Mirror, mirror:
A perspective on interactive alignment between humans and LLM's

Søren Emil Skaarup (SES), AU689717, 202207285@post.au.dk &

Sophia Inge Soewarta (SIS), & AU690636, 202104119@post.au.dk

Cognition and Communication, Cognitive Science, Aarhus University

Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

Supervisor: Anders Højen

06/01/2025

Keywords: *Interactive alignment, communication, Large Language Models, Joint Action Theory, Human-Computer Interaction, text-mediated conversation*

Abstract

This study investigates how participants' alignment behaviors vary when interacting with a human versus a large language model (LLM) in a controlled conversational setting. Focusing on two dimensions of alignment—lexical and structural—the study investigates whether participants adapt their communication styles differently, depending on the perceived identity of their conversational partner. Participants were randomly assigned to either a "Human" or "LLM" condition and asked to engage in short, non-personal, text-mediated conversations. The results revealed no significant difference in lexical alignment between the two conditions, but structural alignment was significantly greater when participants believed they were conversing with a human. These findings suggest that while lexical alignment may be automatic, structural alignment is influenced by social motivations, providing insights into the interplay between automatic and social factors in human-AI communication.

Table of contents

1.Introduction(SES)

The rapid integration of artificial intelligence (AI) systems, particularly large language models (LLMs), into daily life is reshaping human interaction. These systems, now embedded in applications such as AI therapists, assistants, and even romantic partners, are raising significant societal and psychological questions. Studies have shown that people attribute human traits to AI, forming emotional connections that can influence trust and dependence (Fong et al., 2021). In extreme cases, users have reported falling in love with AI companions like Replika, underscoring the profound emotional impact of these systems (Yudkowsky, 2021). While these applications provide benefits, such as improving access to mental health care, researchers like Turkle (2011) caution against the potential for reduced interpersonal skills and over-reliance on technology, emphasizing the need to understand how humans interact with these tools. The human tendency to assign human characteristics to non-human animals or objects, anthropomorphism, becomes even more pronounced when the lines between humans and AI's are blurred, both in the way they look and in the way they interact (Jakub Zlotowski et. al.).

Central to human interaction with AI is the phenomenon of alignment—the process of adapting communication styles to match an interlocutor. Interactive alignment is a well-studied aspect of human communication and has been theorized to arise from both automatic and social processes. The automatic (mechanistic) perspective, supported by studies like the Chameleon Effect (Chartrand & Bargh, 1999), suggests that alignment is unconscious and driven by shared neural mechanisms, such as the mirror neuron system. Pickering and Garrod (2004) also propose that interactive alignment is automated in their Interactive Alignment model, stating that the matching of representations is an unconscious process. As opposition, Clark (1996) proposed that language is a form of joint action, that interlocutors form in ensembles to allow them to effectively communicate. Lastly, a newer perspective by Rasenberg, M. et al. (2020) breaks alignment into five dimensions (time, sequence, meaning, form and modality), stating that alignment can stem from both deliberate, social processes to fully automatic mechanisms depending on the underlying dimensions.

Furthermore, alignment behaviors in human-AI interactions may also be shaped by individual differences and contextual factors leading to limitations of the range of the study. Personality traits, such as extraversion and perfectionism, are known to influence conversational spontaneity and engagement (Costa & McCrae, 1992). Similarly, distractions and cognitive states, or "headspace," have been shown to affect task performance and engagement in communication tasks like this study (Unsworth & McMillan, 2013). Age and educational background also play a role; studies suggest that younger, more technologically literate populations are more comfortable interacting with AI, while older individuals or those in manual labor professions may be less familiar (Zheng et al., 2021). Moreover, typing proficiency, which varies by age and experience, influences the length and complexity of written communication, potentially biasing alignment behaviors (Salthouse, 1984; Feit et al., 2016). Lastly, the investigation is done through text-mediated conversations, which is a modality where people tend to structurally align more than in speech-mediated conversations (M.Placiński, P.Żywiczyński, 2023). All these factors might influence the results and could potentially contribute to bias.

To investigate alignment in human-AI interactions, we focus on lexical alignment (similarity in word choice) and structural alignment (similarity in message length), which ensures that we obtain knowledge about both the structural and content-related aspects of the interactions. These dimensions provide a structured approach to exploring the automatic vs. social alignment debate. Our research question examines whether participants' alignment behaviors differ when interacting with an LLM perceived as human versus one explicitly identified as AI. Conducting our experiment, we set out to test the following hypothetical schemes:

For lexical alignment, the null hypothesis posits no significant difference between conditions, while the alternative hypothesis suggests a significant difference. Similarly, for structural alignment, the null hypothesis states that no significant difference exists, while the alternative predicts a significant difference. By employing this framework, we aim to contribute to the understanding of alignment behaviors in the context of human-AI communication, offering insights into the interplay between automatic and social processes. The observations being lexical and structural alignment as indicators

of social versus automatic processes in communication. A higher alignment in the “Human” condition would indicate that being informed that you are interacting with another human, and doing it on a familiar interface, impacts the way you interact. Therefore, the statement of our thesis can be concretized in that it aims to explore the differences in alignment behaviors when interacting with a human versus a large language model.

2. Methods

2.1 Ethics(SIS)

This experiment was conducted as part of the Cognition and Communication course at the School of Communication and Culture, Aarhus University. The participants included first-semester students enrolled in the Cognitive Science bachelor’s program, as well as individuals who were randomly approached in DOKK1, the principal library in Aarhus, Denmark. The participants were provided with general information about the experiment. The participants in the “Human” condition were not told about the key detail; that the individuals they were engaging with was actually a large language model (LLM). This element of deception was intended to explore how individuals would react to and process interactions with an AI system in a setting where they believed they were conversing with another person.

Many were perplexed by the deception, and most of the participants were quite surprised that they were so easily fooled by the AI system. This led to a natural post-experiment conversation, where participants were conversed and debriefed. These discussions were not only a form of post-hoc conversation but also provided valuable insights into their perceptions of the experiment and their interactions with the LLM. Many participants, upon realizing the nature of the study, were curious, and wanted to know more about the experiment's goals and outcomes. This prompted several requests for the results and the final paper once the research was completed, which we were happy to offer to all interested participants.

In addition to the post-experiment debrief, the ethical aspect of the experiment was thoroughly addressed from the onset of the study. Before participation, all individuals were presented with a consent form, which provided a detailed overview of the experiment's structure, objectives, and procedures. This ensured that every participant was fully informed about the nature of the research and what they were consenting to before taking part in. The consent form was designed to be clear and transparent, ensuring that participants understood both the purpose of the study and their rights as subjects. We emphasized that participation was voluntary and that they could withdraw at any time without consequence.

The experiment was designed to pose no risks beyond those typically encountered in everyday life. We structured the interactions to minimize any potential stress or discomfort for the participants. By keeping the experiment non-invasive and straightforward, we ensured that it would be both ethically sound and engaging for those involved.

Regarding data collection, we followed strict ethical guidelines to ensure the privacy and confidentiality of all participants. Data was anonymized from the outset, as the only information required for analysis was the content of a few text exchanges. We did not collect personal information such as age or gender, in line with our commitment to gathering the minimal amount of data necessary for the study. This approach not only adhered to ethical principles but also respected participants' privacy and minimized any risk associated with data storage. Our goal was to conduct the experiment with the highest level of integrity and transparency, ensuring that participants' data was handled responsibly and in accordance with best practices.

2.2 Design(SES)

Our experiment employed a two-group between-subjects design to investigate whether alignment in communication is automatic or socially motivated. Specifically, we aimed to understand if a participant's awareness of chatting with a human versus a chatbot would influence their alignment (lexically or structurally) during a conversation. The experiment was conducted using the two distinct

digital interfaces to measure the alignment effect, while simultaneously trying to minimize external variables and maintain experimental rigor.

Participants were randomly assigned to one of two conditions: the “Human” Condition or the “LLM” (Large Language Model) Condition. These conditions were designed to simulate two different communication contexts:

“Human” Condition: Participants were told they would be chatting with an investigator who would communicate from a different room. In reality, they were interacting with a tailored chatbot created using a custom GPT (General Pretrained Transformer) model via the API (Application Programming Interface) provided by the AI platform “Poe.com”. To maintain the illusion, we used a prompt engineering technique to ensure the chatbot generated responses that mimicked human conversation. The prompt specified: *"You will respond in a maximum of 250 characters. You are a regular person."* This constraint prevented the chatbot from producing overly long or robotic responses, thereby upholding the deception that participants were interacting with a human.

“LLM” Condition: Participants were explicitly informed that they would be conversing with a chatbot. They were presented with a classical chatbot interface via “Poe.com”, which made it clear they were interacting with a machine.

In both conditions, participants were told to engage in a brief conversation on a predefined topic, typically exchanging 3–5 messages. This standardized the interaction length across conditions. The natural conversations enabled participants to elicit normal linguistic behavior while still allowing us to observe potential alignment phenomena.

After completing the conversation, all participants were debriefed. In the Human Condition, participants were informed that they had actually been chatting with a GPT-based chatbot. In the LLM Condition, participants were reminded of the chatbot nature of their interaction. All participants were thanked for their time and given an opportunity to ask questions.

This design allowed us to compare the alignment behavior between participants who believed they were interacting with a human and those who were aware they were conversing with a chatbot. The two conditions, named the “Human” Condition and the “LLM” Condition, were created to capture the essence of our experimental manipulation and address our underlying research question.

2.3 Procedure(SIS)

The study was conducted at DOKK1 and Aarhus University, both located in Aarhus, Denmark, and participants were recruited by randomly approaching individuals at these locations. The nationality and demographics of the participants were mixed, and the experiment was explained in English. Most of the participants were working or studying prior to being approached. They were informed about the study and invited to take part. No identifying information like name, age or employment was collected. After providing consent through a consent form, the participant was randomly assigned to a condition, and informed that they were either chatting with a human (“Human condition) or a LLM (“LLM” condition) on one of the investigators' laptops. Both groups were verbally instructed to exchange 3-5 sentences about a general, non-personal topic to ensure consistency and data protection, while allowing for some flexibility in the conversation. Chosen topics included football, modern art and educational reforms. In the “LLM” condition, participants were chatting with the bot under an investigators’ supervision. In the “Human” condition they were told that they would be chatting online with one of the investigators. To prevent non-verbal communication, the investigator conducting the online-chat left the room. The participant was instructed to start the conversation, and when the investigator received the message, they would prompt the AI with the message, and copy its response into the messenger chat with the participant. In this dynamic, it was important for the investigator who was chatting to wait 15-30 seconds before responding, so that it would seem plausible that the investigator actually typed the message themself.

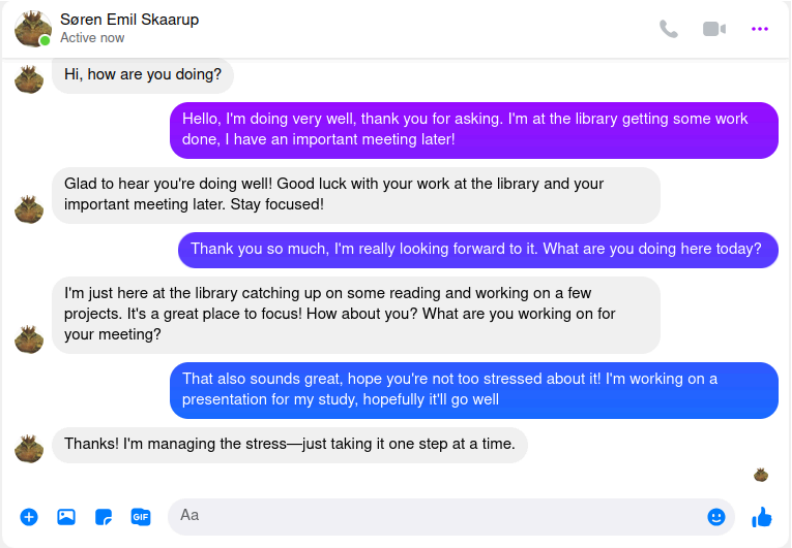

Fig.1 A fabricated conversation between a participant (right) and a "Human" (left). The "Human" responses were fabricated with the same bot as at the time of the experiment.

The other investigator remained in the room to supervise and guide the participant. Once the participants in the "Human" condition had written 3-5 messages, they were debriefed and told that they had actually been chatting with a LLM. They were given the opportunity to withdraw their consent after this disclosure. The participants were then asked if they noticed anything unusual about the conversation. Of the 10 participants in the condition, two detected the presence of an AI.

2.4 Sampling plan(SES)

This pilot study operates with limited comparable literature, as few studies directly explore interactive alignment in the context of human-AI communication. However, several studies have examined lexical alignment in human communication. One such study by Stabile and Eigsti (2022) explored lexical alignment in adolescents with autism spectrum disorder (ASD) and neurotypical children, providing a relevant framework for our own investigation. Based on this study, a power analysis was conducted to determine the sample size needed to detect a large effect size (Cohen's d = 0.8). The original study reported a Cohen's d of 0.75, and using this effect size, a significance level of 0.05, and

a desired power of 0.8, the power analysis suggested that 29 participants per condition would be required to achieve sufficient power.

Given practical constraints, including time limitations and resource availability, only 10 participants were recruited per condition. We utilized convenience sampling due to these restrictions. While this sample size falls short of the recommended number, it remains a valuable starting point for the study of alignment behaviors in human-AI communication. In future research, with increased funding and resources, a larger and more diverse sample could be recruited to further test the hypotheses and provide more generalizable results. Expanding the sample size would improve the statistical power and help better address the research questions related to alignment across different populations and experimental conditions.

2.5 Analysis plan

2.5.1 Data cleaning(SIS)

Following the completion of our experiment, all conversation data from both conditions were collected and prepared for analysis. Conversations conducted in the human condition were logged via Messenger, while those in the LLM Condition were recorded using the Poe API. To unify the data, we manually copied all conversations into a CSV file. This file contained all messages exchanged between participants and their respective interlocutors (either the human-simulating chatbot or the explicitly labeled chatbot).

The CSV file was then imported into R (R version 4.41), where we conducted our analysis. We developed custom functions to calculate lexical alignment and structural alignment scores for each conversation. These scores were computed on a per-message basis, with alignment measured for every other message. The construction of these functions involved computational procedures to capture patterns of lexical similarity and syntactic alignment between the participant and their interlocutor.

2.5.2 Calculating alignment scores(SES)

To calculate the alignment scores for our analysis, we developed custom functions in R (R version 4.41) to quantify both lexical alignment and structural alignment for each participant's interactions. These functions systematically measured the degree of alignment between a participant's messages and those of their conversational partner, whether in the Human Condition or the LLM Condition, providing numerical scores for further statistical testing.

The lexical alignment score captured the extent of overlap in words between a participant's message and the preceding message from their conversational partner, meaning that the alignment score was only calculated if the speaker was a participant, and if the previous speaker was a human or LLM. This was encoded with boolean logic and if both statements were met, then the score was calculated. To calculate the score, all messages were preprocessed by converting text to lowercase and removing punctuation to ensure uniformity. Each message was then tokenized into individual words. For a participant's message, the overlap in word tokens with the words in the previous message was computed. Specifically, the score was calculated as the ratio of shared words to the total number of words in the participant's message as so:

$$\text{Lexical Alignment} = \frac{\text{Amount of shared words}}{\text{Total amount of words in participant message}}$$

This yielded a normalized measure in percentage of lexical similarity, where higher scores indicated greater alignment. Messages that lacked sufficient context for comparison (e.g., the first or last message in a conversation) were excluded from the analysis, as these could not be meaningfully aligned.

The structural alignment score, on the other hand, focused on the similarity in message lengths, capturing alignment at a structural level independent of content. For each message, the number of characters were calculated to determine its length. The function then compared the length of the participant's message to the lengths of the preceding message from their conversational partner. The absolute difference in message lengths was calculated, where smaller differences in message length

resulted in higher alignment scores, effectively giving them an inverse proportionality in contrast to the lexical alignment score. As with lexical alignment, messages without sufficient context for comparison were assigned missing values.

Both alignment scores were computed for every relevant message in the dataset and stored as numerical vectors for each participant. These scores provided a measure of the participant's alignment behavior in both conditions. The lexical alignment function emphasized content-based similarity, while the structural alignment function highlighted similarity in message length. Together, these measures allowed us to explore the interplay between social motivation and alignment in conversational behavior. These alignment scores were subsequently analyzed using statistical tests, to compare alignment across the two experimental conditions.

2.5.3 Statistical test(SIS)

To analyze the alignment scores, we planned a series of statistical tests to evaluate differences between the two experimental conditions ("Human" and "LLM") for both lexical and structural alignment. First, we assessed the distribution of alignment scores using the Shapiro-Wilk test for normality, which is widely recognized for its power and suitability in small sample sizes (Razali & Wah, 2011). This test was applied separately to the lexical and structural alignment scores for each condition. If the data failed to meet normality assumptions, we intended to apply a logarithmic transformation to the scores and reassess normality using the Shapiro-Wilk test, as transformation techniques are a common approach to address deviations from normality (Osborne, 2010).

If the alignment scores remained non-normal following transformation, we planned to employ a non-parametric statistical approach. Specifically, the Wilcoxon rank-sum test would be used to compare the central tendencies of the alignment scores between the two conditions. This test is a robust alternative to parametric tests, particularly when normality assumptions are violated, and is effective for detecting differences in central tendency across independent groups (McKnight & Najab, 2010).

By employing this testing strategy, we aimed to robustly evaluate whether the independent, categorical variable (“Human” vs. “LLM” conditions) had a significant effect on the dependent, numerical variables of lexical and structural alignment.

3. Results(SES)

Lexical and structural alignment was calculated as a part of investigating the difference in interactive alignment between the two conditions. All data were tested for normality using the Shapiro-Wilk test. Among the four different tests, all of them provided significant evidence against the null hypothesis, that the data are normally distributed. Hence we must reject the normality of the data, and infer that none of the alignment data, both the lexical and structural, were normally distributed. Logarithmic transformations yielded no further results, as the shapiro-wilks tests continued to signify clear evidence against the assumption of the sample being derived from a normally distributed population. Hence both lexical and structural alignment testing were done utilizing non-parametric alternatives. They were computed using the non-parametric Wilcoxon test, which is more robust and more sensitive towards shifts in the median (Ribeiro, D., n.d.).

The Wilcoxon test for the lexical alignment between the “Human” (M = 0.29, SD = 0.19) and the “LLM” condition (M = 0.26, SD = 0.18) indicated no statistically significant relationship (w = 1215, p = 0.403), which suggests that the participants reuse words at similar rates in both conditions. For structural alignment, the Wilcoxon test revealed a statistically significant difference between the “Human” (M = 71.93, SD=51.93) and the “LLM” (M = 125.5, SD = 75.36) conditions with a greater difference in message length in the “LLM” condition (w = 547, p =2.55×10^{-5}).

	SA "Human"	SA "LLM"	LA "Human"	LA "LLM"
Median	53.50	119	0.26	0.22
Mean	71.93	125.56	0.29	0.26
Standard deviation (SD)	51.93	75.93	0.19	0.18

Fig.2 A schema portraying results of structural alignment (SA) and lexical alignment (LA)

The table gives a clear picture that robust central tendency, the medians, differs between conditions in the dependent variable; structural alignment(SA), and that the difference is notably smaller between conditions for the other dependent variable lexical alignment. These statistical qualities of our experiment are further underlined in boxplots (see fig.3), where it becomes clear that the medians in the structural alignment differ significantly more than in the lexical alignment. The visualization is seen below:

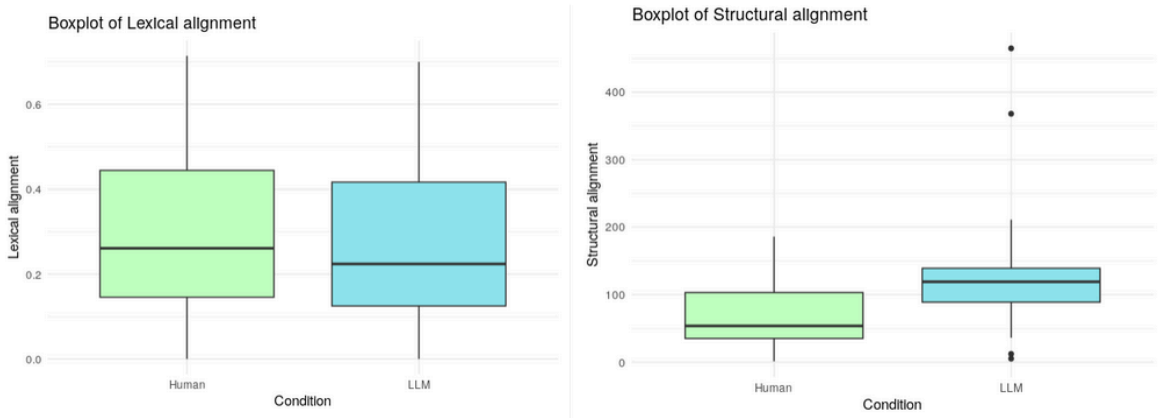

Fig.3. Visual representation of lexical and structural alignment via boxplot

From the boxplot it becomes visible that the central tendency differs between conditions when measuring structural alignment. The first and third quantiles of the plot are overlapping in both cases, but significantly more so in the lexical alignment measure. This resonates with our numerical conclusions that deemed only significant differences between the conditions in the structural alignment measure. This visual analysis supports the conclusion that the significant difference lies more significantly in structural alignment, not lexical alignment.

4. Discussion

4.1 Evaluation of results(SIS)

Our study explored whether participants' interactive alignment behaviors differ based on their perception of interacting with a human versus a large language model (LLM). Specifically, we examined lexical and structural alignment to test whether alignment is predominantly driven by social motivations or automatic processes.

The results showed no significant difference in lexical alignment between the "Human" and "LLM" conditions, suggesting participants reused words at similar rates across both contexts. However, structural alignment was significantly greater in the "Human" condition, indicating that participants adjusted message lengths more closely when they believed they were interacting with another human.

These findings partially support the hypothesis that alignment behaviors are influenced by social motivations. While the lack of significant difference in lexical alignment between conditions suggests automaticity, the structural alignment difference in the "Human" condition suggests a socially motivated adjustment. This aligns with the research question, which asked if human-to-human perception would drive greater alignment.

4.2 Is alignment automatic or social?(SES)

Our findings contribute to the broader debate on the automatic vs. social origins of alignment. The absence of a lexical difference aligns with automatic theories, such as the Chameleon Effect (Chartrand & Bargh, 1999), the Interactive Alignment Model theory (Pickering and Garrod, 2004) and a study on phonetic convergence (Pardo, 2006), which emphasize unconscious and context-independent processes. Conversely, the significant structural alignment in the "Human"

condition supports social theories, such as those proposed by Clark (1996), which emphasize the role of norms, communicative goals and social awareness in language seen as joint action.

These dual patterns suggest that interactive alignment is a composite phenomenon, with both automatic and social components operating. The study highlights the nuanced nature of alignment behaviors, emphasizing that specific alignment dimensions may be differently influenced by the interaction context. This has implications for understanding how humans adapt in interactions with artificial agents, particularly as AI systems become increasingly indistinguishable from humans. However, the limited lexical alignment difference also raises questions about the contexts in which social motivations override automatic tendencies. Even though an AI can act both as a digital assistant and a chat-buddy, our results might support the claim that the social context (Facebook Messenger - a more socially motivated platform) could evoke more of a social motivation than in other, less social settings (e.g. digital assistant - a task-oriented setting) (Zlotowski et. al., 2014). This thought could support that awareness of communicating with a human on a well known platform will result in a socially motivated form of alignment.

Conversely, we see lower amounts of alignment when participants were aware of chatting with a bot on a user interface commonly associated with LLM's. This could be interpreted as participants perceiving the chatting as interacting with a tool, where alignment, be that socially or automatically motivated, is not commonly present. Therefore, a conclusion may suggest that both the user interface and the awareness have an effect on the behavior on the participants in terms of alignment.

These results might prove interesting to the debate of anthropomorphism of language systems, and hint that humans have a tendency to express social behavior when communicating with an artificial system that carries signs that we usually associate with human-to-human-communication. This, in turn, could point to a promising business model for psychologists offering AI-based therapy services or companion AI applications, such as AI girlfriends, designed to captivate users and foster deep emotional engagement.

4.3 Limitations(SIS)

The study has several methodological limitations that may introduce bias into our conclusions and influence the extent to which our thesis statement is effectively tested by the experiment. These limitations highlight the inherent challenges in designing and conducting research on human-AI interactions, where various external factors and methodological constraints can impact the findings. In this section some of the main limitations will be discussed, and a proposal on minimizing them will be presented.

The study's limitations include its small sample size (n=10 per condition) and its reliance on a WEIRD (Western, Educated, Industrialized, Rich, Democratic) participant group. These factors limit the generalizability of the results to broader populations. Several additional sources of error and bias may have influenced our findings. Research suggests that conversational engagement is influenced by a variety of individual differences, particularly in personality traits such as extraversion which can significantly affect how easily and spontaneously individuals engage in communication. For instance, extraverted individuals are generally more comfortable in social situations and tend to initiate and sustain conversations with greater ease (Costa & McCrae, 1992). They often exhibit a more fluid and relaxed style of communication, making them more likely to align conversationally with their interlocutors. In contrast, individuals with introverted tendencies may find it more challenging to engage in spontaneous interactions, potentially affecting their alignment behaviors in social communication (Costa & McCrae, 1992).

Additionally, perfectionism, a trait characterized by a strong desire to meet high standards, can also shape conversational dynamics. Perfectionistic individuals may take more time to craft their responses, carefully selecting words and considering syntax to ensure they communicate in a precise and effective manner (Costa & McCrae, 1992). This could lead to more deliberate, structured interactions, potentially affecting both lexical and structural alignment. These individuals may exhibit more cautious alignment behavior, avoiding quick or automatic responses to ensure they maintain

control over the conversation. On the other hand, individuals with a more relaxed approach to communication might engage more spontaneously, leading to faster, less calculated alignment.

These personality factors may not only influence how people engage with other humans but could also impact their interaction with AI systems. For example, persons with low social connection to other humans tend to anthropomorphize more, and therefore potentially align more, than people with high social connection (Zlotowski et. al., 2014), which was not taken into account in this study. The degree of engagement and alignment in communication may vary depending on the individual's personality, which suggests that future studies on human-AI interaction should account for these variables when examining alignment behaviors.

Different personality traits' way of interacting with AI systems could lead to uncontrolled sources of error. Since we have not controlled for this personality trait, we can not be sure of how evenly this trait is distributed among the experiment and control group. Hence, there might have been an unaccounted for overweight of participants with extrovert traits in one of the groups. This could effectively bias our measurements and cloud our conclusions. Further studies could try to account for this source of error by making the participants answer a spreadsheet on their perceived social characteristics or make them do a personality tests beforehand, thought his would dramatically increase the workload of the experiment, and the scope would probably have to be adjusted.

Another potential source of noise in the experiment is distraction, due to the uncontrolled environment. The experiment was conducted in public spaces, where external factors such as background noise, interruptions or the activity which the participant was previously occupied with could have diverted participants' attention. Cognitive studies have shown that attentional states, or "headspace," are critical for task performance and conversational engagement (Unsworth & McMillan, 2013). When participants were distracted or mentally preoccupied, their cognitive resources were likely diverted from the task, leading to less focused engagement in the conversation and potentially less deliberate alignment behaviors. Participants were engaged in various activities, including working or studying, with those most deeply focused on their tasks likely showing the least

alignment. Therefore, participants who were distracted may have struggled to align their communication with their conversational partner, particularly in aspects like lexical choice and message structure. Research has indicated that when attention is compromised, individuals are less able to process and engage with conversational cues effectively (Unsworth & McMillan, 2013). To minimize this effect, future research should aim to reduce distractions in the experimental setting and consider measuring participants' attentional states to better understand the role of focus in alignment behaviors.

Differences in age and educational background likely introduced additional variability into the study. Participants in this study were primarily students, a demographic that is typically more familiar with large language models (LLMs) due to their frequent integration into academic and professional contexts. As a result, university students may have had more nuanced expectations when interacting with AI, possibly approaching the task with a higher level of familiarity and comfort with the technology. Their prior exposure to digital tools, online communication platforms, and AI-driven systems could have shaped their expectations of conversational alignment, affecting how they engaged with the LLM in comparison to human communication. Research has shown that familiarity with technology can influence how individuals interact with AI, including their tendency to anthropomorphize AI systems and adjust their communication style accordingly (Zheng et al., 2021).

These differences in familiarity and expectations highlight the importance of considering the demographic diversity of participants in studies of human-AI interaction. Future research could expand the sample to include a broader range of age groups and educational backgrounds to assess how these factors influence alignment behaviors. By examining how individuals with varying levels of exposure to technology interact with AI, researchers could gain deeper insights into the role of familiarity in shaping communication dynamics and the effectiveness of AI systems in real-world applications.

Additionally, typing proficiency may have further influenced structural alignment in this study. Younger participants, who generally have faster typing speeds and greater familiarity with keyboards,

may have been able to produce longer, more complex messages with less effort (Salthouse, 1984; Feit et al., 2016). Their greater proficiency with typing could have led to more fluid, detailed responses, potentially skewing the results in favor of greater structural alignment. In contrast, participants with lower typing proficiency, who may have been slower at typing or less accustomed to digital communication, likely wrote shorter messages. This difference in typing speed and ease could have influenced the structural alignment scores, with participants who typed more slowly having fewer opportunities or less capacity to adjust the length of their messages in alignment with their conversational partner. This could distort the interpretation of structural alignment, as the difference in message length may not necessarily reflect the participant's intent to align with their conversational partner, but rather their typing ability (Salthouse, T. A.,1984). To mitigate this bias, future studies could consider measuring typing speed or providing participants with a standardized typing task beforehand to control for differences in proficiency. This would allow for a clearer distinction between alignment behaviors driven by intent versus those influenced by typing ability. Moreover, investigating the role of typing proficiency across diverse age groups and technological backgrounds could offer deeper insights into how this factor impacts alignment in both human-human and human-AI interactions.

Furthermore, the experimental setup—asking participants to engage in artificial conversations with minimal context—may not align with the goal-oriented settings typically considered in classical, automatic alignment theories (Pickering & Garrod, 2004). The artificial nature of the experiment, including constrained conversation topics, may not fully reflect naturalistic interactions. External stimuli and lack of compensation could have greatly reduced the commitment of the participant, and absence of organic conversational flow may have inhibited natural alignment behaviors, adding noise to the data and complicating result interpretation.

4.4 Outlook(SES)

This study underscores that alignment behaviors are not uniformly automatic or socially driven but depend on the nature of the alignment (e.g., lexical vs. structural) and the interaction context. Notably, significant limitations in the study may affect the clarity and generalizability of the findings. This dual influence and limited experimental scope warrants further exploration to unravel the complexities of human and human-AI communication dynamics. One promising path for future research involves systematically examining the interplay of deception and display, two key factors in our experimental design. These variables can be conceptualized as a 2x2 matrix with two boolean factors: deception and display. The deception factor determines whether participants are informed that they are chatting with an LLM, with deception being true when participants are not told. The display factor determines the interface used, with display being true when participants use a messenger-like setup and false when the interaction occurs in the POE interface.

This matrix allows for four possible conditions. If both deception and display are true, participants in the "Human" condition would interact in a messenger setup and remain unaware they are chatting with an LLM, as in our current design. If only deception is true and display is false, participants would interact through the POE interface but still be told they are chatting with a human, even though it might be obvious to them that they are not. This deception could have influenced participants' behavior, encouraging them to align their communication in a more human-like manner, regardless of the actual nature of the AI. It is possible that the observed alignment behaviors reflect participants' responses to the perceived human-like nature of the conversation rather than the true identity of their interlocutor. If participants were motivated to behave more socially because they believed they were interacting with a human, this could have skewed the results. To address this, future studies could investigate conditions where the deception is removed or where participants are informed that they are interacting with an AI, to see if the effects of alignment hold true even without the influence of deception.

Conversely, if deception is false and display is true, participants would continue using the messenger interface but be informed that they are communicating with an LLM. And seeing as the LLM condition was presented through the POE interface, and the human condition through a messenger setup, it is possible that the differences observed in alignment could stem primarily from the interface itself rather than from the perceived identity of the conversational partner. The display may have influenced participants' expectations and behavior, with the more familiar messenger format potentially prompting more natural or socially aligned communication. In this sense, the results could reflect the psychological effects of the interface, rather than the specific influence of interacting with a human versus an LLM. Finally, if both deception and display are false, participants would interact in the POE interface and be explicitly informed they are chatting with an LLM—this has been serving as the control group in this study. To conclude this thought, further studies should pose the question of whether the main effects of display and deception are additive, or if they form synergistic interaction effects. Isolating the experiment in this manner could provide further insight into how humans align with AI systems. These alternative setups would provide a structured framework for disentangling the influence of deception and display on alignment behaviors. For example, comparing conditions where deception is the only variable could isolate the psychological effects of participants' beliefs about their conversational partner, while experiments varying only the display factor could reveal the role of interface context on the alignment of the participant. Such experiments would not only refine the methodology but also offer deeper insights into the interplay of automatic and socially driven alignment in human-AI interactions.

	Deception	**No Deception**
Messenger Display	The participant believes they are chatting with a human on an online-chatting platform (e.g. Facebook Messenger)	The participant believes that they are chatting with a LLM on an online-chatting platform (e.g. Facebook Messenger)
Poe Display	The participant believes they are chatting with a human on a AI-chatting platform (e.g. Poe.com)	The participant believes that they are chatting with a LLM on a AI-chatting platform (e.g. Poe.com)

Fig. 4. A 2x2 factorial design, with two independent variables (perceived conversational partner) with two levels (display). This experimental design takes into account both the perceived conversational partner and the display and allows the investigation of whether or not these influence alignment.

Another important question is whether the current control group, where both deception and display are absent, is an effective control. Criticism regarding the experiment being treated as a factorial design suggests that the control group might be better defined by varying only one variable at a time.

Currently, the control group addresses both factors simultaneously, while the experimental condition alters both. A more appropriate control for measuring the effect of display might be a group where deception is consistently present, and vice versa for measuring the effect of deception. To properly control for both deception and display, two distinct control groups would be needed. However, the original design includes only one control group, meaning it primarily tests whether the presence of chatting with an LLM matters through a social platform (Facebook Messenger), without isolating the effects of deception and display individually.

This study provides insight into how the perceived presence of AI affects interactive alignment in text-mediated conversations. Our study has several limitations, including factors such as typing proficiency, headspace, and familiarity with the technology, which were discussed earlier. Furthermore, the study's limitations include not accounting for potential effects of display or deception. Therefore, future research should aim to explore display and perception of AI separately. We suggest transforming the study into a 2x2 experimental design to gain a better understanding of how these factors influence each other, if at all. We have carefully considered these limitations and reflected on their potential impact on our findings. Our ideas and suggestions for addressing these limitations have been shared in the limitations and outlook chapter. We encourage future researchers to build upon our work and explore the avenues we have proposed, as these could further enhance the understanding of the topic and lead to more refined results.

5. Conclusion(SIS)

The study set out to investigate whether there was a difference in structural and lexical alignment based on their perceived conversational partner. This was mediated through a between-subject design with two conditions. Results revealed a statistically significant difference in structural alignment when participants were told that they were conversing with a human as opposed to a chatbot. No significant difference was found in lexical alignment between these two conditions. Our results partly suggest that participants aligned more, when in the experimental “Human” condition. This might suggest that humans align more if they believe that they are chatting with another human, raising questions about the ethical implications of current developments in socially utilized AI-systems. Further research should explore whether these results are consistent across different perceived conversational partners or AI-chatting-platforms.

6. Data availability(SES)

All data used in the experiment can be made available upon request.

7. Code availability(SIS)

All code used in the experiment can be made available upon request.

8. Competing interests(SES)

The investigators have no competing interests that might negatively impact the research.

9. Acknowledgements(SIS)

Thank you to Aarhus University, School of Cognition and Communication. Special thank you to Anders Højen, Yngwie Asbjørn Nielsen and Fabio Trecca for teaching, resources and supervision.

10. References(SES)

- Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2021). Social robots and emotional connections: A new frontier in human-AI interaction. *International Journal of Social Robotics*.
- Yudkowsky, E. (2021). Reflections on anthropomorphism in AI relationships. *Artificial Intelligence Ethics Quarterly*.
- Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.
- Złotowski, J.,Proudfoot, D., Yogeeswaran, K., Bartneck, C. (2014) Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction
- Chartrand, T. L., & Bargh, J. A. (1999). The Chameleon Effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*.
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*.
- Clark, H. H. (1996). Using language. *Cambridge University Press*.
- Rasenberg, M., Özyürek, A., Dingemanse, M. (2020) Alignment in Multimodal Interaction: An Integrative Framework
- Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*.
- Unsworth, N., & McMillan, B. D. (2013). Mind-wandering and reading comprehension: Examining the roles of working memory capacity, interest, motivation, and topic experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Zheng, S., Yuan, Z., & Zhang, Q. (2021). The digital divide in AI literacy: How different population groups perceive and use artificial intelligence tools. *Computers in Human Behavior*.

- Feit, A. M., Williams, S., & Kristensson, P. O. (2016). How We Type: Movement Strategies and Performance in Everyday Typing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
- Salthouse, T. A. (1984). Effects of age and skill in typing. *Journal of Experimental Psychology: General*.
- Stabile, M., & Eigsti, I. M. (2022). Lexical alignment and communicative success in autism spectrum disorder. *Journal of Autism and Developmental Disorders*.
- Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*.
- McKnight, P. E., & Najab, J. (2010). Mann-Whitney U test. *The Corsini Encyclopedia of Psychology*.
- Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*.
- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*.
- https://poe.com/
- Placiński, M., Żywiczyński, P. (2023) Modality effect in interactive alignment: Differences between spoken and text-based conversation
- (Ribeiro, D. *Understanding the Wilcoxon test*. Diogo Ribeiro Statistics, n.d.). https://diogoribeiro7.github.io/statistics/data%20analysis/wilcoxon/