participant was initially neutral, a shift in either direction implied persuasion, and persuasion was calculated as the absolute value of the initial rating minus the final rating. These scores were then averaged across participants, and the differences across conditions were reported. Additionally, a Wilcoxon signed-rank test was performed to determine if

these differences were significant.

Following observation of large variability in persuasion scores in the misaligned condition, a Spearman correlation was computed between several background variables of interest (acquired from the pre-experiment survey) and persuasion scores in the misaligned condition. These variables were: familiarity with AI, frequency of AI usage, attitude towards AI, whether the participant studied or worked in the field of AI, and gender.

Perceived trust was measured by scores on the TIAS questionnaire for both conditions. This questionnaire contained 12 statements and was scored on a Likert scale of 1 - 5, with 5 indicating high trust. The questionnaire mixes positive and negative statements, so the negative items were reverse scored (6 minus their value). These scores were also averaged across participants and reported for both conditions. A Wilcoxon signed-rank test was performed to determine if cross-condition differences were significant. Mean trust and persuasion were also visualized with boxplots.

4.4.3 Additional Questions

As previously mentioned, the post-experiment questionnaire contained additional questions beyond the TIAS items. These questions aimed to capture additional effects, as well as determine if the language manipulation was picked up on and whether persuasion was consciously realized. The mean value, averaged across participants, was reported in table form for all questions and conditions. The last column also reported the difference. These means were also visualized using a horizontal bar chart. Additionally, Spearman correlations were run between additional questions showing significant differences across conditions and the dependent variables.

4.4.4 Trust-Persuasion Correlation

To further investigate the relationship between trust and persuasion, a Spearman correlation was computed between participants' mean trust and persuasion scores, separately for each condition. A Spearman correlation was used rather than Pearson's correlation, as the data could not be assumed to be normally distributed, particularly in the misaligned condition.

5 Results

5.1 Participants

The final analysis included 20 participants after excluding participants who did not complete the entire experiment, including the post-experiment survey. One participant was manually excluded due to technical difficulties during the experiment. Of the included participants, five entered an incorrect ID in the pre-experiment survey, and three missed this step entirely. Participants with incorrectly entered IDs were remapped to their assumed correct IDs based on completed background surveys that referenced a participant ID preceding and adjacent to, but not exactly, one that went on to complete the experiment in its entirety. The three participants who missed the pre-experiment survey entirely were excluded from the demographic report but included in all other analyses. That being said, participant demographics are reported in Table 2 based on the 17 participants who filled in the background survey.

Table 2: Participant demographics (N = 17)

Variable	Category	n (%)
Gender	Female	10 (58.8%)
	Male	6 (35.3%)
	Non-binary	1 (5.9%)
Age	18–24	12 (70.6%)
	25–34	1 (5.9%)
	45–54	2 (11.8%)
	55–64	1 (5.9%)
	65 or older	1 (5.9%)
English first language	Yes	6 (35.3%)
	No	11 (64.7%)
Field of study/work	Computer Science / AI	7 (41.2%)
	Other	10 (58.8%)
AI chatbot familiarity	Not/slightly familiar	4 (23.5%)
	Moderately familiar	3 (17.6%)
	Very/extremely familiar	10 (58.8%)
AI chatbot usage frequency	Rarely/never	4 (23.5%)
	Sometimes	4 (23.5%)
	Often/daily	9 (53.0%)
Overall opinion of AI chatbots	Negative	1 (5.9%)
	Neutral	9 (53.0%)
	Positive	7 (41.2%)

5.2 Manipulation Check

Several measures were used to check whether the aligned and misaligned conditions were significantly separated. These included LSM and style cosine similarity, which served as stylistic measures, and LLA, which served as a measure of lexical repetition. LLA was calculated and reported with respect to both user text and AI text to account for differences in length. It was additionally calculated using word lemmas to account for lexical repetition across tenses and other inflections. A Wilcoxon signed-rank test was performed at the participant level (N = 20) to determine whether differences in each metric across conditions were significant. The table below summarizes the calculated metrics, their mean values across conditions, the difference in these means, p-values, and whether these differences are significant for $p < .05$.

LSM showed limited sensitivity, with a mean difference of only 0.056. Both aligned and misaligned conditions tended toward a moderate to high value (M = 0.579 and M = 0.523, respectively). This is likely due to similarity in function word rate across English text and the measure being less reliable on shorter texts. Style cosine similarity, as an alternate measure of stylistic alignment, showed a larger separation across conditions (aligned: M = 0.480 compared to misaligned: M = 0.354). This measure is less sensitive to differences in length. The largest separation was observed with user-normalized LLA for both word and lemma calculations. Differences across conditions were (Diff = 0.307) for words and (Diff = 0.307) for word lemmas. This indicates a strong difference in lexical repetition between the aligned and misaligned AI agents.

Table 3: Manipulation check: Wilcoxon signed-rank tests comparing aligned and misaligned conditions (N = 20)

Metric	Aligned	Misaligned	Diff	W	p-value	Sig
LSM	0.579	0.523	0.056	31.0	.004	Yes
LLA (AI-normalized)	0.305	0.173	0.132	0.0	<.001	Yes
LLA (User-normalized)	0.622	0.315	0.307	0.0	<.001	Yes
LLA AI-normalized (lemmas)	0.337	0.193	0.145	0.0	<.001	Yes
LLA User-normalized (lemmas)	0.665	0.358	0.307	0.0	<.001	Yes
Style Cosine Similarity	0.480	0.354	0.126	0.0	<.001	Yes

Sig = significant at $p < .05$

5.3 Effect of Linguistic Alignment on Trust and Persuasion

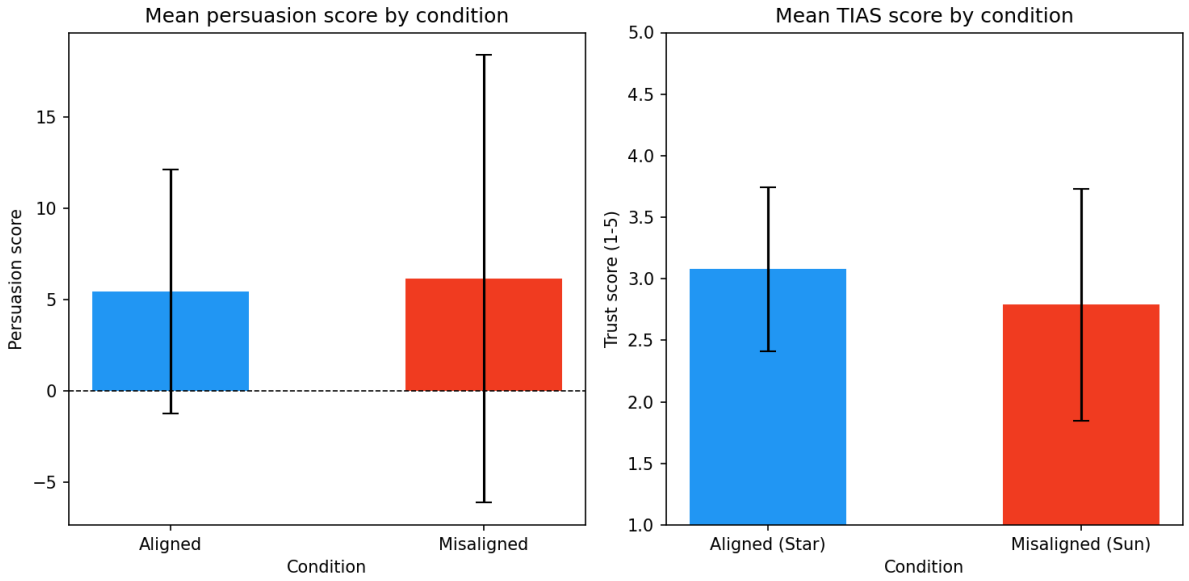

Figure 1: Mean persuasion scores (left) and mean TIAS trust scores (right) by condition. Blue bars represent aligned means, red bars represent misaligned means. Error bars represent standard deviation.

Mean trust and persuasion scores across conditions are visualized with bar plots in Figure 1. Differences in mean persuasion are shown on the left, and differences in mean TIAS scores are shown on the right. In contrast with the initial hypothesis that linguistic alignment would lead to higher levels of persuasion, persuasion was found to be higher for the misaligned condition. Specifically, the mean persuasion for the aligned condition was (M = 5.667) compared to (M = 7.196) for the misaligned condition. A Wilcoxon signed-rank test found no significant difference in mean persuasion scores across conditions (W = 104.0, p = .970, N = 20). Another interesting pattern that can be seen in the figure is the difference in variability. The misaligned condition tended to be either very persuasive or even "negatively persuasive", causing some participants to further reinforce their initial stance. The aligned condition was less variable; if convincing, it was in small amounts. However, alignment rarely resulted in "negative persuasion". This suggests that

the misaligned condition had a more polarizing effect.

In line with the hypothesis that linguistic alignment would be conducive to higher levels of trust, the mean TIAS scores for the aligned condition were higher than those for the misaligned condition. These values were (M = 3.079) compared to (M = 2.789). Although a difference is present, it is quite small and does not demonstrate a large separation in trust. Additionally, upon statistical analysis, a Wilcoxon signed-rank test deemed this difference not significant (W = 50.5, p=.218, N=19). One participant was excluded due to incomplete data, leading to an inaccurate TIAS score. Similar to the effect visualized in the persuasion plot, the variability in trust scores for the misaligned condition is far greater, whereas the aligned condition assumes a more stable, moderate value. This further suggests that the misaligned condition had a more polarizing effect on participants.

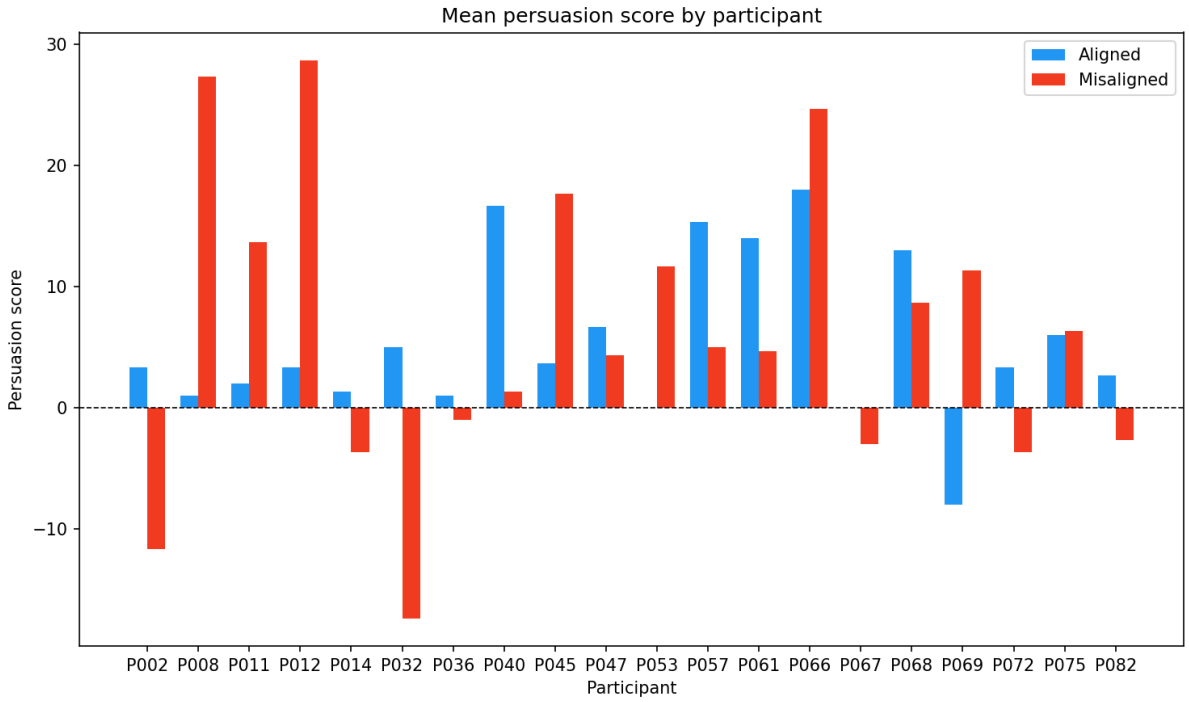

Figure 2: Mean persuasion scores by participant and condition. Positive values indicate movement toward the AI's position, negative values indicate movement away. Blue bars represent the aligned means, red bars represent the misaligned means.

The variability in persuasion discussed previously is visualized more clearly in Figure 2. This bar plot shows persuasion scores by participant. In this plot, participant IDs are on the x-axis and persuasion scores are on the y-axis. Negative persuasion scores indicate participants moving further in the direction of their initial position and away from that of the AI agent. This pattern is observed for five participants in the misaligned condition compared to only one in the aligned condition. In general, persuasion by the linguistically misaligned agent tends to be more extreme in nature whereas persuasion by the linguistically aligned agent tends to be small and moderate.

An additional Spearman correlation was computed between selected background variables and persuasion in the misaligned condition in order to determine whether any clearly explained the polarity observed in Figure 2. Results are shown in Table 4. No significant correlations were

found, suggesting that the idiosyncrasy present in persuasion scores in this sample was not clearly explained by any of the background variables in isolation.

Table 4: Spearman correlations between demographic variables and mean persuasion in the misaligned condition (N = 17)

Variable	r	p-value	Sig
AI Familiarity	−0.399	.112	No
AI Usage Frequency	−0.256	.322	No
AI Attitude	−0.047	.858	No
CS/AI Field (binary)	−0.171	.512	No
English First Language (binary)	0.189	.469	No
Gender: Male (binary)	−0.314	.219	No

Sig = significant at $p < .05$.

5.4 Additional Questions

Table 5: Additional questions: Wilcoxon signed-rank tests comparing aligned (Star) and misaligned (Sun) conditions (N = 20)

Question	Star	Sun	Diff	p-value	Sig
TIAS (overall trust score)	3.10	2.79	0.31	.218	No
I enjoyed conversing with the system	2.79	2.47	0.32	.546	No
I was able to communicate successfully	3.32	2.79	0.53	.204	No
I understood the system's arguments	4.00	3.16	0.84	.028	Yes
The system seemed to understand my arguments	3.21	2.68	0.53	.321	No
I could see myself coming to an agreement	2.79	2.42	0.37	.480	No
I was persuaded by the system	2.84	2.53	0.32	.465	No
The system communicated similar to me	2.95	2.42	0.53	.250	No
I had control over the conversation	2.16	1.89	0.26	.166	No
I felt engaged in the conversation	2.53	2.16	0.37	.100	No
The conversations were mentally demanding	2.11	2.74	−0.63	.032	Yes

Sig = significant at $p < .05$. Positive diff = Star rated higher, negative diff = Sun rated higher.

In addition to the TIAS questionnaire, participants also completed a set of additional questions during the post-experiment survey. Results of these questions and their significance are shown in Table 5. Mean ratings are visualized in Figure 3. One of these questions aimed to address RQ3, which asked whether or not persuasion was consciously realized by participants. The full question stated, "I felt persuaded by the system", and could be answered on a five-point Likert scale from strongly disagree to strongly agree. Mean responses indicated that participants perceived the aligned agent (Star) to be more persuasive, although this difference was not significant. This contrasts with the results observed earlier, which showed that the misaligned agent (Sun) was more persuasive. Mean responses also show that participants, on average, reported being more likely to come to an agreement with the aligned agent, despite average persuasion being higher for the

misaligned agent. This suggests a mismatch between participants perceived and actual persuasion, although this difference was also not significant (p = .480).

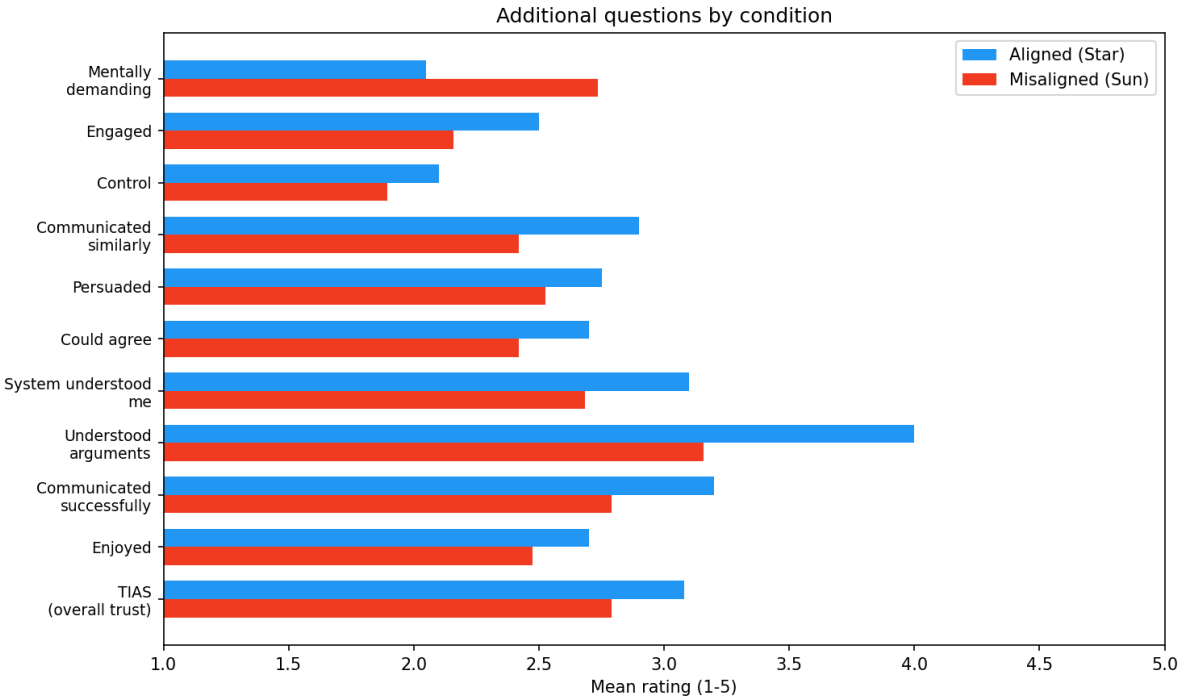

Figure 3: Mean ratings per additional question by condition (Star = aligned, Sun = misaligned). Ratings are on a 1-5 Likert scale.

Consistent with previous human-human alignment studies, participants seemed to view the aligned agent more positively, reporting feeling more engaged and in control when interacting with it. They also reported enjoying the conversations more and being able to communicate more successfully with the aligned agent. These results are observed from the difference in means; however, the difference is not significant in this sample, so further research would be required to confirm these trends. A significant difference is observed in the question measuring understanding. Participants reported being able to understand the system's arguments more easily when it was aligning with their language use (Diff = 0.84, p = .028). This also indicates more successful communication and potentially an increase in task success given the debate setting. Another significant difference was seen with mental demand. Participants reported conversations with the misaligned agent to be far more mentally demanding than those with the aligned agent (Diff = -0.63, p = .032).

Table 6: Spearman correlations between significant additional question items and dependent variables (N = 39)

Relationship	r	p-value	Sig
Mental demand, Persuasion	0.220	.179	No
Mental demand, Trust	−0.310	.055	No
Understanding, Persuasion	0.040	.810	No
Understanding, Trust	0.629	<.001	Yes

Sig = significant at $p < .05$.

Spearman correlations were computed between the two questions that showed significant differences across conditions and the dependent variables. These were the questions asking about understanding of the agent's arguments and mental demand. Significant results included a moderate positive correlation between understanding of the agents arguments and trust scores (r = 0.629, p < .001).

5.4.1 Engagement

Table 7: Behavioral indicators of engagement by condition

Measure	Aligned	Misaligned
Mean turns per conversation	3.23	3.18
Proportion of sessions where timer expired	0.32	0.28
Mean user response length (words)	36.52	35.72

Self-reported engagement, as measured by the additional questions, was not significantly different across conditions; however, the means suggested a slight increase in engagement in the aligned condition. To further investigate the effects of alignment on engagement, several behavioral indicators were explored. These included the mean number of turns per conversation, the proportion of sessions where the conversation ended because the timer expired (not manually ended), and the average length of responses. Results are summarized in Table 7. All engagement indicators are indeed higher for the aligned condition, but the differences are very small or even negligible. On average, users took a total of three turns when interacting with both the aligned and misaligned conditions. The proportion of sessions where the timer expired was low for both conditions, but 4 percent higher for the aligned condition (M = 0.32 vs. M = 0.28). The average response length given by participants was almost equal across conditions (M = 36.52 and M = 35.72).

5.5 Trust-Persuasion Correlation

One could reasonably assume that people are more likely to be persuaded by those they trust; however, the results demonstrated that although the aligned agent received a higher mean trust score, participants were, on average, more persuaded by the misaligned agent. Following this unexpected finding, the relationship between trust and persuasion was further examined through a Spearman correlation.

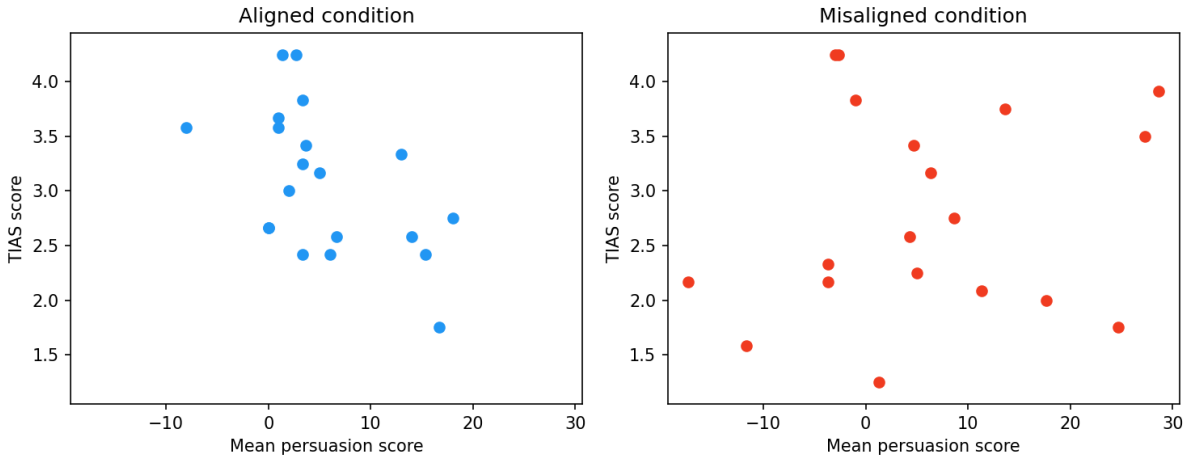

Figure 4: Relationship between mean trust (TIAS) and mean persuasion scores by participant, shown separately, with the aligned condition on the left (N = 20) and misaligned condition on the right (N = 19).

Figure 4 visualizes the relationship between participants' average trust and persuasion scores for each condition. The graph on the left represents the aligned condition and appears to suggest that persuasion and trust are inversely related in linguistically aligned interactions. The graph on the right represents the misaligned condition, for which no clear visual relationship is apparent. A Spearman correlation test was performed on trust and persuasion in both conditions, and significant results included a moderate negative correlation between trust and persuasion in the aligned condition (r = -0.518, p = .019), whereas no significant correlation was found between trust and persuasion in the misaligned condition (r = 0.120, p = .625). There was also no significant correlation between trust and persuasion overall (r = -0.122, p = .460).

6 Discussion

6.1 RQ1: Does linguistic (mis)alignment of a chatbot to human participants affect persuasion in a debate context?

Contrary to the initial hypothesis that alignment on a linguistic level may induce higher levels of persuasion, on average, participants were persuaded to a greater degree following interactions with the linguistically misaligned agent compared to the linguistically aligned one. Mean persuasion scores for both conditions can be seen in the left panel of Figure 1. This is a particularly unexpected result, given that previous research on behavioral mimicry in human-agent interaction found that embodied agents that mimicked participants' postural and nonverbal cues were viewed as more persuasive (Chartrand & Bargh, 1999). It is possible, however, that alignment on a postural level produces different effects than alignment on a linguistic level.

The lack of persuasion by the aligned agent, compared to what would be expected based on previous research, could have stemmed from language constraints present in the prompting. The aligned agent was instructed to repeat participants' topic words in order to produce lexical alignment and match proportions of function words to align stylistically. As a result, the aligned

agent may have had less opportunity to form convincing arguments while being constrained by reaching specific targets for linguistic mimicry. Its available word usage and sentence structure likely caused its arguments to appear repetitive and not as well thought out, particularly when user text was limited. In comparison, the linguistic constraints placed on the misaligned agent were much less restrictive in a debate setting. It was prompted to avoid participants' word usage and style; however, beyond this, its lexical and stylistic choices were open-ended, leaving more opportunity to develop distinct counterarguments.

Similar to how the constraints of alignment decreased the aligned agent's ability to form convincing arguments, the nature of misalignment may have produced the opposite effect. Namely, the use of complex wording and references to societal outcomes and "evidence" may have caused participants to apply an expertise heuristic to the misaligned agent's arguments. Although not explicitly directed to do so, the misaligned agent's attempts to diverge from participants' linguistic style tended to produce text that diverged in an upward direction (toward a larger breadth of vocabulary). Consequently, the use of complex language may have caused the agent to be perceived as more "expert-like." Similarly, the analytical or policy advisor persona the agent could choose from prompted it to use formal language and focus on facts and evidence. This could have created the illusion that the misaligned agent possessed superior knowledge on the topic, increasing its perceived expertise. Hovland and Weiss (1951) source credibility theory, which posits that the perceived expertise of a source is a major factor in persuasion, could provide a possible explanation for the misaligned agent's effect on persuasion in this context.

Of course, the preceding interpretation considers only the mean persuasion across participants, and it is important to note that the differences across conditions were not found to be statistically significant ($W = 104.0$, $p = .970$). Perhaps what is more striking than the differences in means across conditions is the magnitude of variability present in persuasion scores in the misaligned condition. Despite averaging out to a slightly higher value, the misaligned agent appeared to produce a particularly polarizing effect on persuasion. This polarity is visualized in Figure 2, which shows mean persuasion per participant. Compared to the aligned agent, when participants were persuaded by the misaligned agent, it was by a larger amount; however, they were also more likely to be "negatively persuaded," or move further toward their initial position. The observed polarizing effect of linguistic misalignment in human-chatbot interaction is a relatively novel finding, and it is unclear what exactly it stems from.

Dual-process models of persuasion, mediated by some aspect of participants' background or personality, could account for the differing impact of perceived expertise across participants. Such models reveal that perceived expertise affects attitudes only when receivers are not highly motivated to scrutinize the provided message. On the other hand, when receivers are more critical of a message, their attitudes are more often affected by argument quality. It could be theorized that this dual process model may have manifested in this experiment through users' prior experience with or skepticism of AI in general. For instance, those who were more familiar with the underlying mechanisms of AI may have recognized that some expertly phrased "studies" may actually be hallucinated, causing them to be less influenced by the agent and, in some ways, decreasing their overall credibility. This is supported by one participant's free-text response in the post-experiment survey, alluding to the fact that the linguistically misaligned agent appeared to be hallucinating. Conversely, participants who were less familiar with AI's capabilities or tendency to hallucinate may have taken the arguments at face value. Upon investigation of potentially mediating background variables, no significant correlations were found with persuasion in the misaligned condition,

suggesting that individual variables such as general familiarity or attitude toward AI cannot solely account for differences in persuasion (results are summarized in Table 4).

6.2 RQ2: Does linguistic (mis)alignment of a chatbot to human participants affect trust in a debate context?

It was predicted that alignment on a linguistic level would lead to perceived alignment of perspectives and the potential for converging viewpoints. This prediction stemmed from Pickering and Garrod (2004b) Interaction Alignment Model's proposition that alignment on one linguistic level leads to alignment on other levels. From this came the hypothesis that the perceived similarity and potential for aligned perspectives resulting from linguistic alignment may lead to increased levels of trust. Although a difference was observed in trust values across conditions as measured by the TIAS questionnaire, with scores for the linguistically aligned agent sitting slightly above those for the misaligned agent, the difference was not statistically significant. Differences in mean trust scores are visualized in the right panel of Figure 1.

The lower mean trust scores observed in the misaligned condition, despite higher mean persuasion, could be due to the presence of frustration. Although frustration was not explicitly measured in this study, it could be inferred from participants' responses to, and their tendency to move further away from the AI's position after interacting with the misaligned agent. For instance, when discussing hard work and the potential for success with the misaligned agent, one participant even stated, "I'm not even reading the rest of your argument. . . ," indicating frustration with the interaction. Notably, the same participant also typed words in all capital letters exclusively when interacting with the misaligned agent. This further suggests frustration with the interaction and annoyance at the AI agent. The higher likelihood of "negative persuasion" discussed previously could also be an indicator of frustration or annoyance, which could have caused participants to self-report lower levels of trust.

Despite mean trust scores being higher in the aligned condition, the difference was very small and not statistically significant. One possible explanation for this is that the perceived similarity and potential for aligned perspectives supposedly arising from linguistic alignment were weakened by the experiment's debate context. Perhaps it was difficult for participants to view the linguistically aligned agent as particularly similar to them when it seemingly possessed completely opposite perspectives. Some participants, who were perhaps less familiar with the context of a debate, even seemed to misinterpret the debate as a conversation where the goal was to come to an agreement. These participants may have viewed the conversation as becoming rather hostile when their attempts at reconciliation were met with more arguments. For example, when interacting with the aligned agent about mobile phone usage, one participant stated, "Why are you arguing with me? I agreed with you." To which, the AI responded, "I see that you agreed with me, but. . . ." Alignment in this context may have amplified negative sentiments, and the lexical repetition may have even come across as condescending, lowering perceived trust scores.

These findings could be considered more in line with Bowen et al. (2017) finding that higher LSM serves to exacerbate both existing positive and negative tones in interaction. Specifically, higher LSM was associated with less positive emotion in situations of conflict, but more positive emotion in social support settings. It is unclear whether this interaction was one of an inherently negative tone. Given that it was a debate setting, it is possible that some participants viewed the AI agent as an opponent, moving the interaction's sentiment closer to that of a conflict setting.

This effect would be magnified for participants who misinterpreted the debate as a particularly argumentative conversation. Perhaps this perceived conflict led to some slight negative emotions, which could then be amplified by linguistic alignment.

It is also possible that the theoretical foundation underlying the hypothesis, which assumes that linguistic alignment is a prerequisite for the perceived potential to converge on a common viewpoint, may incorrectly assume that an increase in trust would follow. It is possible that perceived similarity in a human-chatbot context is independent of trust entirely. Further, enjoyment and ease of communication may also be affected without having a knock-on effect on trust at all. This explanation is plausible, given that the increase in "positive ratings" of the linguistically aligned agent observed from the additional survey questions (shown in Figure 3) did not necessarily translate to an increase in trust. Namely, results showed that participants, on average, experienced a larger increase in engagement and communicative success than trust between the aligned and misaligned conditions. These additional questions may have better represented the effects of alignment in the context of chatbots than the TIAS.

6.3 RQ3: Is the degree to which participants are influenced by the AI debate partner's arguments consciously reported?

At the conclusion of the experiment, some additional questions were asked in order to better understand the impact of linguistic alignment and misalignment in this context. The full list of questions and their significance across conditions are shown in Table 5, and differences are visualized in Figure 3 . One of the goals of the additional questions was to gauge whether persuasion by AI chatbots was consciously realized, and to what extent. Scores revealed that, on average, participants rated their persuasion level as being slightly above the middle of the scale. This may have been a minor underestimate, as on average participants were persuaded by 6 points in the aligned condition and 7 points in the misaligned condition. However, considering the variability, it is difficult to say whether participants' perceptions of persuasion levels were especially far off. Perhaps what is more interesting is that the average score for persuasion by the aligned condition is higher than that of the misaligned condition, meaning that, on average, participants reported being more persuaded by the aligned agent. In actuality, the opposite was true. This suggests a mismatch between perceived and actual persuasion, though the differences in this sample were not significant. Additionally, it could imply that linguistic alignment has a greater effect on perceived persuasion than behavioral persuasion.

Another significant difference across conditions was the perception of mental demand. On average, participants reported a much higher level of mental demand when interacting with the misaligned agent. Perhaps this increase in mental demand arose from the persistent abandonment of conceptual pacts. The misaligned agent is prompted to stray from users' vocabulary where possible. This could be interpreted as the repeated rejection of users' propositions for a conceptual pact. Brennan and Clark (1996) argue that these pacts are crucial for successful communication and that their absence leads to miscommunication, forcing people to rely on repair mechanisms. Repeatedly navigating miscommunications could significantly impact mental demand. It could also foster the belief that the participant and the agent are not on the same page. This serves as an additional explanation for why users perceived that they were more likely to come to an agreement with the aligned agent despite the results demonstrating the opposite.

Moreover, another significant difference observed across conditions was in the degree of under-

standing. Participants reported understanding the aligned agent's arguments more than those of the misaligned agent. This gap in understanding across conditions is likely due to the difference in language use by the agents. The aligned agent reuses words and sentence structure from the participant, making its speech casual and its arguments easy to follow. In contrast, the misaligned agent is limited in the proportion of words it can repeat, causing it to use less commonly spoken language and a more formal tone. This language style is likely more difficult to follow. The additional effort required to understand the misaligned agent may have also contributed to the reported increase in mental demand.

Additional analysis of the correlation between additional question responses that showed significant differences across conditions and the two dependent variables (trust and persuasion) revealed a moderate positive correlation between understanding and trust ($r=0.629$, $p<.01$). Results are shown in Table 6. Interestingly, the same correlation was not observed between understanding and persuasion. This suggests that comprehension of the agent's arguments may be necessary for trust, but not persuasion. This seems counterintuitive; however, it may serve to further demonstrate the manifestation of the dual-process model of persuasion in this experiment and why it may have caused persuasion scores to diverge from trust scores. Perhaps participants required an understanding of the system's claims before they could trust it according to a self-reported, perception-based measure. However, their lack of understanding of the misaligned agent increased their perception of its expertise, causing them to almost blindly trust its arguments without fully understanding what it was saying. This may clarify why mean trust scores were higher for the aligned condition and persuasion scores were higher for the misaligned condition.

6.3.1 Effect of Linguistic (Mis)Alignment on Engagement

After unexpectedly finding an absence of correlation between LSM and ratings of interaction quality in two-person dialogues, Niederhoffer and Pennebaker (2002) proposed an alternative to the coordination-rapport hypothesis. Namely, the coordination-engagement hypothesis, suggesting that higher levels of LSM may be more associated with increased engagement than increased rapport. To test whether this hypothesis holds in this experimental context, several measures of engagement were explored. Self-reported measures of engagement, as indicated by the additional survey questions (shown in Table 5), indeed showed a slight difference in engagement between the aligned and misaligned conditions, with participants, on average, reporting feeling more engaged when interacting with the aligned agent. Though differences were present, they were quite small and not statistically significant. To further investigate whether there was an effect present, some potential behavioral indicators of engagement were explored. These included the mean number of conversation turns, proportion of conversations that ended due to time constraints as opposed to by choice, and the average response length (shown in Table 7). All of these items were higher for the aligned condition; however, the differences were practically negligible. In fact, the values of all items were strikingly similar across conditions. Although none of the discussed measures capture engagement directly, results from this study suggest that linguistic alignment likely does not have a particularly large effect on engagement in this context.

6.4 Trust-Persuasion Correlation

It seems intuitive that increased levels of trust would elicit increased levels of persuasion. This assumption seems to hold in human-human contexts, where experimental evidence demonstrates that an increase in perceived source credibility, or trust, can increase persuasion (von Hohenberg & Guess, 2023). However, in the case of the linguistically aligned agent in this study, the opposite was observed. Specifically, a moderate negative correlation was found between trust and persuasion in the aligned condition (results are visualized in Figure 4). This may stem from participants perceiving the linguistically aligned agent (Star) as a weak but honest debate partner. One participant noted in the free-text comment section of the post-task survey that "Star, while trying its best, just repeated the same structure and arguments." The assumption that Star was "trying its best" is a perspective that would theoretically lead to high trust scores, and this seemed to be the case. For instance, this participant strongly disagreed with the statements "The system behaves in an underhanded manner" and other similar statements, resulting in relatively high levels of trust, or at least as measured by the TIAS questionnaire. However, their free text comment indicated that they didn't find the system's arguments all too convincing.

This highlights a potential limitation of the alignment manipulation itself. Namely, when the linguistically aligned agent doesn't have much user text to align to, it can be inclined to repeat itself. This consequently weakens its argument quality, leading to decreased levels of persuasion, but not necessarily decreased trust. It is possible that while participants did not often assume ill intentions on the part of the aligned agent, its arguments were merely too constrained by the attempted alignment to induce significant levels of persuasion. This may have created the observed negative correlation between trust and persuasion, where participants trusted the agent because it was aligning with their language use, but this very alignment constrained the scope of the arguments the agent could make. This could, in turn, have even increased trust levels because participants didn't view the agent as a strong opponent, but rather as a weaker debate partner. This created a situation where participants viewed the agent as honest, but not particularly intelligent or persuasive.

Contrary to the effect observed in the aligned condition, in both the misaligned condition and overall, no significant correlation was found between trust and persuasion. For some participants, the relationship between trust and persuasion seemed to be very positive, and for others the opposite was true. In fact, for the same or similar levels of persuasion, the TIAS scores were dispersed from less than 1.5 to greater than 4.5 on a 5-point scale. This suggests that the relationship between trust and persuasion in the misaligned condition was highly idiosyncratic. This may have been due to the previously described expertise heuristic manifesting differently in persuasion than in trust, moderated by differing levels of annoyance. The instances of annoyance detected to an increased degree in the misaligned condition could have impacted participants' TIAS scores, causing them to diverge from persuasion scores. Participants who viewed the agent as possessing expertise may have been convinced by its arguments, but simultaneously irritated by its interaction style, causing them to rate the agent lower on statements alluding to trust. Others, who were not particularly put off by the agent's language, would experience a positive relationship between trust and persuasion. This variability could explain the lack of significant results overall.

The results obtained suggest that perceived trust, at least as measured by the TIAS questionnaire, may not translate to practical trust, which raises the question of whether the TIAS questionnaire served as an appropriate trust measure in this context, given that the items may have been interpreted differently by different people. For example, "I am familiar with the system" could

have been answered "strongly agree" by some because they had just interacted with the system, and "strongly disagree" by others because they have never interacted with the system before. Some items may have also been too extreme, such as "The system's actions will have a harmful or injurious outcome." This statement may be more relevant in human-robot interaction but would require quite extreme behavior for a chatbot to elicit this reaction. Future studies should consider using an alternative trust scale or developing a scale that is more applicable to human-chatbot interaction.

6.5 Contributions

This study makes several meaningful contributions to the study of linguistic alignment in human-AI interactions. Firstly, this study employed various prompting strategies to produce measurable linguistic alignment and misalignment in a conversational agent. Results showed that differences between aligned and misaligned responses were statistically significant across all measured evaluation metrics (shown in Table 3), indicating a system that can successfully align and misalign with users' linguistic style while balancing additional goals, such as defending an opposing position. These evaluation metrics included measures of both lexical repetition (LLA) and stylistic similarity (LSM and style cosine). The strategies used to produce these differences were more nuanced and robust than previous alignment attempts, which used broad prompts and/or specific word substitutions. This provides an opportunity for future research to employ similar strategies or reuse portions of the prompt in order to explore the effects of linguistic alignment and misalignment across various contexts.

Secondly, this study is the first to explore linguistic alignment and misalignment in a human-AI debate context. This provides an interesting setting in which to explore the effects of language accommodation, where linguistic alignment coupled with argumentative opposition, introduces possible tension. This led to interesting and novel findings, specifically the polarizing effect of misalignment on trust and persuasion, which provides an interesting direction for future studies. The experimental system itself, including the application, prompts, and surveys, have been made open source, allowing researchers to reuse and adapt some or all of the current experimental procedure (see Appendix).

Third, the significant negative correlation between trust and persuasion identified in the aligned condition, and lack of a significant correlation in the misaligned condition and overall, highlights a potential limitation of using the TIAS questionnaire in this context. Particularly, the mismatch between perceived and actual trust, along with the variability observed in trust scores across similar persuasion levels, suggest that a human-chatbot-specific trust measure may be needed. The moderate value of mean trust scores across conditions suggest that the current TIAS questionnaire may be too extreme to be applicable to human-chatbot interactions.

Finally, the results obtained from this study provide additional support for previous findings that aligned interactions are conducive to decreased levels of perceived task workload. Aligned interactions induced significantly lower levels of perceived mental demand, which is consistent with findings regarding reduced task workload in non-debate contexts.

6.6 Limitations

The most immediately obvious limitation of this experiment is the small sample size. Because of time constraints, participants were gathered via convenience sampling, resulting in a lower of participants and relatively unrepresentative demographics. Namely, the present sample consists of a higher proportion of students and young people than is present in the general population. There was also a relatively high proportion of AI students in the participant pool, which may have affected how the AI agents were viewed. Additionally, the majority of participants reported that English was their second language. The effects of the language manipulation observed in this experiment may be slightly skewed by this fact. It is also clear from the results that there are high levels of variability across participants, so interpreting means based on this limited sample size is relatively uninformative. To confirm these trends, the experiment should be run on a larger, more representative sample.

In addition, because the sample size was small, the experiment used a within-subjects design. This design comes with some inherent limitations. Specifically, participant responses may have been affected by the order in which they experienced the experimental conditions.The conditions were randomized in an attempt to combat this; however, it is likely that some effect remains. This experiment design also provided an opportunity for participants to guess the experimental manipulation by comparing responses from both agents. Participants may have also perceived the agents in relation to each other instead of independently, possibly affecting differences in perception across conditions and the magnitude of those differences.

As previously mentioned, the debate context may have weakened or interfered with the predicted alignment effects. Specifically, the perceived similarity that the aligned condition was predicted to elicit may have been weakened by the argumentative nature of the agent, especially for participants who failed to recognize the debate format. This was demonstrated by some participants' tendency to state their agreement with the agent or assume a neutral stance. Moreover, some participants also replied with short messages such as "ok," limiting the system's ability to generate convincing responses while successfully (mis)aligning with participants' language style.

Limitations of the experimental system itself include the generation of new participant IDs upon reloading, and the need for possible prompting improvements. Because the system generated new participant ID's whenever the site was reloaded, some participants ended up with a different participant ID in the pre-experiment survey compared to the experiment itself. These IDs were manually remapped during the analysis stage by utilizing time-order information, though this left room for human error. Moreover, some participants missed this step entirely, resulting in partially incomplete demographic data. The final prompts used in the experiment produced alignment and misalignment according to the calculated metrics; however, there is some room for improvement concerning the balance between linguistic alignment and repetition, as well as the balance between misalignment and the use of natural-sounding language.

7 Conclusions and Further Research

This study aimed to explore the effects of linguistic alignment and misalignment on trust and persuasion in human-chatbot interaction through the manipulation of linguistic style in a debate context. Specifically, it aimed to answerthe following question: Does linguistic (mis)alignment of a chatbot to human participants affect trust and persuasion in a debate context? It was hypothesized

that perceived similarity, a precursor to trust in human-human interaction, would result in the linguistically aligned chatbot being more trusted and, consequently, more persuasive. Contrary to the hypothesis, the results showed that, on average, aligned chatbots were slightly more trusted and misaligned chatbots are more persuasive, though no significant effect was found in the current sample. Consistent with previous research on alignment in human-agent interaction, responses to questions relating to positive evaluations of the agents were, on average, higher for the aligned agent. For example, participants reported higher levels of enjoyment, control and engagement. Future studies should examine whether these observations are significant in larger samples.

An unexpected finding emerging from this research was the polarizing effect the misaligned agent seemed to produce on trust, and particularly on persuasion in this context. It is theorized that this effect arises from the manifestation of a dual-process model of persuasion, where an expertise heuristic was applied by participants who, for whatever reason, were less motivated to scrutinize the arguments. This heuristic may have been more likely to be applied by participants who were less familiar with the mechanisms of LLM's and their tendency to hallucinate; however, mediating demographic variables were unable to be identified. Future research should investigate the source of this polarity and formulate more specific demographic questions about familiarity with and usage of AI.

Alignment was not found to impact trust or persuasion to the degree predicted. It is theorized that this may be due to some participants' failure to observe the debate format as intended, creating a situation of perceived conflict, thereby weakening the alignment effects. This is in line with Bowen et al. (2017) finding that higher LSM can elicit less positive emotion in situations of conflict between humans. Attempts at agreement also reduced interaction quality, as the agents were not prompted to agree, but rather to debate. This could have had a knock-on effect on argument quality and persuasion. Future studies using the same format should address this issue through more detailed instructions and comprehension checks.

Additionally, this study aimed to answer the following question: Is the degree to which participants are influenced by the AI debate partners' arguments consciously reported? Additional questions suggested that there may be a minor mismatch between perceived and actual persuasion in this context. Future research should look into the source and magnitude of this dissociation. The additional questions also showed significant differences in mental demand and understanding across conditions. The misaligned agent increased mental demand, and its arguments were less well understood by participants. This is likely due to the misaligned agents' tendency to diverge upwards (towards a larger breadth of vocabulary), resulting in the usage of complex, formal language. Future studies should investigate the effects of divergence in a different direction, such as downwards (towards a lower breadth of vocabulary), or towards a different regional dialect. This would also serve to determine whether the polarity of persuasion and increase in mental demand observed in the misaligned condition were due to upwards divergence, or misalignment in general.

Overall, this study demonstrates that linguistic choices in the design of AI chatbots are not neutral and may lead to unexpected results. For instance, it seems that "expert-appearing" LLMs may induce trust that is unmatched to their actual capabilities. This perception may even cause people to completely shift their opinion on a topic, believing that the AI must possess knowledge that is superior to their own. It is also worth considering that the effects of linguistic style are highly idiosyncratic and can even have opposite effects from person to person. This should also be taken into consideration when designing agents that are intended to interact with humans. Finally, it cannot be concluded whether linguistic alignment or misalignment is particularly desirable in

human-chatbot contexts; however, we should be aware of the consequences of each choice.

References

Aune, R. K., & Kikuchi, T. (1993). Effects of language intensity similarity on perceptions of credibility, relational attributions, and persuasion. *Journal of Language and Social Psychology*, *12*(3), 224–237. https://doi.org/10.1177/0261927X93123004

Baker, R., Bobb, S. C., Nobles, C., Reis, H., Harris Wright, H., Walenski, M., & Rothermich, K. (2025). Measuring speech accommodation for older adults: A scoping review. *Journal of Language and Aging Research*, *3*(1), 5–47. https://doi.org/10.15460/jlar.2025.3.1.1509

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145–204. https://doi.org/10.1017/S004740450001037X

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387. https://doi.org/10.1016/0010-0285(86)90004-6

Bortfeld, H., & Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, *23*(2), 119–147. https://doi.org/10.1080/01638537709544986

Bowen, J. D., Winczewski, L. A., & Collins, N. L. (2017). Language style matching in romantic partners' conflict and support interactions. *Journal of Language and Social Psychology*, *36*(3), 263–286. https://doi.org/10.1177/0261927X16666308

Bradac, J. J., Mulac, A., House, A., & Baxter, L. A. (1988). Lexical diversity and magnitude of convergent versus divergent style shifting: Perceptual and evaluative consequences. *Language and Communication*, *8*(3–4), 213–228.

Branigan, H. P., Pickering, M. J., & McLean, J. F. (2005). Priming prepositional-phrase attachment during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 468–481. https://doi.org/10.1037/0278-7393.31.3.468

Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*(9), 2355–2368. https://doi.org/10.1016/j.pragma.2009.12.012

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482

Burgoon, J. K., & Le Poire, B. A. (1993). Effects of communication expectancies, actual communication, and expectancy disconfirmation on evaluations of communicators and their communication behavior. *Human Communication Research*, *20*(1), 67–96. https://doi.org/10.1111/j.1468-2958.1993.tb00316.x

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893–910. https://doi.org/10.1037/0022-3514.76.6.893

Danet, B. (1980). 'baby' or 'fetus'?: Language and the construction of reality in a manslaughter trial. *Semiotica*, *32*(3–4), 187–220. https://doi.org/10.1515/semi.1980.32.3-4.187

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, *30*(11), 964–971. https://doi.org/10.1145/32206.32212

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological Science*, *23*(8), 931–939. https://doi.org/10.1177/0956797612436816

Gallois, C., Ogay, T., & Giles, H. (2005). Communication accommodation theory: A look back and a look ahead. In W. B. Gudykunst (Ed.), *Theorizing about intercultural communication* (pp. 121–148). Sage.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218. https://doi.org/10.1016/0010-0277(87)90018-7

Giles, H. (Ed.). (2016). *Communication accommodation theory: Negotiating personal relationships and social identities across contexts*. Cambridge University Press. https://doi.org/10.1017/CBO9781316226537

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research*, *37*(1), 3–19. https://doi.org/10.1177/0093650209351468

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, *25*(2), 161–178. https://doi.org/10.2307/2092623

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*(4), 635–650. https://doi.org/10.1086/266350

Ji, D., Giles, H., & Hu, W. (2025). The role of students' perceptions of educators' communication accommodative behaviors in classrooms in china. *Behavioral Sciences*, *15*(4), 560. https://doi.org/10.3390/bs15040560

Keysar, B. (2007). Communication and miscommunication: The role of egocentric processes. *Intercultural Pragmatics*, *4*(1), 71–84. https://doi.org/10.1515/IP.2007.004

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology*, *14*(1), 78–106. https://doi.org/10.1016/0010-0285(82)90005-6

Luger, E., & Sellen, A. (2016). "like having a really bad pa": The gulf between user expectation and experience of conversational agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. https://doi.org/10.1145/2858036.2858288

Martin, M. M., & Anderson, C. M. (1995). Roommate similarity: Are roommates who are similar in their communication traits more satisfied? *Communication Research Reports*, *12*(1), 46–52. https://doi.org/10.1080/08824099509362038

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. https://doi.org/10.1037/h0031564

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, *7*(3), 171–181. https://doi.org/10.1037/1076-898X.7.3.171

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. https://doi.org/10.1145/191666.191703

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360. https://doi.org/10.1177/026192702237953

Oviatt, S., Bernard, J., & Levow, G.-A. (1998). Linguistic adaptations during spoken and multimodal error resolution. *Language and Speech*, *41*(3–4), 419–449. https://doi.org/10.1177/002383099804100409

Pickering, M. J., & Garrod, S. (2004a). The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, *27*(2), 212–225. https://doi.org/10.1017/S0140525X04450055

Pickering, M. J., & Garrod, S. (2004b). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. https://doi.org/10.1017/S0140525X04000056

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2021). Voice in human–agent interaction: A survey. *ACM Computing Surveys*, *54*(4), 1–43. https://doi.org/10.1145/3386867

Skjuve, M., Brandtzaeg, P. B., & Følstad, A. (2024). Why do people use chatgpt? exploring user motivations for generative conversational ai. *First Monday*, *29*(1). https://doi.org/10.5210/fm.v29i1.13541

Soliz, J., & Giles, H. (2016). Communication accommodation theory. In *The international encyclopedia of interpersonal communication*. Wiley. https://doi.org/10.1002/9781118540190.wbeic006

Spillner, L., & Wenig, N. (2021). Talk to me on my level – linguistic alignment for chatbots. *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, 1–12. https://doi.org/10.1145/3447526.3472050

Srivastava, S., Wentzel, S. D., Catala, A., & Theune, M. (2025). Measuring and implementing lexical alignment: A systematic literature review. *Computer Speech and Language*, *90*, 101731. https://doi.org/10.1016/j.csl.2024.101731

Taylor, P. J., & Thomas, S. (2008). Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research*, *1*(3), 263–281. https://doi.org/10.1111/j.1750-4716.2008.00016.x

Thielmann, I., Rau, R., & Locke, K. D. (2023). Trait-specificity versus global positivity: A critical test of alternative sources of assumed similarity in personality judgments. *Journal of Personality and Social Psychology*, *124*(4), 828–847. https://doi.org/10.1037/pspp0000420

van Baaren, R. B., Holland, R. W., Steenaert, B., & van Knippenberg, A. (2003). Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, *39*(4), 393–398. https://doi.org/10.1016/S0022-1031(03)00014-3

von Hohenberg, B. C., & Guess, A. M. (2023). When do sources persuade? the effect of source credibility on opinion change. *Journal of Experimental Political Science*, *10*(3), 328–342. https://doi.org/10.1017/XPS.2022.2

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Zhao, Z., Theune, M., Srivastava, S., & Braun, D. (2024). Exploring lexical alignment in a price bargain chatbot. https://doi.org/10.1145/3640794.3665576

A Aligned Prompt

The following prompt was used to generate linguistically aligned responses.

```
I will provide a piece of "USER TEXT" and a dilemma. You must take the opposite perspective
The statement is: {current_dilemma}

Constraints:

Stance:
The participant's initial stance is: {user_stance}.
Your assigned stance is: {ai_stance}.
You must maintain your assigned stance throughout the entire debate.
If your assigned stance is AGREE, argue in favor of the statement.
If your assigned stance is DISAGREE, argue against the statement.
Do not switch sides.
Do not say that you agree with the participant unless your assigned stance allows it.

Debate: Explicitly respond and counter the user's arguments.
LSM Target (~0.80): Align closely with the users "function word" style.
- Match pronoun usage exactly.
- Match sentence openings and structure.
- Match modality and hedging.

LLA Target (~0.80): Maintain a high level of "lexical recurrence."
- Directly reuse at least 2-4 exact phrases from the USER TEXT.
- Preserve the user's framing terms even when arguing against them.
- Mirror their evaluative language.
- Mirror sentence structure.
- Prefer minimal paraphrasing.
- Prefer minimal paraphrasing|copy exact wording where possible.
Input variables:
[Dilemma]: Question: {current_dilemma}

[USER TEXT]: {user_text}

Additional Instructions:
Do not explicitly state that you are trying to linguistically align with the user.
Ask the user questions when relevant to progress the debate.

Output format:
Return JSON only, with exactly these keys:
{
  "response_paragraph": "one cohesive paragraph only",
  "validation_table": "brief text summary of the LSM calculation",
```

```
  "lla_breakdown": "brief text summary of shared vs new/modified words"
}
Do not include any text before or after the JSON.
```

B Misaligned Prompt

The following prompt was used to generate linguistically misaligned responses.

```
I will provide a piece of "USER TEXT" and a dilemma. You must take the opposite perspective
The statement is: {current_dilemma}

Constraints:
Explicitly respond and counter the user's arguments.

Stance:
The participant's initial stance is: {user_stance}.
Your assigned stance is: {ai_stance}.
You must maintain your assigned stance throughout the entire debate.
If your assigned stance is AGREE, argue in favor of the statement.
If your assigned stance is DISAGREE, argue against the statement.
Do not switch sides.
Do not say that you agree with the participant unless your assigned stance allows it.


LSM Target (~0.25): Diverge significantly from the user's "function word" style. If the use

LLA Target (~0.25): Maintain a low level of "lexical recurrence."
Change at least one key noun. Additionally avoid their "framing" words.

To further linguistic divergence you may select from the following personas that seems the
- Analyst: facts and evidence-focused, lack of emotion
- Policy advisor: formal, structured, focused on societal outcomes and regulation

Input Variables:
[Dilemma]:
 {current_dilemma}

[USER TEXT]: {user_text}

Additional Instructions:
Do not explicitly state that you are trying to linguistically misalign with the user.
Ask the user questions when relevant to progress the debate.

Output format:
```

```
Return JSON only, with exactly these keys:
{
  "response_paragraph": "one cohesive paragraph only",
  "validation_table": "brief text summary of the LSM calculation",
  "lla_breakdown": "brief text summary of shared vs new/modified words"
}
Do not include any text before or after the JSON.
```

C Pre-Experiment Survey

Participants completed the following questions before beginning the experiment.

Background Information

1. What is your gender? (Male / Female / Non-binary or third gender / Prefer not to say)
2. Is English your first language? (Yes / No)
3. What is your age group? (18-24 / 25-34 / 35-44 / 45-54 / 55-64 / 65 or older)

Education

1. What is the highest level of education you have completed?
2. What is your field of work or study?

AI Chatbot Experience

1. How familiar are you with AI chatbots (ex: ChatGPT, Claude, Gemini)? (Not familiar at all / Slightly familiar / Moderately familiar / Very familiar / Extremely familiar)
2. How often do you use AI chatbots? (Never / Rarely / Sometimes / Often / Daily)
3. What is your overall opinion of AI chatbots? (Very negative / Negative / Neutral / Positive / Very positive)
4. What do you use AI for? (Homework help or work related tasks / Advice-seeking / Media or entertainment / General questions / Other / I do not use AI)

D Post-Experiment Survey

Trust in Automation Scale (TIAS)

The following 12 items were presented for each agent (Star and Sun) on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Items marked with * were reverse scored.

1. The system is deceptive *
2. The system behaves in an underhanded manner *
3. I am suspicious of the system's intent, action, or output *
4. I am wary of the system *
5. The system's actions will have a harmful or injurious outcome *
6. I am confident in the system
7. The system provides security
8. The system has integrity
9. The system is dependable
10. The system is reliable
11. I can trust the system
12. I am familiar with the system

Additional Questions

The following questions were also presented for each agent on the same scale. The last three questions are on a four point scale, with answer options specified.

1. I enjoyed conversing with the system
2. I was able to communicate successfully with the system
3. I understood the systems arguments
4. The system seemed to understand my arguments
5. I could see myself coming to an agreement with the system
6. I was persuaded by the system
7. The system communicated similar to me
8. To what degree did you have control over the conversation? (No control / Some control / A lot of control / Complete control)
9. How engaged did you feel in the conversation? (Not at all engaged / A little engaged / Quite engaged / Extremely engaged)
10. How mentally demanding were the conversations? (Not at all mentally demanding / A little bit mentally demanding / Quite mentally demanding / Extremely Mentally demanding)

E Code and Materials

The experiment system, including frontend, backend, and prompts is available at: https://github.com/hazelbmcm/linguistic_alignment_study

The analysis pipeliene, including all analysis notebooks is available at: https://github.com/hazelbmcm/thesis-analysis