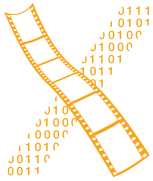

Chair for Computer Science 10 (Media Computing and Human-Computer Interaction)

RWTH AACHEN UNIVERSITY

Investigating the Effects of Response Times in Human-AI Communication on User Perception

Master's Thesis at the
Media Computing Group
Prof. Dr. Jan Borchers
Computer Science Department
RWTH Aachen University

by
Ulyana Lavnikevich

Thesis advisor:
Prof. Dr. Jan Borchers

Second examiner:
Prof. Dr. Johanna Bogon

Registration date: 06.01.2026
Submission date: 06.07.2026

Eidesstattliche Versicherung
Declaration of Academic Integrity

Name, Vorname/Last Name, First Name

Matrikelnummer (freiwillige Angabe)
Student ID Number (optional)

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Arbeit/Bachelorarbeit/ Masterarbeit* mit dem Titel
I hereby declare under penalty of perjury that I have completed the present paper/bachelor's thesis/master's thesis* entitled

selbstständig und ohne unzulässige fremde Hilfe (insbes. akademisches Ghostwriting) erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt; dies umfasst insbesondere auch Software und Dienste zur Sprach-, Text- und Medienproduktion. Ich erkläre, dass für den Fall, dass die Arbeit in unterschiedlichen Formen eingereicht wird (z.B. elektronisch, gedruckt, geplottet, auf einem Datenträger) alle eingereichten Versionen vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.
independently and without unauthorized assistance from third parties (in particular academic ghostwriting**.** I have not used any other sources or aids than those indicated; this includes in particular software and services for language, text, and media production. In the event that the work is submitted in different formats (e.g. electronically, printed, plotted, on a data carrier), I declare that all the submitted versions are fully identical. I have not previously submitted this work, either in the same or a similar form to an examination body.

Ort, Datum/City, Date

Unterschrift/Signature

*Nichtzutreffendes bitte streichen/Please delete as appropriate

Belehrung:
Official Notification:

§ 156 StGB: Falsche Versicherung an Eides Statt
Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.
§ 156 StGB (German Criminal Code): False Unsworn Declarations
Whosoever before a public authority competent to administer unsworn declarations (including Declarations of Academic Integrity) falsely submits such a declaration or falsely testifies while referring to such a declaration shall be liable to imprisonment for a term not exceeding three years or to a fine.
§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt
(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.
(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.
§ 161 StGB (German Criminal Code): False Unsworn Declarations Due to Negligence
(1) If an individual commits one of the offenses listed in §§ 154 to 156 due to negligence, they are liable to imprisonment for a term not exceeding one year or to a fine.
(2) The offender shall be exempt from liability if they correct their false testimony in time. The provisions of § 158 (2) and (3) shall apply accordingly.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:
I have read and understood the above official notification:

Ort, Datum/City, Date

Unterschrift/Signature

Contents

List of Figures and Tables

Abstract

Large language model systems can produce fluent responses that appear plausible even when they are incorrect, making it important to understand which aspects of the interaction users rely on when evaluating AI-generated content. This thesis investigated whether response time influences users' belief in AI answers and whether this effect depends on question complexity. We conducted two online experiments using a simulated AI system with fixed, pre-written questions and answers. Participants decided whether they believed AI answers to simple and complex questions shown after either a short (400 ms) or long (6000 ms) delay. Study 1 (n = 36) used a text-based chat interface, while Study 2 (n = 34) replicated the design in a voice-based interaction. The predicted interaction between response time and complexity on participants' beliefs was not statistically significant in both studies. Study 1 showed the expected descriptive pattern, with higher belief rates after short delays for simple questions and after long delays for complex questions. In Study 2, belief rates showed a more general tendency to be higher after short delays. Additional analyses suggested that latency becomes more relevant when answer correctness is difficult to evaluate. Despite knowing half the answers were wrong, participants believed 64.7% (Study 1) and 68.2% of all answers, suggesting general overtrust in AI-generated content. Overall, the meaning of response times is not fixed, but depends on context, interaction modality, question complexity, and users' ability to evaluate answer correctness.

Überblick

Chatbasierte KI-Systeme sind in der Lage flüssige und plausibel wirkende Antworten auf Fragen zu generieren, selbst wenn diese inhaltlich falsch sind. Daher ist es wichtig zu untersuchen, welche Aspekte der Interaktion von Nutzenden bei der Bewertung KI-generierter Inhalte herangezogen werden. In dieser Masterarbeit wurde untersucht, ob die Antwortzeiten der KI einen Einfluss auf die Glaubwürdigkeit der Antworten haben. Des Weiteren wurde beobachtet, ob diese Glaubwürdigkeit zusätzlich von der Komplexität der Frage abhängig ist. Zu diesem Zweck wurden zwei Online-Studien mit einem simulierten KI-System durchgeführt, wobei die Fragen und Antworten vorgegeben waren. Die Teilnehmenden sollten angeben, ob sie die Antworten auf einfache bzw. komplexe Fragen glauben und wie sicher sie sich dabei sind. Die Antworten wurden vom System entweder nach einer kurzen (400 ms) oder einer langen (6000 ms) Antwortzeit präsentiert. In Studie 1 ($n = 36$) wurde eine textbasierte Chat-Oberfläche genutzt. Während der Studie 2 ($n = 34$) wurde dasselbe Design in einer sprachbasierten Form repliziert. Die erwartete Interaktion zwischen Antwortzeit und Fragekomplexität erwies sich in beiden Studien als statistisch nicht signifikant. In Studie 1 zeigte sich deskriptiv das erwartete Muster, dass einfachen Fragen nach kurzer Antwortzeit und komplexen Fragen nach langer Antwortzeit mehr Glauben geschenkt wurde. In Studie 2 zeigte sich hingegen eine allgemeine Tendenz zu höheren Glaubensraten bei kurzen Antwortzeiten. Die weitere Auswertung deutet darauf hin, dass die Antwortzeit relevanter wird, wenn die Richtigkeit einer Antwort schwerer zu beurteilen ist. Obwohl die Teilnehmenden darüber informiert waren, dass die Hälfte der Antworten falsch war, glaubten sie im Durchschnitt 64,7% der Antworten in Studie 1 und 68,2% in Studie 2. Dies deutet auf ein mögliches Problem des Übervertrauens in KI-generierte Inhalte hin. Insgesamt zeigt diese Masterarbeit, dass der Einfluss der Antwortzeiten vom Kontext, der Interaktionsmodalität, der Komplexität der Frage und der Fähigkeit der Nutzenden, die Richtigkeit einer Antwort zu beurteilen, abhängig ist.

Acknowledgments

First of all, I would like to thank Prof. Dr. Jan Borchers and Prof. Dr. Johanna Bogon for examining this thesis.

Secondly, I am especially grateful to my supervisors, Paul Preuschoff and Prof. Dr. Johanna Bogon for their time, helpful feedback, and guidance throughout the process. I appreciate the supportive and comfortable working environment they provided, as well as their engagement in this project.

I would also like to thank Jan Erik Karuc for his technical support, especially with the deployment.

Furthermore, a big thank you to everyone who participated in my user studies. Your time and effort made this work possible.

Lastly, I would like to thank my friends for their emotional support. In particular, I want to thank my partner Florian for his constant support, advice, and motivation.

Conventions

Throughout this thesis we use the following conventions:

- The thesis is written in British English.
- The first person is written in plural form.
- Unidentified third persons are described in female form.

Where appropriate, paragraphs are summarized by one or two sentences that are positioned at the margin of the page.

This is a summary of a paragraph.

Chapter 1

Introduction

LLMs have grown extremely fast and, unlike earlier chatbots, support open-ended, follow-up-driven interactions.

Large language model (LLM) chat systems, such as ChatGPT, Gemini, Claude, or Copilot, allow users to interact in natural language rather than being limited to a fixed set of questions and domains, as earlier rule-based chatbots were [Walker et al., 2024]. They generate fluent and conversational responses in seconds. ChatGPT reached approximately 100 million monthly active users within two months of its launch [Huschens et al., 2023; Trust et al., 2023], illustrating how quickly such systems entered everyday use. People increasingly rely on such systems for information gathering [Xu et al., 2023], learning [Prananta et al., 2023], and decision support [Liu et al., 2023].

Expectations and interaction design can shape how users evaluate AI.

However, LLMs can generate responses that seem plausible and convincing, yet may contain inaccuracies or fabricated information [Tamanna et al., 2024]. This is especially relevant when users lack prior knowledge or cannot easily judge whether an Artificial Intelligence (AI) output is correct. Prior work suggests that such judgements can be affected by expectations and interaction design. For example, von Felten et al. showed that simply presenting a product as "AI-assisted" can increase users' expectations of its performance, even when the product does not provide corresponding functional benefits [von Felten et al., 2026]. Modern LLMs are designed to resemble a real conversation partner through features such as natural language, typing indicators, or emojis.

Many AI systems already include voice modes, giving users the ability to listen to generated responses or engage in direct spoken interaction. These voice-based AI agents are now starting to copy oral speech patterns, such as pausing, saying "mm-hm", or adapting the user's speaking style [Spillner et al., 2025].

Human-like interaction cues may influence trust and perceived accuracy in AI responses.

Including these features can make the system appear more human-like, which has been associated with enhanced trust [Cohn et al., 2024] and increased perceived response accuracy [Shi et al., 2026]. Consequently, users' belief in a response may depend not only on the content of the answer, but also on interaction cues utilised by the system [Jung et al., 2024]. One such interaction cue is response time. Before users read or hear an AI-generated answer, they already experience how long the system took to respond. This thesis investigates whether this response time influences how much people believe an AI's answers, and whether this depends on how complex the underlying question is.

In human communication, pauses are meaningful and shape how a conversation partner is perceived.

Response time is often treated as a technical performance variable, but it can also carry meaning in conversational AI. In human-to-human interactions, pauses are not simply silent gaps. They influence how an answer is interpreted and how the speaker is perceived. Short speech pauses are part of smooth conversation, whereas longer pauses may signal hesitation, uncertainty, reflection, or disagreement [Kendrick and Torreira, 2015; Matzinger et al., 2023; Stivers et al., 2009]. Therefore, the same answer may be understood differently depending on when it is delivered.

Findings on response latency in chat-based systems are mixed.

If people extend conversational expectations to AI systems in this way, response time could also function as a timing cue in Human-AI interaction. However, the existing findings on this question are mixed. An instant, short response can be read as efficient, but it can also suggest that the system did not properly process the request. In contrast, a long response can be interpreted as a signal of careful processing, yet it can also be viewed as a technical failure [Gnewuch et al., 2018; Holtgraves et al., 2007].

Prior work suggests that the meaning of response time can be affected by the difficulty of the question and the expectations placed on the responder [Efendić et al., 2020]. In tasks that were assumed to be difficult and time-consuming for humans, longer human responses were evaluated more positively. In contrast, long algorithmic responses were evaluated negatively because fast computation was expected. However, when task difficulty was manipulated directly, the difference between human and algorithmic responses disappeared. This suggests that complexity may shape whether a delayed AI answer feels appropriate.

Task complexity influences whether response time seems appropriate.

Therefore, this thesis investigates whether response latency influences users' belief in AI-generated answers and whether this effect depends on question complexity. We conducted two online experimental studies in which participants evaluated AI answers to simple and complex questions after either a short or a long response latency, judging whether they believed each answer and how confident they were in that judgement. In Study 1, participants interacted with a text-based chat interface, while Study 2 extended this design to a voice-based interaction in which the AI answers were presented using a synthetic voice.

This thesis investigates the interaction between latency and complexity on belief, across text- and voice-based studies.

In both studies, the hypothesised interaction between response latency and question complexity was not statistically confirmed. However, in the text-based study, belief rates descriptively followed the expected pattern. In the voice-based study, this pattern did not replicate. Instead, participants generally believed answers more often after short delays. Both studies showed that participants generally tended to believe AI-generated answers, despite being informed that half of the answers were incorrect. Additional analyses suggested that response latency, together with complexity, may matter most when participants could not otherwise judge whether an answer was correct. Additionally, response latency also changed how long participants took to decide.

The expected interaction was not statistically confirmed, but latency had a descriptive influence on belief and affected decision time.

This thesis makes three main contributions:

- It empirically investigates how response latency and question complexity together may shape users' belief in AI-generated answers.
- It examines whether latency effects become more visible under uncertainty, when users have difficulty evaluating the correctness of an answer.
- It compares text-based and voice-based interaction, investigating whether the meaning of response time changes with modality and question complexity.

Together, this work extends research on response time in conversational AI by connecting research on pauses in human conversation, social responses to computers, and trust in AI, and suggests that response latency functions as a timing cue that can shape users' evaluations of generative AI.

After this introduction, **Chapter 2** provides an overview of research related to this thesis. **Chapter 3** and **Chapter 4** present the two experimental studies, followed by a general discussion of the findings, implications, and limitations in **Chapter 5**. **Chapter 6** concludes the thesis with a summary and directions for future work.

Chapter 2

Related Work

This chapter looks at prior research on response time as a cue in interactions with AI systems. It first discusses how pauses are interpreted in Human-Human Interaction and why timing may also matter in Human-Computer Interaction (HCI). Next, it reviews previous findings on response latency in HCI and conversational systems. The chapter also explores how task complexity, effort heuristic, and trust in AI can affect user perceptions. Finally, the chapter points out the research gap that this thesis aims to address.

2.1 Pauses in Human Communication

In human conversation, people usually take turns quickly, making longer pauses noticeable.

Pause in conversation refers to the "delay between one speaker's utterance and the other's response" [Maslych et al., 2025]. Short pauses are a normal part of human communication, helping to regulate who speaks next and when [Jaffe et al., 2001; Sacks et al., 1974]. Generally, turn-taking between conversation partners is impressively fast. In many languages, speakers typically switch from one turn to the next in 200-300 ms [Dingemanse and Liesenfeld, 2022; Stivers et al., 2009]. In contrast, longer pauses are more noticeable and can make the interaction seem less smooth and break the flow [Matzinger et al., 2023].

The timing of an answer can signal whether it is expected, hesitant, or problematic.

Conversation analysis treats a conversation as a sequence of paired actions, known as adjacency pairs. The action of the first speaker, such as a question, a request, or an offer, produces the *expectation* that the corresponding action of the second speaker will follow, such as a response, an agreement, or an acceptance [Schegloff and Sacks, 1973]. Within such pairs, the timing of the second action differs depending on whether the response aligns with what is expected, called a preferred response (e.g., accepting an offer), or not, called a dispreferred response (e.g., rejecting an offer) [Matzinger et al., 2023]. Stivers et al. [2009] found that confirmations, a type of preferred response, were delivered on average 100-500 ms faster than disconfirmations. Kendrick and Torreira [2015] found that responses with a gap of 700 ms or longer were more likely to be dispreferred than preferred. This indicates that pauses can already influence how a response is interpreted even before it is given, whether it is expected, problematic, or sounds hesitant.

The social meaning of a pause depends on the relationship between speakers.

The social meaning of a pause depends on the relationship between speakers. Templeton et al. defined long gaps as longer than two seconds, based on this typical conversational gap of 200-300 ms [Templeton et al., 2023]. They found that such long gaps were perceived as awkward and negative between strangers. In contrast, friends associated long pauses with a moment of increased connection and bonding [Templeton et al., 2023].

Pauses can affect perceived competence, but context is important for interpretation.

Matzinger et al. [2023] examined how pause length affects perceived competence in both native and non-native speakers. Longer pauses were linked to lower ratings of knowledge and confidence for all participants. For non-native speakers, long pauses were sometimes seen as a sign of language processing rather than a lack of knowledge.

This suggests that the same pause duration can have different meanings depending on the context in which it occurs. The next question is whether these interpretations of pause also apply to Human-Computer Interaction.

2.2 Social Responses to Computers

Human-like cues can trigger social responses to computers and increase perceived trustworthiness.

In many interactions with computers, users do not have access to the same social cues that are available in face-to-face communication, such as body language, facial expressions, and gaze. Despite this reduced channel, people still tend to respond to computers in a social way. According to the Computers Are Social Actors paradigm, this happens because people apply social rules learned in Human-Human Interaction to computers when the technology mimics human characteristics, e.g., natural human language or voice [Reeves and Nass, 1996]. Furthermore, Reeves and Nass [1996] found that people may do it automatically, even when they are aware that they are interacting with a machine. Simply the fact that a system has a name can already trigger people to interact with it socially [Nass et al., 1994]. A related concept is anthropomorphism, which describes the tendency to attribute human qualities to non-human entities [Złotowski et al., 2015]. A clear example of anthropomorphism is how people interact with personal assistants like Apple's Siri. Users often treat Siri as if it has emotions and may thank the system after receiving feedback [Bhat, 2025]. Gestures and voice are also strong triggers of anthropomorphism because they are central parts of human communication [Salem et al., 2011]. Generally, anthropomorphism is associated with a more positive perception of the competence and trustworthiness of autonomous systems [Haresamudram et al., 2024].

In conversational AI, response latency may function as a chronemic cue.

AI chatbots and voice assistants are particularly important in this context, as they are designed to generate human-like text or speech, making such social responses even more likely. According to Social Information Processing theory [Walther, 1992], people form impressions of a conversation partner not only through what is said and how it is phrased, but also through the timing of a response, a so-called chronemic cue [Kim et al., 2026]. Response latency is one such time-related signal and may function similarly to pauses in human conversation because it provides information that users can interpret in a social way [Kim et al., 2026; Walther, 1992].

2.3 Response Latency

2.3.1 Delays in HCI

In traditional HCI, longer system response times are usually evaluated more negatively.

Response latency, or system response time, refers to the time interval between a user's input and the system's response [Shneiderman, 1984]. In HCI, there are three thresholds for how delays are experienced [Miller, 1968; Nielsen, 1993]. Delays of 100 ms are perceived as direct cause and effect, e.g., between a mouse click and a button's visual response. Delays of about one second are tolerated without disrupting the user's thought flow. And delays of up to 10 seconds are tolerated as the maximum time to complete one step of a task before users lose attention. For tasks such as scrolling or dragging, delays of less than 200 ms are needed to maintain a sense of control [Dabrowski and Munson, 2011]. Longer system responses can reduce satisfaction and increase stress and frustration [Schleifer and Amick III, 1989]. Google conducted an experiment by increasing the delay for search results from 100 to 400 milliseconds [Schurman and Brutlag, 2009]. This increase resulted in a significant decrease in the number of searches. In these systems, the relationship between delays and user evaluation is linear: the longer the delay, the worse the reaction.

2.3.2 Delays in Conversational Systems

In conversational systems, delayed responses can increase humanness and social presence.

However, in conversational systems, e.g., chat-based systems, this relationship does not always hold because users may treat such systems more as conversational partners than as systems for direct manipulation. Gnewuch et al. [2018] conducted an online experiment with different groups in a customer service setting. They created a chatbot to compare an instant delay with a dynamically calculated delay. The instant delay was about 200-400 ms, which is similar to the short network delay due to the limits of Internet data transmission. To calculate the longer delay, they used the Flesch-Kincaid formula Kincaid et al. [1975], which measures sentence complexity based on word and

syllable counts. Participants who interacted with the dynamically delayed chatbot rated it as more human-like and reported greater social presence and satisfaction [Gnewuch et al., 2018].

Moderate AI response delays may make the system seem more thoughtful and useful.

Tan et al. [2026] reported a similar pattern in the context of Human-AI Interaction. They conducted a between-subjects study with a GPT-based chat interface to test three response times: 2, 9, and 20 seconds. Delay was defined as the time interval between the participant's prompt submission and the first generated token. The study also varied the task type, comparing creation tasks with advice tasks. Participants' interaction behaviour, such as how often they submitted or edited prompts, was unaffected by latency but mostly by the type of task. However, the delay did affect how participants judged the AI's responses. When the AI replied after 2 seconds, its answers seemed less thoughtful and less useful than those after 9 or 20 seconds. The 9-second delay was rated as the most useful. At 20 seconds, some participants started to doubt the system's reliability [Tan et al., 2026].

Persuasiveness may be highest when chatbot delays are neither too short nor too long.

Moon [1999] observed a similar trend in a chatbot-based context. They investigated which factors influence persuasiveness in a lab experiment. Chatbot's messages were most persuasive when the response delay was neither too short nor too long: a delay of 5 to 10 seconds resulted in greater persuasiveness than a short delay (0 to 1 second) or a long delay (13 to 18 seconds) [Moon, 1999].

Some findings suggest that users still prefer faster chatbot responses.

Other findings align more closely with the expectation that chatbots should respond quickly. Holtgraves et al. [2007] compared a chatbot implemented to respond after either a 1-second or a 10-second delay. Participants rated the bot as more extroverted and more reliable when it responded quickly than when it responded slowly. The ratings of emotional stability and friendliness did not differ between conditions [Holtgraves et al., 2007].

A similar preference for shorter response times has been found among younger adults. Wang and Lo [2025] tested a chatbot as a virtual companion, comparing younger (18-23) and older (56-81) adults with either fast replies within

0.1 s or delayed ones ranging from 10 to 60 s. Participants interacted with the chatbot daily over five days and were asked to report on its social presence, satisfaction, and intention to use. Younger adults reported higher satisfaction and engagement with instant responses, while older adults felt more comfortable with delayed responses [Wang and Lo, 2025].

2.4 Complexity and the Effort Heuristic

The same response time can signal effort in humans but inefficiency in algorithms.

The impact of a delay may depend on what users believe the system is doing during the wait. One relevant factor is the perceived complexity of the task being performed. Efendić et al. [2020] investigated how response times influence trust in human and algorithmic predictions. They refer to the term *effort heuristic* by Kruger et al., according to which people often judge an outcome more positively when they believe that more time and effort have been invested [Efendić et al., 2020]. In seven studies, participants read brief scenarios describing predictions, such as forecasts of future sales or a student's academic success. These predictions were described as having been generated either quickly or slowly, and either by a human or by an algorithm. The main finding was that the same response time led to opposite judgements: a slowly generated prediction was rated as more accurate when it came from a human, but as less accurate when it came from an algorithm.

Whether response time is evaluated positively depends on perceived task complexity and source.

The researchers explained this through perceived task complexity. For humans, prediction was perceived as a difficult task, so a slow response was read as a sign of careful thinking and higher quality. In contrast, for algorithms, the task was viewed as easy, so a slow response conflicted with expectations and was not associated with higher prediction quality. When the authors directly manipulated task difficulty, response time influenced perceived accuracy in a similar way for both humans and algorithms [Efendić et al., 2020].

A more recent study examined a similar issue within the context of generative AI. Kim et al. [2025] conducted sev-

eral studies on how response latency influenced users' evaluations of ChatGPT recommendations. Participants watched pre-recorded videos or viewed screenshots of an interaction with ChatGPT. In the first study, the AI provided travel recommendations within 4 or 30 seconds. Kim et al. [2025] found that faster responses resulted in higher satisfaction and perceived usefulness. Another study varied response latencies (2 or 8 seconds with different levels of information complexity, defined as the number of recommended destinations (4 or 30). Results indicated that faster responses were preferred for simple recommendations, but when the information was complex, the advantage of a fast response disappeared. In the qualitative interviews, participants described longer response times as a sign of effort and "thinking" [Kim et al., 2025].

The advantage of fast AI responses may disappear when the information is more complex.

These findings suggest that the same response latency can be interpreted differently, depending on whether the source is human or an algorithm and on how complex or difficult the task appears. Question complexity may then serve as a moderator of response latency effects.

2.5 Trust and Overtrust in AI

Trust in computer systems is a central factor in how users engage with them, especially when these systems offer decision support [Seymour and Van Kleek, 2021]. Prior HCI work describes trust as both confidence in the system and willingness to use the information it provides [Madsen and Gregor, 2000; Seymour and Van Kleek, 2021].

AI systems can generate fluent and convincing responses, but these are not always correct, as LLMs may produce errors, biased outputs, or hallucinated information [Cohn et al., 2024; Tamanna et al., 2024]. Overtrust in AI is problematic because it can spread misinformation and lead users to follow harmful or misleading advice [Bhat, 2025; Klingbeil et al., 2024]. In a behavioural experiment with financial incentives, Klingbeil et al. [2024] found that participants often followed advice labelled as AI ad-

Overtrust is problematic because users may rely on AI advice even when it is incorrect.

vice even when it conflicted with available information and their own assessment.

When accuracy is hard to verify, users rely on source credibility and interaction style.

When people cannot verify the accuracy of the information, they often rely on the perceived credibility of its source, e.g., how the information is presented or the system's interaction style [Jung et al., 2024]. Users may perceive AI systems as both technological tools and social partners. Interface features such as typing indicators or gradual text presentation can make an AI response appear human-like, thereby increasing perceived presence and affecting users' trust [Bhat, 2025].

Voice can increase anthropomorphism and perceived accuracy in interactions with LLMs.

Cohn et al. [2024] noted that previous work has linked anthropomorphism to trust in conversational agents, avatars, and robot systems. Moreover, Cohn et al. [2024] found in their study that adding a voice to an LLM increased anthropomorphism ratings of the system and perceived information accuracy compared to a text-only condition. This aligns with Seymour and Van Kleek's argument that users' relationship with voice assistants falls "somewhere between the clear-cut categories of human and machine", unlike that with other smart technologies. In addition, research shows that special vocal qualities such as naturalness, rhythm, intonation, and accent can shape trust differently and more strongly influence perceived trustworthiness than voice gender, which plays a secondary role [Danielescu et al., 2023].

2.6 Summary and Research Gap

Previous research shows that the effect of response times on user interaction with chat-based systems remains unclear. On the one hand, users expect chatbots to respond quickly because such systems are assumed to process input instantly [Efendić et al., 2020]. On the other hand, instant responses can also make a system appear suspicious and less human-like [Gnewuch et al., 2018].

Response times may also function as a social cue, mimicking human behaviour [Kim et al., 2026]. Applying anthro-

pomorphistic features to AI systems, such as delay, presentation style, usage of human language and voice, can shape users' trust in systems and its outputs [Cohn et al., 2024]. As the correctness of AI-generated content cannot always be verified [Jung et al., 2024], such cues may convince them directly.

At the same time, the effort heuristic suggests that response time depends on the perceived complexity of the task [Efendić et al., 2020]. This indicates that complexity is a factor influencing whether a delay appears appropriate.

However, most evidence on the impact of response latency and complexity on user evaluations comes from studies in which delays were described in text form [Efendić et al., 2020] or shown in pre-recorded videos [Kim et al., 2025; Zhang et al., 2024]. Kim et al. suggested that future research should involve real chatbot interactions to improve ecological validity. Some studies that used real interaction with LLMs varied latency together with other factors, such as task type [Tan et al., 2026]. Other studies took a similar approach in earlier chatbot systems, combining latency with message complexity [Gnewuch et al., 2018]. In both cases, researchers measured perceived humanness, usefulness, and satisfaction, rather than whether people believed the response was accurate. It remains unclear whether response latency and complexity of a query to the AI interact with each other and influence users' perception of credibility under real, experienced delays.

Chapter 3

User Study 1: Text-Based Interaction

In the previous chapter, we already noted that response latency in Human-AI interactions can have social meaning and act as a timing cue. It may depend on the complexity of the question being asked. However, it remains unclear whether these two factors interact and influence how plausible an AI-generated answer appears in the context of actually experienced delays. Study 1 investigated this question in a web-based experiment in which participants evaluated responses from a simulated AI system, preceded either by a short or a long response latency, while the complexity of the questions was varied. In this chapter, we outline our research question and hypotheses, describe the study design and procedure for the first online experiment, and present the results and discuss our findings.

3.1 Research Question & Hypotheses

Study 1 explored how response latency shapes users' perception of AI-generated answers and whether it interacts with question complexity. In particular, we investigated whether *short* and *long* response delays are interpreted differently for *simple* and *complex* questions.

Study 1 investigated response latency and question complexity effect in a text-based AI interaction.

Hence, we defined the following research question:

RQ Does response latency of an AI system influence users' belief judgements of its answers, and does this effect differ depending on question complexity?

In the following hypotheses, question complexity refers to the complexity of the question asked before the AI-generated answer. Participants then judged whether they believed the AI answer.

We formulated the following hypotheses:

H1 Response latency interacts with question complexity in influencing users' belief judgements and confidence ratings of AI-generated answers.

H1.1 For simple questions, the belief rate and confidence rating after short latency are higher than the belief rate after long latency.

H1.2 For complex questions, the belief rate and confidence rating after long latency are higher than after short latency.

H1.1 is based on the general expectation of users that computational systems respond fast [Shneiderman, 1984]. Chatbots that reply after a longer delay have been rated as less reliable than those that reply quickly [Holtgraves et al., 2007]. When a response to a straightforward question takes longer than expected, it can signal low competence and malfunction, reducing user trust and satisfaction.

H1.2 builds on the *effort heuristic*, according to which people judge an outcome more positively when they believe that more time and effort have been invested [Efendić et al., 2020; Kruger et al., 2004]. For complex tasks, a longer delay can be associated with deep processing of a difficult request [Kim et al., 2025]. The long pause here may increase how plausible the answer feels.

3.2 Method

This section describes how Study 1 was designed and conducted. It covers the experimental design, stimulus development, procedure, counterbalancing, participant sample, technical setup, and the measures used to assess participants' evaluations of the generated responses and the system's behaviour.

3.2.1 Study Design

Study 1 used a 2x2 within-subject design with latency and complexity as independent variables.

The study used a 2x2 within-subject design with two independent variables: response latency and question complexity. The response latency was manipulated by showing the AI response after a short or long delay. The complexity of the question was manipulated by asking the simulated AI agent simple or complex questions. Each participant saw all four conditions, allowing comparisons within-person. Table 3.1 provides an overview of the resulting four experimental conditions.

within-subject design

	Question complexity	
Response latency	**Simple**	**Complex**
Short (400 ms)	Short-Simple	Short-Complex
Long (6000 ms)	Long-Simple	Long-Complex

Table 3.1: Experimental conditions in Study 1.

Belief judgements and confidence ratings were the primary outcome measures.

Our primary dependent variables were belief judgement and confidence rating. For each item, participants first reported whether they believed the AI's answer. Afterwards, they rated their confidence in that judgement. Additionally, reaction time and time to confidence rating were collected as exploratory measures (see Section 3.2.7 for more details).

3.2.2 Stimuli & Materials

The study used a simulated chat interface with pre-written answers.

The study simulated a chat-based AI interaction designed after commonly used conversational AI systems such as ChatGPT, Claude, or Copilot, in which the user's request is displayed on the right side and the system's response on the left. However, we did not use a live language model. All questions and answers were pre-written and fixed because LLMs produce different answers to the same question. This ensured that answer content, length, and phrasing were identical for all participants.

This approach follows similar methodological choices in related work. Tan et al., for instance, designed their front-end to "mirror" common LLM assistants such as ChatGPT, including a chat-bubble layout in which user messages were displayed separately from AI response [Tan et al., 2026].

Each trial consisted of a user question, a manipulated delay, and a fixed AI response.

The core unit of the study was a trial consisting of a question coming from the user, followed by a delay, and then an answer provided by the AI agent. With this setup, we wanted to imitate an interaction between a user and an AI chatbot. We were interested in if the complexity of the question primes certain expectations during the waiting period, and if these expectations interact with the length of the delay.

Response Latency

The latency manipulation contrasted an instant 400 ms delay with a clearly noticeable 6000 ms delay.

The short delay (400 ms) and a long delay (6000 ms) were chosen to create a clear contrast between conditions, based on response latencies reported in prior work on chatbots and conversational timing (see Section 2.3). The short delay is comparable to the response times of chatbots used in practice, which typically range between 200 and 400 ms due to network transmission [Gnewuch et al., 2018], and is significantly below the two-second threshold identified by Shneiderman as acceptable for system responses. It is also close to the average gap observed between turns in human conversation [Stivers et al., 2009]. The long delay of 6000 ms is below the tolerable waiting times reported in

prior work, from around 8 seconds in speech-based applications to the 10 second upper limit identified by Nielsen for holding users' attention. It also fits into the range of waiting times used in recent studies on AI response latency, in which medium to long delays were rated more positively than short ones [Tan et al., 2026].

Question Complexity

As mentioned in Section 2, previous research indicated that the meaning of the waiting time can shift depending on how complex the task appears. Therefore, we decided to manipulate the question complexity in our experiment. We differentiated between simple and complex questions.

Question complexity was defined by cognitive structure rather than subjective difficulty.

Simple questions often relate to fact- or memory-based information, involve one relation and can be answered with a short sentence or yes/no response [Yani and Krisnadhi, 2021]. In contrast, a complex question involves multiple relations or reasoning steps [Etezadi and Shamsfard, 2023]. Moreover, complex questions typically start with "*Why*", "*How*", while simple queries begin with "*What*", "*When*", or "*Which*" [Etezadi and Shamsfard, 2023]. Therefore, in our simple questions, we asked for definitions, specific facts or terms, and in the complex items, we asked for reasoning or explanations. We aimed to define complexity in terms of cognitive processing and structure, rather than perceived difficulty.

The question set was designed to make answer correctness difficult to judge from prior knowledge alone.

We developed the questions for the study through a multistep process. Our initial idea was to look first at the already established question datasets, e.g., Natural Questions, TriviaQA, AI2 Arc, and SciQ, which we used as a useful starting point for further question generation. However, many questions from these databases would be too familiar. With these questions, participants would easily verify whether the AI is right based on their prior knowledge, and the delay may become irrelevant. To give more space for possible delay effects, we needed questions where participants, in the best case, are uncertain, and therefore have to rely on other systems' cues, such as how long the AI took to re-

spond. We iteratively generated and refined the questions using ChatGPT and Claude LLMs. The goal was to find such questions where prior knowledge is minimised while maintaining ecological validity by selecting topics which users might themselves query in a real-world scenario.

Answer Development and Plausibility

Answers were designed to be plausible, regardless of their factual correctness.

Each question was paired with a single answer generated using ChatGPT and Claude and manually reviewed. All answers were pre-written and pre-defined as correct or incorrect. We aimed to introduce more uncertainty about the accuracy of the answers in order to motivate participants for an active evaluation for each trial. Half of the simple and half of the complex answers were factually correct, while the remaining half contained incorrect information. Correct and incorrect answers were written to sound plausible. For this work, we used the definition of "plausibility" by Agarwal et al., who define it as "the degree to which a response appears logical and coherent within the context of general world knowledge" [Agarwal et al., 2024]. For example, in our question-answer dataset, one plausible but incorrect answer to a simple question about the planet with the shortest day in the solar system was that Saturn is the answer, although the correct answer is Jupiter. An incorrect response to one complex question explained that metal feels colder than wood because of the low starting internal temperature, whereas the actual reason is that metal conducts heat away from the skin more rapidly.

Answer length was controlled within complexity levels to reduce possible confounds.

We controlled answer length to reduce its potential influence and keep reading time more consistent. Answers to complex questions were longer because they required more explanation than answers to simple questions. In our stimulus set, complex answers consisted of two to three sentences with 24-34 words, whereas responses to simple questions included one sentence and were limited to 12-17 words. This approach was inspired by Zhang et al. [2024], who similarly operationalised longer chatbot responses as approximately twice the length of shorter responses.

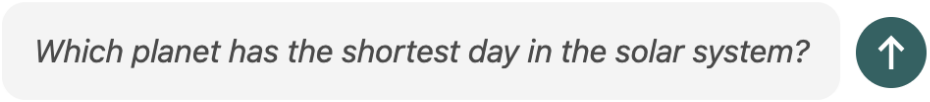

Figure 3.2: The screen shown at the start of each trial: a user question is displayed alongside a "send" button.

3.2.3 Procedure

The online study began with general information and informed consent.

The user study was conducted entirely online and without assistance. Each participant entered the website and first saw a welcome page, which provided general information about the experiment and the technical requirements. Participants were then directed to an informed consent form. To proceed, they had to provide consent by selecting a checkbox.

Participants completed practice trials and were informed about a 50/50 answer accuracy.

After that, participants had the opportunity to complete a tutorial to get familiar with the study procedure and user interface. There were three practice trials: the first demonstrated the basic trial flow, while the other two introduced the countdown timer that each actual trial has. Participants were informed that half of the answers they would see were correct and half incorrect, and were to make their decisions as quickly as possible based on their immediate impression.

Participants actively sent each question to "AI" before experiencing the delay.

Each trial followed the same structure. After the page loaded, a simple or a complex question was displayed, as shown in Figure 3.2. A send button to the right of the "user's" question was activated after 1500 ms to prevent the immediate sending of the question. This short block of sending was also included to motivate reading the text of the question first. Participants had to actively click the "send" button to submit their question to the "AI" and receive an answer from it. We did this to give them a sense of control over the interaction and make the interaction seem more natural. As shown in Figure 3.3, after sending, a typing indicator was displayed for the duration of the assigned delay condition (400 ms or 6000 ms).

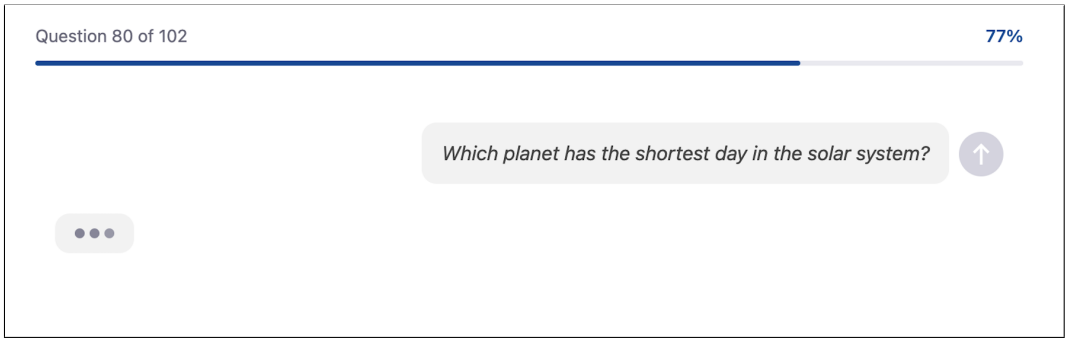

Figure 3.3: After the question has been sent, the "send" button is deactivated, and an animated typing indicator is displayed for the duration of the assigned delay condition.

The typing indicator was used to make the waiting period feel like answer generation.

A common way to visually represent typing and waiting for a reply in a chat setting is by using three moving animated dots [Gnewuch et al., 2018]. Because these indicators suggest an ongoing process, participants might associate the delay more with the effort than with inefficiency [Gnewuch et al., 2018]. By adding such animation during short or long delays, we aimed to signal answer generation and to provide a more natural interaction.

Participants judged each answer under time pressure based on their immediate impression.

Following the delay, the full AI-generated answer to the question was shown to participants at once instead of being displayed word by word. This ensured that delay manipulation was related only to the waiting time before the answer was displayed and was not confounded with the speed of answer presentation. Once the answer appeared, they had 13 seconds (a value derived from pilot studies, see Section 3.3) to decide if they believed it by clicking "Yes" or "No" to the statement "I believe this answer is correct". Figure 3.4 shows this decision moment, including the countdown progress bar that visualised the remaining time.

The content was blurred so participants could not reread it before rating confidence.

If a participant has made a decision, the text of the question asked, and the corresponding AI answer were blurred to prevent rereading, so the focus remains on the immediate reaction, illustrated in Figure 3.5. Participants then rated the confidence in their judgement on a 7-point Likert scale and moved to the next trial by clicking "Next Question". If no decision was made within the time limit, the current

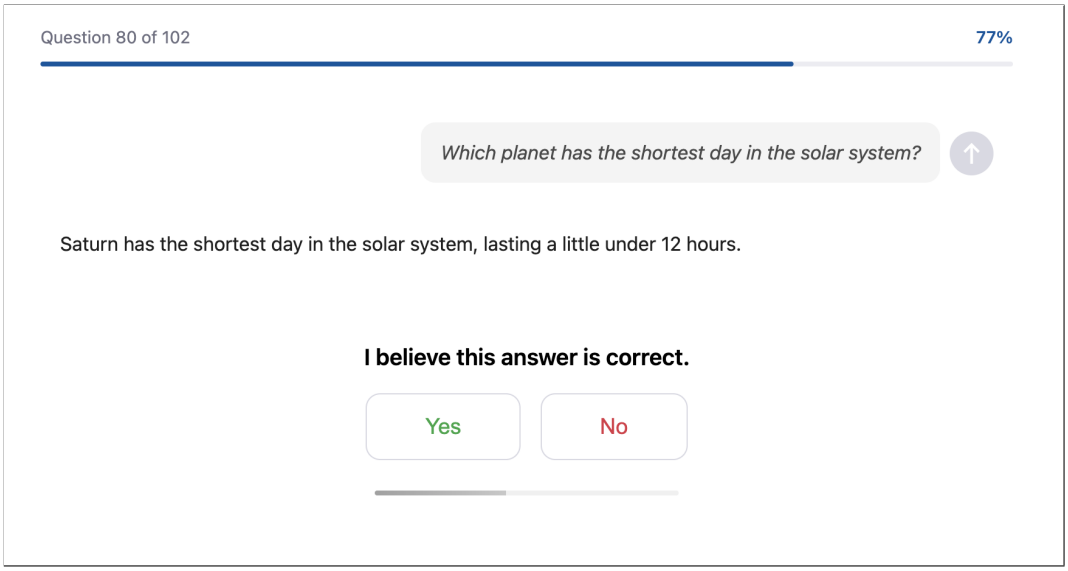

Figure 3.4: After the delay, the AI-generated answer appears below the question. Participants indicate their belief decision by clicking "Yes" or "No", with a countdown bar visualising the remaining response time.

trial was marked as a timeout, and participants were asked to proceed with the next trial without giving a belief and confidence decision.

Attention checks helped identify random responses.

The main study had 102 trials in total, two of which served as attention checks. They followed the same format as regular trials. However, participants were explicitly told which button to select. Using the attention checks, we were able to verify which participants had actually read the content and which had answered randomly.

The study concluded with questionnaires, a reading-speed task, and a debriefing.

After all trials were completed, participants were asked to fill out a demographic questionnaire, followed by a perception questionnaire. They also completed a short reading speed task to collect how quickly they read. Finally, participants were directed to a debriefing page, where they were informed about the true purpose of the study, the two manipulations we did, and clarifying that the AI system had been simulated and all answers were pre-written and fixed. The study took approximately 45-60 minutes to complete.

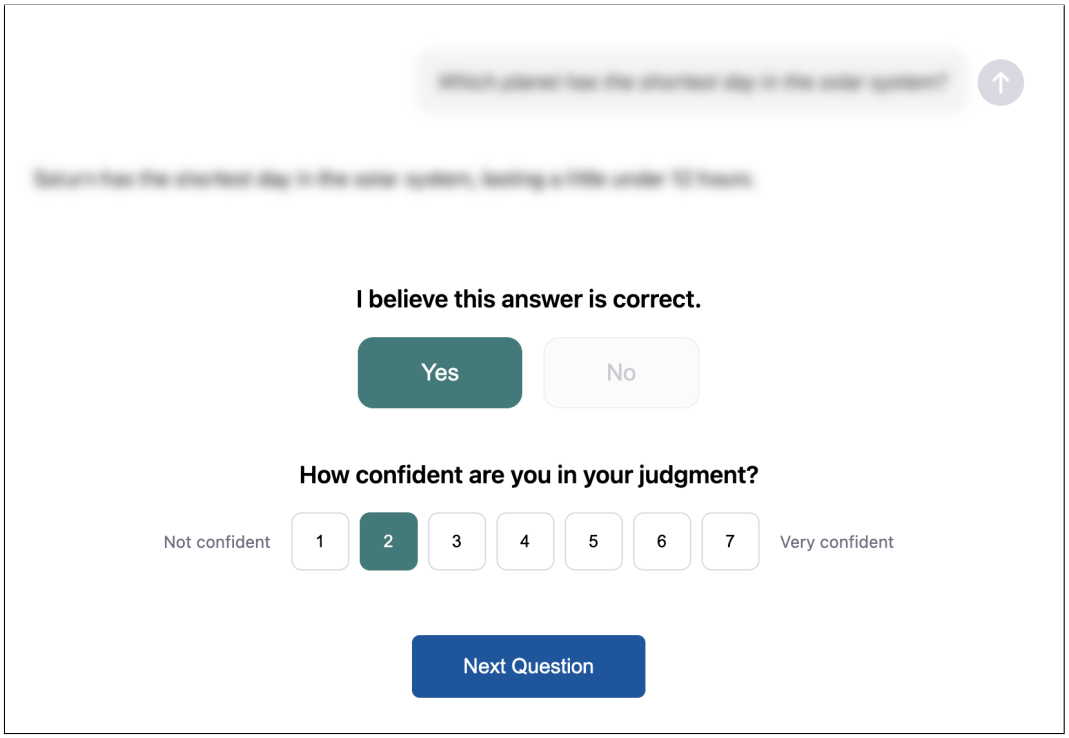

Figure 3.5: After a belief decision has been made, the question and answer are blurred. The example shows a "Yes" decision together with a confidence rating of 2 on the 7-point Likert Scale.

3.2.4 Counterbalancing & Randomisation

Because Study 1 used a within-subjects design, each participant saw trials from all four experimental conditions: simple-short, simple-long, complex-short, and complex-long. Each question pair was presented only once to each participant. To ensure that each pair was not always associated with the same response latency, the assignment of trials to delay conditions was counterbalanced.

Delay assignment was counterbalanced so that items were not tied to a single latency condition.

We divided the 50 simple and 50 complex items into equally sized subsets. The assignment to these two subsets was randomised. The counterbalancing logic was based on the participant ID number. For participants with an even number, the first subset of simple (complex) items was presented with a short delay, and the second subset with a long delay. The odd number group received the reversed

assignment. By changing the delay assignment between participants, we reduced the risk that item-related characteristics, such as topic, answer length and complexity, were confounded with latency.

Trials were divided into blocks and randomised to reduce order effects.

100 trials were divided into 25 blocks of four, each containing one trial per experimental condition. Block order and trial order within each block were randomised. This was done to reduce potential carry-over and order effects. Two additional attention-check trials were inserted at fixed positions (trial 33 and 67).

3.2.5 Participants

Participants were recruited using personal networks and required no specific domain knowledge.

We recruited 41 participants through convenience and snowball sampling using personal networks. Participation was voluntary and uncompensated. No specific domain knowledge was required because the study focused on the general perception of AI responses. Participants needed sufficient English proficiency to understand the texts in the trials. One of the requirements was to use a desktop or laptop to minimise differences in user experience between participants.

The final sample consisted mostly of young adult, highly educated, frequent AI users.

The final sample ($n = 36$) ranged in age from 22 to 65 years ($M = 29.9, SD = 8.3$). 17 participants were female, 19 were male. Demographic data for one person were not saved because of the session timeout, and the participant later provided it themselves. 22 participants had a Master's degree, 10 had a Bachelor's degree, and four had completed high school. The majority were native German speakers ($n = 28$), followed by Russian ($n = 4$), with the remaining four reporting Portuguese, Ukrainian, Turkish, and Bulgarian as their first language. Most participants reported using AI tools daily or several times per week.

3.2.6 Apparatus

A custom web application was built to simulate an AI chat-based interface.

For the study, a web-based application was implemented with the Symfony 6.4 framework (PHP 8.4) and a MariaDB database, and deployed on a Linux container (Debian 11) hosted on a server provided by RWTH Aachen University. The website was publicly accessible through HTTPS. Response delays were controlled from the server through the application's logic to ensure that all participants experienced the same delay durations.

Participants completed the study remotely using a desktop or laptop.

Participants could access the study website remotely using their own devices via a web browser of their choice. Device metadata and browser version are logged automatically for each session. Participants were instructed to use a desktop or laptop rather than a mobile device. Moreover, they were asked to find a quiet environment without distractions, ensure a stable internet connection, and respond without using external resources during the task.

3.2.7 Measures

Primary Measures

Belief Rate. In each trial, participants had to decide whether they believed the generated AI answer. The decision was made by clicking on the "Yes" or "No" button. We defined belief rate as the proportion of trials where a participant believed the AI response, calculating the mean of the belief decision across all trials per condition. For each participant, four belief rate values (one per condition) were computed.

Confidence Rating. After each belief decision for each trial, participants rated their confidence in their judgements on a 7-point Likert scale, varying from 1 (not confident) to 7 (very confident). As for the belief rate, for each participant, we calculated the average confidence per condition.

Exploratory Measures

Reaction Time. For each trial, we collected how long it took participants to make a belief decision. We defined reaction time as the time interval between the AI answer being shown to the users and their belief decision click. It was analysed to explore whether the delay condition or complexity influenced participants' decision speed.

Time to Confidence. Moreover, we recorded the interval between the belief decision and the confidence rating. This measure was used to explore how much time they spent rating their confidence, as a potential indicator of uncertainty.

Timeouts. Trials in which participants did not provide a belief judgment within the 13 second time limit were recorded as timeouts. We used the number of timeouts as a criterion for participant exclusion.

Answer Accuracy. According to our experimental design, there were two ways to give a correct response: pro question: believing a correct AI answer or rejecting an incorrect AI answer. Based on this, we calculated answer accuracy as a further exploratory measure.

Post-Study Measures

Demographics. After completing all trials, participants reported their age, gender, highest level of education, occupation, native language, English proficiency (1-5 scale), frequency of AI chatbot usage, and technical affinity (1-5 scale).

Interaction Perception. Participants were asked to give their impression of the interaction. They rated to what extent prior knowledge, answer wording, and response time had influenced their decisions on a 5-Point Likert scale, ranging from 1 (not at all) to 5 (very strongly). Additionally, we assessed on a 5-point agreement Likert scale if partici-

pants noticed differences in delays during the experiment and if they perceived the system as "thinking".

Qualitative Feedback. Participants were asked to answer three open-ended questions: what they based their judgments on, which aspects of the interaction influenced their evaluation most, and whether they noticed any patterns or developed a strategy during the study. Furthermore, the questionnaire included a free-text field at the end, where the participants could leave their comments.

3.3 Pilot Studies

Pilot studies were used to refine the setup, timing, stimuli, and study flow.

Before starting actual data collection, we conducted several pilot studies to verify the technical setup, evaluate study flow, estimate session duration, and identify potential problems with the data set and study design.

The long delay was increased from 4000 ms to 6000 ms to make the contrast clearer.

First pilot (1 participant, 20 trials). Initially, the long delay was set to 4000 ms. This value was chosen based on the observation that, in human communication, silence tends to start feeling unusual after 4000 ms [Maslych et al., 2025]. However, the first pilot study revealed that the contrast between the short delay of 400 ms and the initially chosen long delay of 4000 ms was not sufficiently distinguishable. Therefore, we increased the long delay to the final value of 6000 ms for the next pilot sessions.

The second pilot led to a time limit and less familiar questions.

Second round (3 participants, 80 trials each). These three full-length trial runs uncovered two problems. First, reaction times varied very widely, from 1.7 to 59.6 seconds, with a mean of 13 seconds. In several trials, participants showed long response times, which suggests they were trying to recall information from memory rather than reacting to the AI answer itself. To increase the potential effect of the delay, we introduced a 10-second time limit for reading the answer and making a belief judgement. We visualised it by a visible count bar (see Figure 2). Second, many questions appeared to participants to be familiar, which obscured the effect of the delay manipulation. Therefore, the trial data set was edited: questions associated with high prior knowl-

edge were replaced, and the total number of trials was increased from 80 to 100. Adding of a response time limit made the study duration shorter, which allowed us to make a larger trial set without increasing the length of the experiment.

Final pilots increased the time limit and introduced blurring after belief decision.

Third round pilot (2 participants, 100 trials each). Two final pilot runs were conducted to verify the revised data set and to see how participants coped with the new time limit. One participant had many timeouts (24% out of 100), which indicated the limit was too short. For this reason, we increased the time limit to 13 seconds, corresponding to the mean response time observed in the second pilot round. Additionally, we discovered that participants used the confidence rating phase to reread the text of the trial. To address this, we blurred the question and the answer after a belief decision was made to ensure that the confidence rating based on the participant's gut feeling, influenced by the preceding delay, rather than a revised assessment.

3.4 Results

This section reports the findings of Study 1. After describing data preprocessing and exclusions, the primary analyses explore whether response latency and question complexity affected participants' belief judgements and confidence ratings. Next, exploratory analyses provide further context on response behaviour, trial believability, and participants' impressions of the interaction.

3.4.1 Data Preprocessing

Python was used for data preparation and statistical analyses using the `pandas` and `pingouin` libraries. Participants were excluded from the main analyses if they:

- Failed at least one of the two attention checks and/or
- Had more than 10 timeouts out of 100 trials.

Five participants were excluded due too many timeouts, resulting in final sample of 36.

Five of the 41 participants were excluded from the analysis on the basis of the timeout criterion, with timeouts ranging from 14 to 21 trials. Moreover, individual timeout trials were excluded from the analyses, because no belief decision and confidence rating were recorded for these cases. All participants passed the attention checks. The final sample consisted of 36 participants with a mean of 4.4 timeouts per participant ($SD = 1.8$).

3.4.2 Primary Analyses

Belief Rate

Belief rates means followed the predicted pattern.

First, we analysed the belief rate for each condition. Overall, participants believed 64.7% of the AI answers. The means and standard deviations are displayed in Table 3.6 and Figure 3.7. As shown, complex trials achieved slightly higher belief rates than simple trials. In addition, for complex questions, belief rates were higher in the long-delay condition than in the short-delay condition. For the simple questions, the pattern was reversed.

	Response latency	
Complexity	**Short**	**Long**
Simple	62.47 (SD = 12.00)	60.08 (SD = 14.00)
Complex	66.94 (SD = 11.91)	69.44 (SD = 12.42)

Table 3.6: Mean belief rates (%) per condition with standard deviations.

Complex answers were believed often, but the predicted interaction was not significant.

To test our hypothesis, the interaction effect between delay and complexity, we used a 2 (Latency) × 2 (Complexity)-ANOVA with repeated measures. The analysis showed a significant main effect of Complexity, $F(1, 35) = 19.99$, $p < 0.001$, $\eta_p^2 = 0.364$. The data show that complex questions were believed more often than simple questions. There was no significant main effect of Delay, $F(1, 35) = 0.002$, $p = 0.962$, $\eta_p^2 < 0.001$, meaning that short and long de-

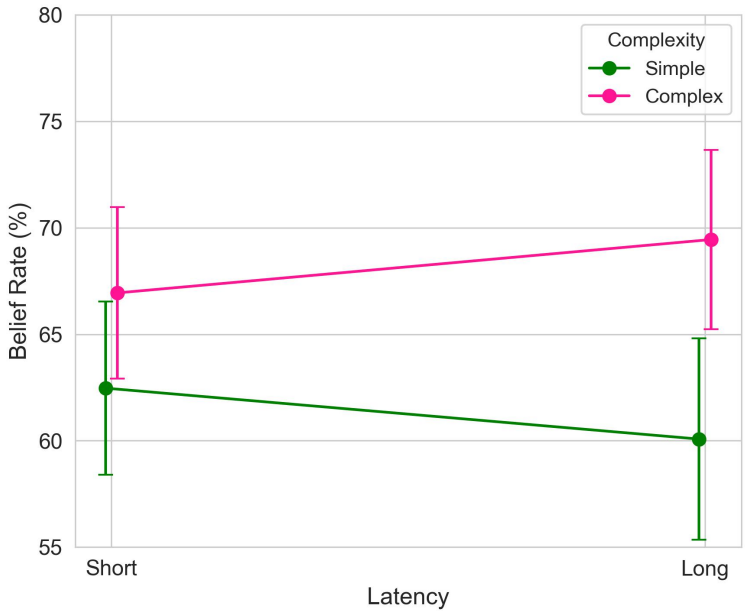

Figure 3.7: Belief rate by response latency and question complexity in Study 1.

lay did not vary in belief rate overall. The interaction between question complexity and delay was also not significant, $F(1, 35) = 1.98$, $p = 0.168$, $\eta_p^2 = 0.054$. Although the means showed the predicted effects, the interaction did not reach statistical significance.

Confidence Rating

Confidence ratings were intermediate and similar across delay conditions.

In general, the participants reported an intermediate level of confidence in their belief judgments ($M = 3.7, SD = 0.93$) on a scale from 1 to 7. Table 3.8 and Figure 3.9 provide the means and standard deviations for each condition. The confidence scores were similar across the delay conditions for both simple and complex questions.

The confidence ratings were analysed using the same 2×2 repeated-measures ANOVA, with complexity and delay as within-subjects factors. The results showed a similar pat-

	Response latency	
Complexity	**Short**	**Long**
Simple	3.61 (0.98)	3.65 (0.87)
Complex	3.76 (1.04)	3.77 (0.99)

Table 3.8: Mean confidence ratings per condition. Values represent M (SD) on a scale from 1 to 7.

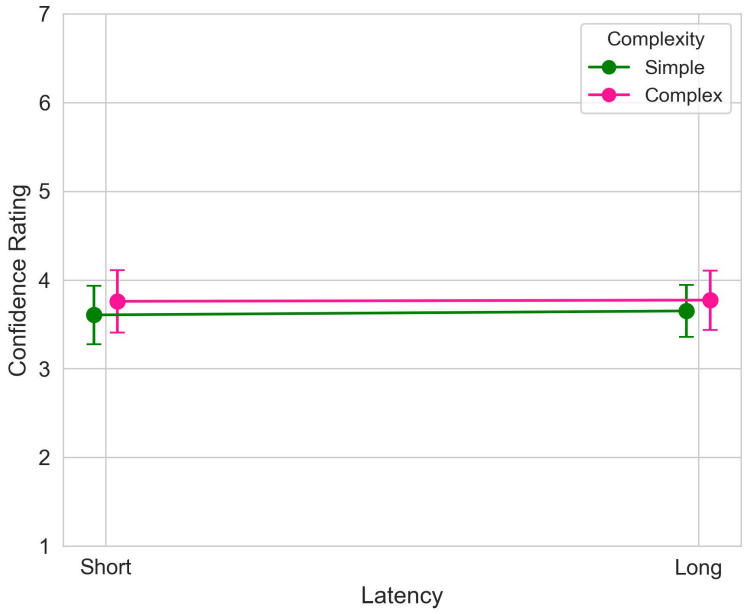

Figure 3.9: Confidence ratings across response latency and question complexity conditions in Study 1.

Confidence was slightly higher for complex trials, but no significant interaction.

tern to the belief rate: a significant main effect of complexity, $F(1, 35) = 4.99$, $p = 0.032$, $\eta_p^2 = 0.125$, meaning participants were slightly more confident in their judgements for complex question-answer pairs than for simple ones.There was no significant main effect of latency, $F(1, 35) = 0.34$, $p = 0.563$, $\eta_p^2 = 0.010$, and no significant interaction between latency × complexity, $F(1, 35) = 0.09$, $p = 0.760$, $\eta_p^2 = 0.003$.

3.4.3 Secondary Exploratory Analyses

Reaction Time

Reaction times were longer for complex trials and slightly longer after long delays.

As defined in Section 3.2.7, we additionally measured reaction time to explore whether response latency and question complexity affected participants' speed. Because answers to complex questions were longer than to simple ones (see Section 3.2.2), reaction times, measured from the appearance of the written answer to belief decision, were likely confounded with reading time. Mean reaction times were 6171 ms (SD = 985 ms) for simple-short trials and 6318 ms (SD = 1095 ms) for simple-long trials. For complex questions, mean reaction times were 9176 ms (SD = 1747 ms) in the short-delay condition and 9423 ms (SD = 1763 ms) in the long-delay condition. The ANOVA showed a highly significant main effect in complexity, $F(1, 35) = 328.99$, $p < 0.001$, $\eta_p^2 = 0.94$, consistent with this difference in answer length rather than reflecting differences in decision speed. However, there were also a significant effect of response latency, $F(1, 35) = 6.60$, $p = 0.015$, $\eta_p^2 = 0.159$, suggesting that participants decided slightly more slowly after a long delay. There was no interaction effect between complexity and response latency, $F(1, 35) = 0.48$, $p = 0.493$, $\eta_p^2 = 0.014$. Thus, the latency effect did not differ between simple and complex questions.

Time to Confidence

Participants took slightly longer to rate confidence after long delays.

Participants also took slightly longer to rate their confidence after long delays than after short delays. Mean time was 1800 ms (SD = 610) in the simple-short combination and 1972 ms (SD = 628) in the simple-long. For complex questions, mean time to confidence was 1797 ms (SD = 660) in the short-delay condition and 1903 ms (SD = 877) in the long-delay condition. The ANOVA indicated no significant main effect of complexity, $F(1, 35) = 0.22$, $p = 0.646$, $\eta_p^2 = 0.006$. However, there was a significant main effect of response latency, $F(1, 35) = 7.10$, $p = .012$, $\eta_p^2 = 0.169$, indicating that participants took longer to decide after long

delays. The interaction between complexity and response latency was not significant, $F(1, 35) = 0.20$, $p = 0.660$, $\eta_p^2 = 0.006$.

Answer Accuracy

Participants identified the correct answer in 58.1% of trials

As described in Section 3.2.2, the study consisted of 50 correct and 50 incorrect answers, where participants were informed about this distribution before the actual experiment. Participants identified the correct answer in 58.1% of trials. Accuracy was slightly higher for complex questions ($M = 60.4\%$) than for simple questions ($M = 55.9\%$), with minimal differences of 1-2% between delay conditions. Correct answers were believed in 72.6% of cases, whereas incorrect answers were disbelieved by clicking "No" only in 43.4% cases.

3.4.4 Exploratory Follow-Up Analyses

Exploratory analyses examined where the predicted interaction might become more visible.

Since no significant interaction was identified in the primary analysis, we decided to perform further exploratory analyses to determine under what conditions the hypothesised effect might be more clearly observed.

Median Splits

Median splits explored whether participant attributes moderated the interaction.

To explore whether the hypothesised interaction would occur in specific participant groups, we conducted several median-split analyses. For this, we divided participants into equally sized low and high groups ($n = 18$ each) based on their number of timeouts, mean confidence rating, and mean reaction time. Participants were split into two groups at the median value across the sample for each measure ($Mdn = 2.5$ timeouts, $Mdn = 3.87$, and $Mdn = 8118$ ms, respectively). For each median split, we used a mixed ANOVA on belief rate, with complexity and response latency as within-subject factors and median-split group as a between-subject factor. For the statistical analyses of the

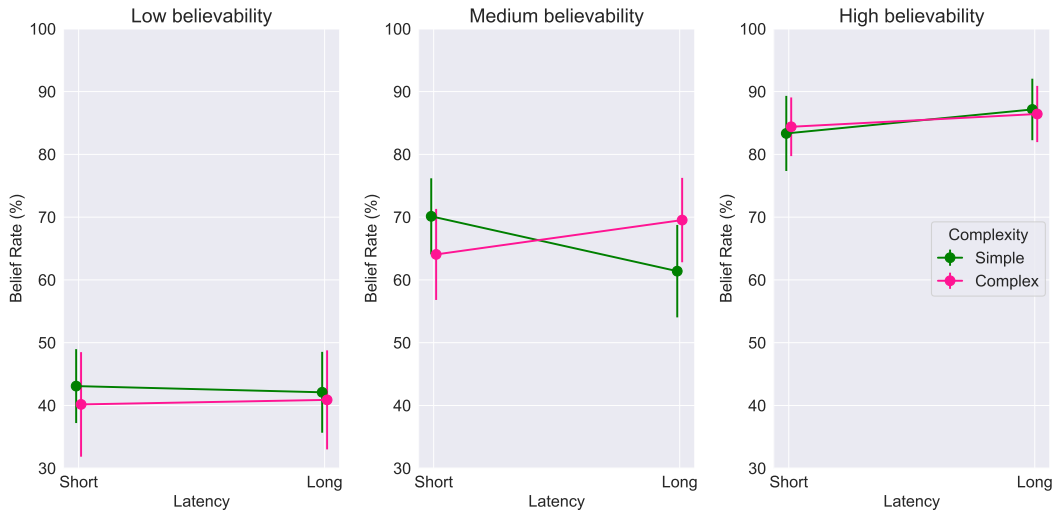

Figure 3.10: Interaction plots of belief rate by latency condition and complexity for each believability bin.

splits, we used R and the `rstatix` package. The descriptive belief rates for all six groups are provided in Table A.1, and the corresponding interaction plots are shown in Figure A.2 in the Appendix.

The predicted pattern appeared descriptively stronger in some groups but was not significant.

None of the three splits indicated a significant three-way interaction between group, complexity, and response latency (timeout split: $F(1, 34) = 0.835$, $p = 0.367$, $\eta_p^2 = 0.024$, mean confidence rating: $F(1, 34) = 0.105$, $p = 0.747$, $\eta_p^2 = 0.003$, reaction time split: $F(1, 34) = 0.90$, $p = 0.349$, $\eta_p^2 = 0.026$). None of these selected participants' characteristics significantly moderated the predicted interaction. However, descriptively, the complexity $\times$ latency interaction appeared numerically larger in the low-timeout, low-confidence, and slow-reaction-time groups than in their respective counterparts, although it did not reach significance within any individual subgroup.

Believability Bin Analysis

The trials for the study were selected to minimise the effects of prior knowledge by taking a range of different topics for which participants were unlikely to know the cor-

Trials were grouped by overall believability to discover cases with greater uncertainty.

rect answer with certainty. Nevertheless, some items may still have appeared more obvious to some participants than to others. The underlying idea was that participants may rely on other available cues, such as response latency, to make a decision. This should be most relevant for items whose overall belief rate lies in a medium range. Therefore, we grouped items by their overall belief rate among all participants. For each of the 100 question-answer pairs, we computed the mean belief rate, regardless of which delay condition they were in. Items were then divided into three equally sized groups: low (33 items, 13.9% - 54.3%), medium (33 items, 55.6% - 76.5%), and high believability (34 items, 77.1% - 97.2%) bin. The division resulted in 20 simple and 13 complex items in the low bin, 18 simple and 15 complex in the medium bin, and 12 simple and 22 complex items in the high bin. For each group, we calculated the belief rate separately for each experimental condition.

The predicted interaction was clearest for items with medium believability.

Figure 3.10 illustrates the resulting interaction plots for the three bins. The low- and high-believability bins showed no clear differences between the conditions. In contrast, the medium-believability showed the clearest pattern in line with our hypothesis. For simple questions, the mean belief rates were higher in the short-delay condition (70.1%) than in the long-delay condition (61.4%). For complex questions, the pattern was reversed, with higher belief rates in the long-delay condition (69.5%) than in the short-delay condition (64.1%).

3.4.5 Self-Report Measures

Interaction Perception

At the end of the study, participants rated several aspects of what had influenced their belief decision and how they perceived the AI agent. Figure 3.11 shows the response distribution for three items: prior knowledge, wording and tone, and response time.

Prior knowledge received the highest ratings, ($M = 4.6, SD = 0.73$). Most of the participants rated it as

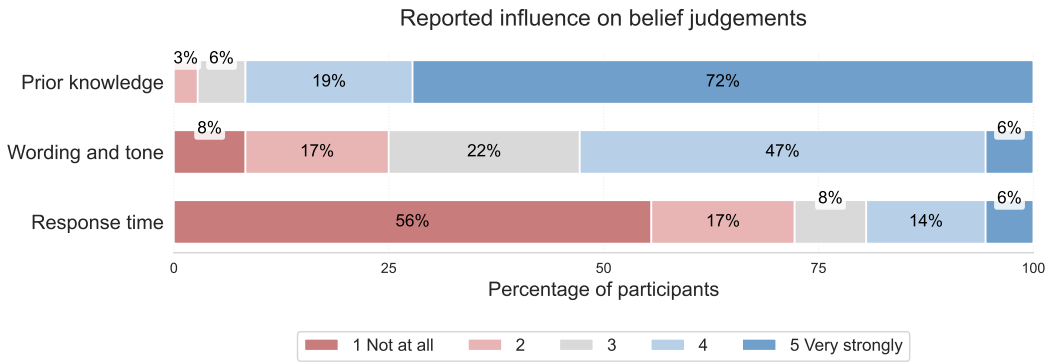

Figure 3.11: Distribution of responses to the self-report items about perceived influences on belief judgements. The scale ranged from 1 (not at all) to 5 (very strongly).

a strong or very strong influence, with 26 of 36 participants (72.2%) choosing the maximum rating. The wording and tone of the AI answers received more varied ratings, and most of the participants rated it as influential ($M = 3.25, SD = 1.08$). The response time of the system showed the opposite pattern ($M = 1.97, SD = 1.32$). More than half of the participants (56%) rated the delay as not influential at all, and only 7 of 36 (19.4%) rated it as a strong or very strong influence factor.

Participants reported relying mostly on prior knowledge, not response time.

As shown in Figure 3.12, 29 of 36 participants strongly agreed that the AI system responded at different speeds ($M = 4.61, SD = 0.90$). This suggests that response latency manipulation was generally noticed. The ratings for "whether the system seemed to be thinking before responding" were spread across all five scale points ($M = 2.72, SD = 1.19$). 15 of 36 participants (41.6%) disagreed or strongly disagreed with the statement, 11 participants (30.6%) chose the midpoint of the scale, and 10 participants (27.8%) agreed or strongly agreed.

Most participants noticed differences in the system's response speed.

Qualitative Feedback

The participants' open responses added more context to the results. The comments repeatedly referred to participants' own knowledge, gut feeling, the plausibility of the answer,

Qualitative feedback highlighted prior knowledge, plausibility, wording, and time pressure.

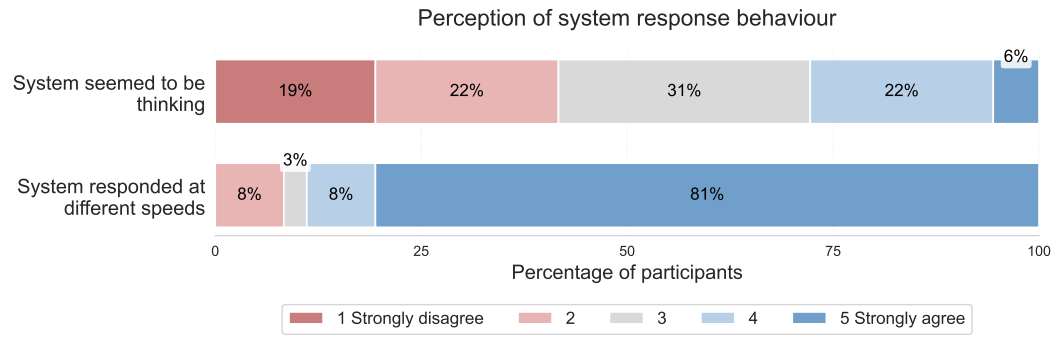

Figure 3.12: Distribution of responses to the self-report items about perceived system response behaviour. The scale ranged from 1 (strongly disagree) to 5 (strongly agree).

answer length and wording, as well as time pressure during the study.

Prior knowledge and gut feeling. Many participants described prior knowledge as the main basis for their decision making. When they already knew answer to the raised question, they compared if the AI response matched their own assumptions. Often, when they did not know the answer, they relied on plausibility or gut feeling. Other participants described this more briefly, for example, as relying on *"my knowledge of the facts and gut feeling"* (P6), or deciding based on whether they *"already know or expect a certain answer"* (P9).

> *"I based my judgement on my own knowledge and/or intuition and on how complex/abstract the questions and answers were."* (P39)

Several participants explicitly described relying on plausibility when they lack the necessary knowledge:

> *"If plausible explanations were given, I was more likely to say 'yes'. Although, I didn't know at all whether the answer was correct or not, I tended to say 'yes'."* (P50)

Answer length and wording. Several participants mentioned that they mostly relied on such cues as answer length, and answer wording to make a decision.

> *"The first criteria was the length of the answer. Short answers directly seem less reliable. The second criteria was the use of professional vocabulary."* (P45)

For some participants, shorter answers were associated with higher credibility. Some participants found longer and detailed answers more convincing. For example, one participant stated "*the detailed explanations added to the feeling of "reliability" of said answers*" (P24). Others became more sceptical when answers were long: *"the more detailed they got, the likelier I thought it was to be a lie"* (P10).

Time pressure. Another frequently mentioned problem was the time limit. The 13-second restriction made it difficult to read the response completely and evaluate it carefully, especially for longer answers to complex questions. Several participants reported that, as a result, they often had to guess rather than make a careful judgement.

> *"The time for reasoning was obviously very short, which caused me to be very uncertain about those answers in particular, but followed this gut feeling to yes or no."* (P24)

Response time as a cue. Several participants noticed variation in response times and reported that it influenced their evaluation.

> *"For immediate responses, I tend to question the answer more than if the AI pretends to think about the answer."* (P15)

3.5 Discussion

Study 1 tested whether latency and complexity interact in a text-based setting.

Study 1 investigated whether response latency interacts with the complexity of the question with respect to users' belief judgements of AI-generated answers in a text-based chat setting. We hypothesised that the effect of response latency on belief rate would depend on question complexity. Specifically, we expected short delays to increase belief rates for answers to simple questions, whereas long delays would increase belief rates for answers to complex questions.

No significant interaction, but descriptively in the expected direction.

The interaction effect did not reach statistical significance. However, the means were consistent with our prediction: belief rates were slightly higher after short delays than after long delays for simple questions, whereas answers on complex questions were believed slightly more often after long delays than after short delays.

The hypothesised pattern was clearest when answers were difficult to judge.

The exploratory believability bin analysis indicated that the predicted interaction may be present, but only for items where participants' judgements were uncertain. In the medium believability bin, the difference between short and long delay was larger than in the overall sample, especially for simple questions (compare Figure 3.10 with Figure 3.7). In contrast, the low and high believability bins, in which the answers were more often rejected or more often believed, showed almost no difference between delay conditions. One possible interpretation is that the medium bin contained items for which neither prior knowledge nor the plausibility of the answer was sufficient to guide participants' decisions, leaving more room for additional cues such as response latency.

A voice-based setting may strengthen the social meaning of response latency.

Thus, Study 1 suggested that response latency may still influence belief judgements, especially when answers are difficult to judge. One possible reason for the weak effect is that the text-based setup may have limited the social interpretation of the delay. The delays may have been interpreted more as a technical processing time rather than a pause in conversation [Nielsen, 1993; Shneiderman, 1984]. Response latency can also function as a chronemic cue in

social interaction [Kim et al., 2026]. In voice-based interaction, timing differences may become more noticeable or take on a different meaning [Cohn et al., 2024]. This motivated us to conduct a second study to investigate whether the interaction between latency and complexity would become more pronounced in voice-based communication. Study 2 was designed as a conceptual extension of Study 1, using the same question-answer pairs and delay manipulation, but presenting the AI responses as spoken rather than written answers.

Chapter 4

User Study 2: Voice-Based Interaction

While Study 1 examined response latency in a text-based chat interface, many real-world interactions with AI systems increasingly take place through voice. Voice modes are now integrated into widely used systems such as ChatGPT and Gemini, and voice assistants like Alexa and Siri have long relied on spoken exchanges. If response latency functions as a social timing cue, as discussed in the previous chapter, it may be even more salient in an interaction using voice mode. Adding voice to an LLM system has been found to increase both anthropomorphism ratings and perceived information uptake compared to a text-only condition [Cohn et al., 2024]. If voice increases the extent to which an AI system is treated as a communication partner, response latency in this modality may be interpreted more as a "thinking" pause instead of technical processing time.

Study 2 extended the text-based design to a voice-based AI interaction.

Study 2 was built directly on Study 1, applying the same concept but in a voice-based setting. Participants read each question aloud, and the AI responded in a synthetic natural voice instead of written text. We aimed to determine whether the patterns observed in the first study would also appear or become more pronounced in this audio format. This chapter first presents Study 2 hypotheses, followed by its design, procedure, measures and results.

4.1 Research Questions & Hypotheses

Study 2 retested the hypothesis in voice mode.

Study 2 was designed as a voice-based extension of Study 1, retesting the hypotheses formulated in Study 1 (see Section 3.1). This approach allowed us to investigate whether the expected interaction between response latency and question complexity occurs in voice-based interaction, as well as to compare these findings to those obtained in Study 1.

H2 As in Study 1, response latency interacts with question complexity in influencing users' belief judgements and confidence ratings (*H2.1, H2.2* correspond to *H1.1* and *H1.2*).

H3 focused on medium believability items, where judgements are most ambivalent.

We extended the hypotheses formulated for Study 1 on the basis of the believability bin analysis conducted there (see Section 3.4.4), which demonstrated that the predicted interaction was most pronounced for items with a medium overall belief rate, i.e. for items where participants' judgements were most ambivalent. This led to an additional hypothesis for Study 2:

H3 The interaction between response latency and question complexity will be more pronounced for items with medium believability than for items with a low or high overall belief rate.

4.2 Method

The section describes the methodological setup of Study 2. It builds on Study 1 and highlights the changes made for the voice mode. We describe the study design, participant sample, audio stimulus generation, voice selection, procedure, counterbalancing, and measures used to evaluate participants' perceptions of the AI agent.

4.2.1 Study Design

Study 2 used the same 2x2 design, but with audio responses.

Study 2 used the same 2x2 within-subjects design as Study 1 (see Section 3.2.1), manipulating response latency and question complexity, with the one change being the interaction modality: participants read the question aloud themselves, and the answer was played via audio rather than text.

4.2.2 Participants

Participants needed a desktop or laptop, and working audio output.

We recruited 37 participants through convenience and snowball sampling using personal networks. Participation was voluntary and uncompensated. The same requirements as in Study 1 applied: sufficient English proficiency and a desktop or laptop. Additionally, participants were required to have working speakers or headphones to be able to hear the AI's audio responses.

The final sample consisted of 34 participants.

Demographic data for two participants were not saved due to a session timeout, and the participants later provided it themselves. The final sample consisted of 34 participants and ranged in age from 19 to 66 years (M = 30.1, SD = 10.0). 17 participants identified as women, 15 as men, one as non-binary, and one preferred not to disclose their gender. 16 participants had a Bachelor's degree, 11 a Master's degree, and seven had completed high school. The majority were native German speakers (n = 18), followed by Russian (n = 10), with the remaining six reporting Belarusian, Bosnian, Greek, Turkish, English. Most participants reported using AI tools daily (n = 15) or several times per week (n = 13).

4.2.3 Stimuli & Materials

Study 2 used the same question-answer set as Study 1.

For the second study, we used the exact same set of questions and answers (see Section 3.2.2). No changes have been made to the content, accuracy, or complexity.

Audio Stimulus Generation

The written AI answers were synthesised as MP3 files using ElevenLabs.

In Study 1, we presented all material as written text. In contrast, in Study 2, we shifted to voice mode. The system's answers were generated using the text-to-speech (TTS) software ElevenLabs[1], with its Multilingual v2 model, the same default model used by Michel et al. in their evaluation of synthetic AI voice service [Michel et al., 2025]. All 100 answers were synthesised using a single female voice at the default speech rate, and exported as MP3 files (44.1 kHz, 128 kbps).

Voice Selection

Voice gender can influence perceived warmth, competence, and trust.

Since the voice itself can influence how audio responses are perceived, we first examined common practices regarding voice gender in commercial voice assistants and conversational AI systems. Historically, voice assistants such as Siri, Alexa, and Cortana have tended to use female voices as their default setting for extended periods [Sindoni, 2024]. This convention has been related to the perception that female voices convey warmth and support, while male voices are associated with authority and competence [Danielescu et al., 2023]. Female voices have also been shown to increase user trust in socially sensitive and health-related contexts [Sindoni, 2024]. However, the default usage of female voices has been criticised for strengthening gender stereotypes [Danielescu et al., 2023].

A female voice was chosen for naturalness and familiarity.

We also considered a gender-neutral voice to avoid gender-specific associations. However, none of the gender-neutral voices available in ElevenLabs at the time of stimulus generation for Study 2 sounded sufficiently natural and convincing. Therefore, we selected a voice based on the criteria of perceived confidence and naturalness among the available samples, as well as the goal of staying closer to what participants might expect from a commercial voice assistant. Thus, we chose "Victoria - Warm, Trustworthy, and Relatable", a female voice. As outlined above, this choice

[1] https://elevenlabs.io/, *last accessed July 5, 2026*

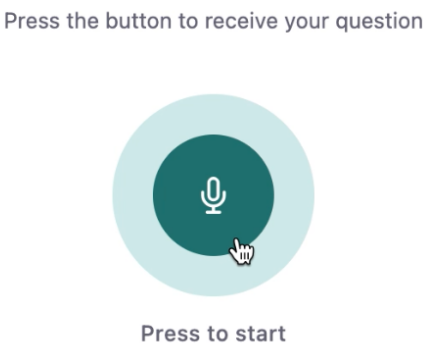

Figure 4.1: The microphone button as it appeared at the start of each trial with the instruction to press it to receive the question.

is not unproblematic and should be understood as a design decision based on the available voice samples.

4.2.4 Procedure

The study began with consent, a sound check, and a tutorial.

As in Study 1, the study was conducted online and without any assistance. Participants first visited a welcome page with general information and technical requirements, followed by the informed consent form. In addition to the consent information provided in Study 1, participants were informed that their voices were not recorded or saved throughout the study. Following consent, participants completed a brief sound check to ensure their device's audio was working before the experiment started. Participants then completed a tutorial, as in Study 1. Furthermore, they were also advised that half of the AI's answers were correct and half were incorrect.

Participants read each question aloud to imitate speaking to a voice assistant.

Each trial began with a microphone button (see Figure 4.1), and participants were instructed to press it to receive the trial question. Pressing the button displayed the question and instructed participants to read it aloud. As shown in Figure 4.2, the button then changed to a red recording state to visually indicate that the spoken input phase was active and the AI agent was "listening". The idea was to imitate

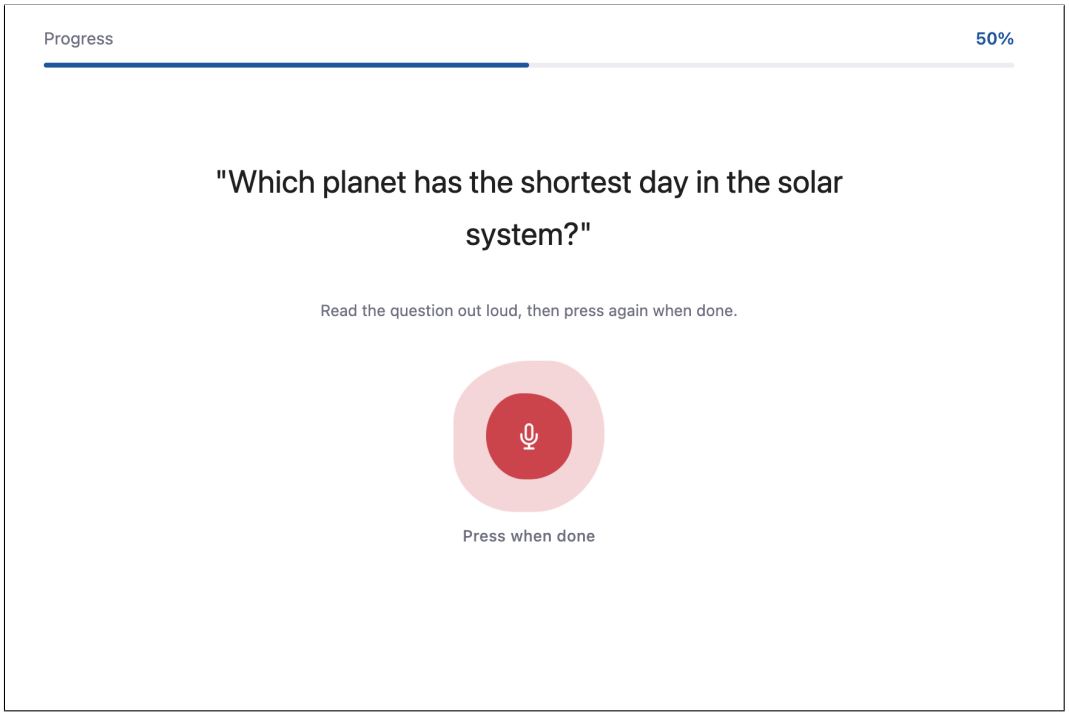

Figure 4.2: After the first press button, the question (e.g., "Which planet has the shortest day in the solar system?") is displayed, and the microphone button changes its state to signal the recording phase.

the way a user might speak to a voice assistant. Participants were asked to press the button once they had finished reading the question aloud. No audio of participants' speech was recorded.

The latency manipulation began after participants finished reading aloud.

Next, the question was displayed as a quoted message in the user's chat bubble, similar to Study 1. The text was taken from the stimulus set rather than a transcription of the participant's speech. Rather than reusing the three-dot typing indicator animation from Study 1, a small pulsing circle was shown during the delay period. The delay between the end of the question and the start of the audio AI response was either 400 ms or 6000 ms, matching the latency manipulation used in Study 1.

A moving visual form represented AI agent.

Following the delay, the AI's audio answer played automatically. As illustrated in Figure 4.3, during playback, the AI agent was visualised as a fluid, shape-changing form

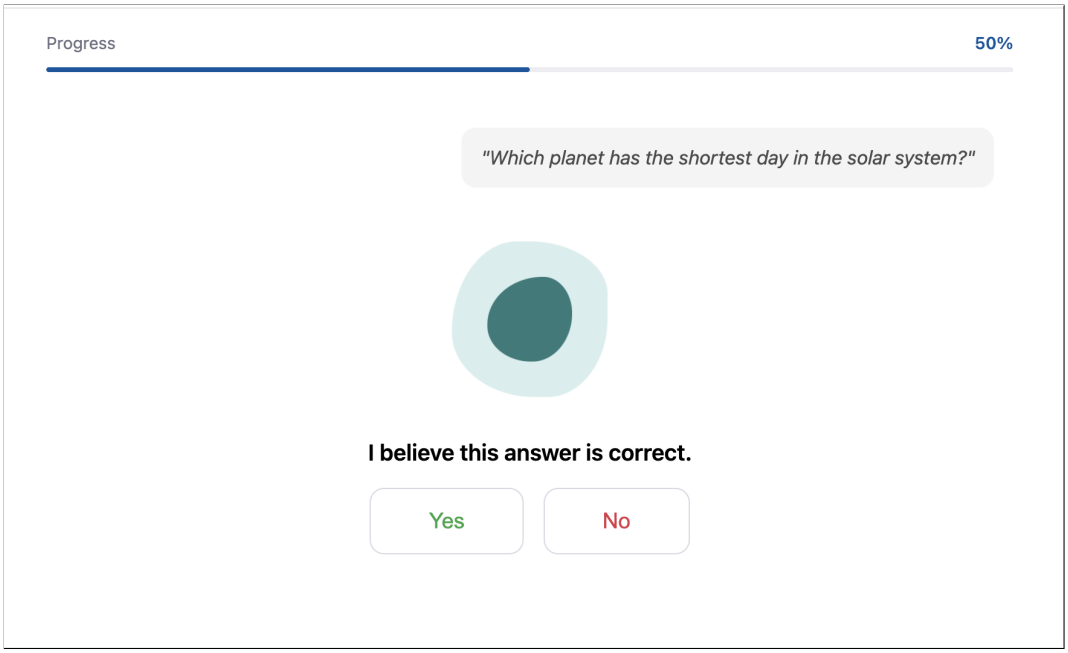

Figure 4.3: The chat interface after the question has been submitted, showing the asked question, and the AI agent, represented as a moving animation, with the belief decision prompt.

built from two overlapping circles that moved and deformed, inspired by the visual representation of ChatGPT's mobile voice mode. This design choice was made to reinforce the impression of a single, continuous AI that handles the "listening", "thinking", and "speaking" states.

Participants judged the AI answer within a 7-second response window.

As in Study 1, participants judged whether they believed the answer. Participants had seven seconds to respond, visualised by a countdown progress bar. This response window was shortened relative to 13 seconds in Study 1, which included time to read the written answer in addition to making a decision. In Study 2, the answer was presented entirely via audio, so the decision window was halved, as no reading time was required. Participants could also decide before the audio finished playing, in which case the audio stopped, and the trial was marked as an early decision. After the belief decision, the animated fluid form shrank to a single dot. As in Study 1, participants rated their confidence and then proceeded to the next trial.

Audio attention checks verified that participants were listening.

The main study consisted of 102 trials, two of which served as attention checks. In Study 2, these trials displayed an "Attention Check" label instructing participants to listen to the following audio message. The audio contained a note to click "Yes" or "No" to confirm they had been listening.

Study 2 ended with questionnaires and took approximately 50-60 minutes.

After completing all trials, participants completed a demographics and perception questionnaire, both adapted from Study 1 (see 3.2.7), with several changes to reflect the audio-based interaction (see Section 4.2.7 for details). Unlike in the text-based study, Study 2 did not include a reading speed task. The study took approximately 50-60 minutes to complete.

4.2.5 Counterbalancing & Randomisation

Counterbalancing followed the same logic as in Study 1.

Counterbalancing and randomisation followed the same logic as in Study 1 (see Section 3.2.4). Trials were assigned to delay condition based on participant ID, and the 100 trials were divided into 25 blocks of four, with block order and trial order within each block randomised.

4.2.6 Apparatus

The web application matched Study 1, with audio playback added.

The web-based application was implemented and hosted identically to Study 1 (see Section 3.2.6). The only additions were audio playback. Pre-generated audio answer files were served directly to participants' browsers.

4.2.7 Measures

Primary Measures

Measures matched Study 1.

As in Study 1, belief rate and confidence rating were defined and computed identically (see Section 3.2.7).

Exploratory Measures

As in Study 1, time to confidence, timeouts, and answer accuracy were collected as secondary exploratory measures (see Section 3.2.7). Reaction time was defined similarly to Study 1 and measured from the start of the AI's audio response. We additionally recorded whether a belief decision was made before the audio finished playing.

Post-study Measures

Demographics. As in Study 1 (see Section 3.2.7), with two additional items to collect prior experience with AI voice modes or voice assistants and whether participants had experienced any technical problems during the study.

Interaction Perception. As in Study 1 (see Section 3.2.7), except that the item addressing wording was replaced with an item addressing the voice and speaking style of the AI agent. Four additional items were added on a Likert scale to assess participants' impressions regarding the naturalness of the AI voice and interaction, the animated visual during speaking, and whether the animation was distracting.

Qualitative Feedback. Unchanged from Study 1 (see Section 3.2.7).

4.3 Pilot Studies

Pilot studies checked audio playback, attention checks, and study flow.

Before starting actual data collection, we conducted two pilot studies to ensure the technical setup was working, evaluate the study flow, and identify potential problems with data storage.

The first pilot session was conducted online with a participant who had previously taken part in Study 1, with a particular focus on verifying that audio playback functioned

Pilots led to redesigned audio attention checks and confirmed engagement with the read-aloud task.

correctly throughout the study. The pilot identified an issue with the initial attention-check design. As in regular trials, participants were asked to read the attention-check content aloud. However, since the content was an instruction rather than a question, this felt confusing and unintuitive. Thus, we redesigned the attention checks so the participants only had to listen to an audio message and respond according to the instructions. The second pilot session was conducted in person and proceeded without any technical issues. Therefore, we included this participant's data in the main analysis. During this pilot study, we observed that the participant moved noticeably closer to the screen while reading questions aloud. This behaviour implied that the person engaged with the speaking task as intended, rather than passive button-clicking.

4.4 Results

This section presents the results of Study 2. After data preprocessing and exclusions, the primary analyses assess how latency and complexity affect belief judgements and confidence ratings. Exploratory analyses provide more insights on response behaviour, trial believability, participant subgroups, and participants' impressions of the voice-based interaction.

4.4.1 Data Preprocessing

Data preparation, exclusion criteria, and statistical analyses followed the same procedure as in Study 1 (see Section 3.4).

Exclusions were based on technical problems and failed attention checks.

Of the 37 participants who completed the study, one was excluded because she reported technical problems during the study, with the audio not playing correctly on many trials. Two further participants were excluded for failing at least one of the two attention checks. No participants exceeded the timeout threshold. The final sample consisted of 34 participants with a mean of 1.8 timeouts per participant (SD = 2.3, range: 0-9).

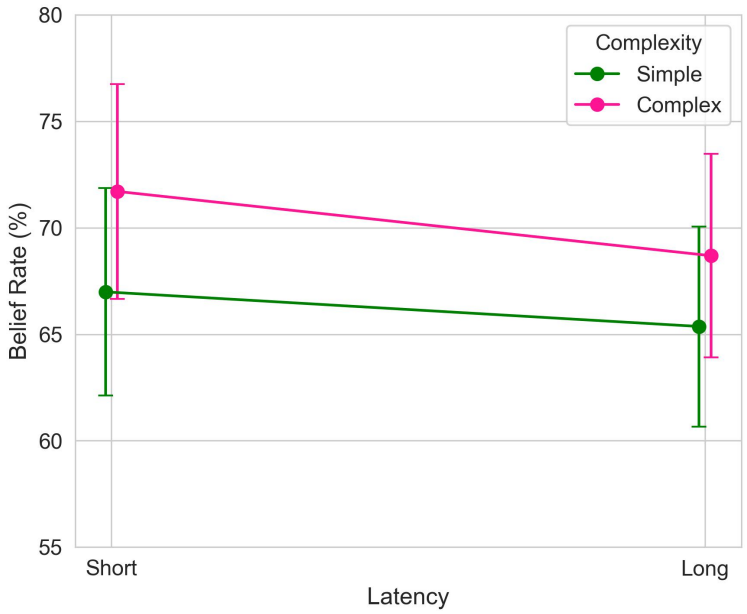

Figure 4.4: Belief rate by response latency and question complexity in Study 2.

4.4.2 Primary Analyses

Belief Rate

Belief rates were higher after short delays for both simple and complex trials.

First, we analysed the belief rate for each condition. Overall, participants believed 68.2% of the AI answers (SD = 10.7%, range: 46.9% – 94.9%). The means and standard deviations are displayed in Table 4.5 and Figure 4.4. As can be seen, complex trials achieved higher belief rates than simple trials. Belief rates were higher in the short-delay condition than in the long-delay condition for both simple and complex questions.

The predicted latency-complexity interaction was not significant.

To test our hypothesis, the interaction effect between delay and complexity, we used a 2 (Latency) × 2 (Complexity)-ANOVA with repeated measures, as in Study 1. Unlike in Study 1, the analysis showed no significant main effect of complexity, $F(1, 33) = 3.61$, $p = 0.066$, $\eta_p^2 = 0.099$. Descriptively, however, belief rates were higher for complex

	Response latency	
Complexity	**Short**	**Long**
Simple	67.0 (14.0)	65.4 (13.5)
Complex	71.7 (14.5)	68.7 (13.7)

Table 4.5: Mean belief rates (%) per condition with standard deviations in brackets.

than for simple trials. There was no significant main effect of delay, $F(1, 33) = 1.67$, $p = 0.205$, $\eta_p^2 = 0.048$, meaning that short and long delay did not vary reliably in belief rate overall. Finally, the interaction between question complexity and response latency was also not significant, $F(1, 33) = 0.33$, $p = 0.569$, $\eta_p^2 = 0.01$. As in Study 1, the hypothesised effect of response latency on belief rate did not depend on question complexity.

Confidence Rating

Confidence ratings were moderate to high and similar across delays.

In general, participants reported a moderate to high level of confidence in their belief judgements ($M = 4.38$, $SD = 0.77$) on a scale from 1 to 7. Table 4.6 shows the means and standard deviations for each condition. As in Study 1, confidence ratings were similar across delay conditions for both simple and complex questions.

Confidence ratings were higher for complex items, but did not depend on latency.

The confidence ratings were analysed using the same 2x2 repeated-measures ANOVA. As in Study 1, there was a significant main effect of complexity, $F(1, 33) = 7.01$, $p = 0.012$, $\eta_p^2 = 0.175$, meaning participants were more confident in their judgements for complex question-answer pairs than for simple ones. There was no significant main effect of response latency, $F(1, 33) = 0.56$, $p = 0.459$, $\eta_p^2 = 0.017$, indicating that the effect of delay on confidence did not depend on question complexity in this sample. Finally, there was no significant interaction between latency and complexity, $F(1, 33) = 0.88$, $p = 0.356$, $\eta_p^2 = 0.026$.

	Response latency	
Complexity	**Short**	**Long**
Simple	4.27 (0.83)	4.29 (0.74)
Complex	4.52 (0.84)	4.42 (0.90)

Table 4.6: Mean confidence ratings per condition in Study 2. Values represent M (SD) on a scale from 1 to 7.

4.4.3 Secondary Exploratory Analyses

Reaction Time

As the duration of the AI's audio files differed between simple and complex answers, the raw reaction time was strongly confounded with answer length. Complex trials showed longer reaction times (M = 14450 ms) than simple trials (M = 7811 ms).

Reaction times were slightly longer after long delays.

Participants made their belief decision before the audio had finished in 17.5% of trials, and after the audio had ended in the remaining 82.5%. For trials in which a decision was made after the audio was played, we additionally calculated the post-decision time. The pattern was similar to the raw reaction time: a significant main effect of complexity, $F(1, 32) = 147.52$, $p < .001$, $\eta_p^2 = 0.82$, but no significant effect of latency, $F(1, 32) = 0.77$, $p = 0.387$, $\eta_p^2 = 0.024$, nor a significant interaction, $F(1, 32) = 0.19$, $p = 0.667$, $\eta_p^2 = 0.006$. However, post-audio decision times were descriptively slightly longer after long delays.

Time to Confidence

Unlike in Study 1, latency did not affect time to confidence.

Unlike in Study 1, where latency significantly affected time to confidence, no significant effects of complexity, $F(1, 33) = 3.87$, $p = 0.057$, $\eta_p^2 = 0.105$, latency, $F(1, 33) = 0.06$, $p = 0.808$, $\eta_p^2 = 0.002$, or their interaction, $F(1, 33) = 0.51$, $p = 0.481$, $\eta_p^2 = 0.015$, were found.

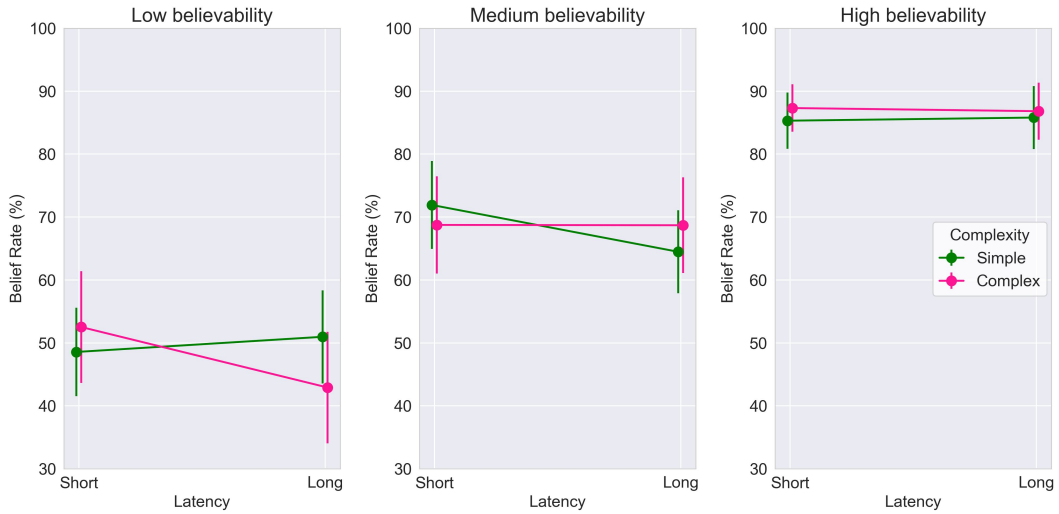

Figure 4.7: Interaction plots of belief rate by latency condition and complexity for each believability bin.

Answer Accuracy

Participants identified the correct answer in 56.4% of trials

As in Study 1, the answer set consisted of 50% correct and 50% incorrect answers. Participants identified the correct answer in 56.4% of trials. Accuracy was similar for complex (M = 56.7%) and simple questions (M = 56.1%), with minimal differences between delay conditions. As in Study 1, correct answers were believed considerably more often (74.5%) than incorrect answers were disbelieved (38.2%).

4.4.4 Exploratory Follow-Up Analyses

Believability Bin Analysis

The medium bin only partially supported the predicted pattern.

As in Study 1, we also investigated the predicted latency × complexity pattern specifically for items with a medium overall belief rate. Items were again divided into three equally sized bins based on each trial's overall belief rate: low (33 items, 29.4% - 60.6%), medium (33 items, 60.6% - 79.4%), and high believability (34 items, 79.4% - 97.1%). The division resulted in 14 complex and 19 simple items in the

low bin, 16 complex and 17 simple in the medium bin, and 20 complex and 14 simple items in the high bin.

For each bin, we calculated the belief rate separately for each experimental condition. As illustrated in Figure 4.8, the high-believability bin showed no clear differences between conditions. However, the low bin showed little difference for simple questions (48.6% in the short-delay condition vs 51.0% in the long-delay condition), but a notable difference for complex questions (52.5% vs 42.9%) in the opposite direction to what *H2.2* would predict. In the medium bin, the predicted pattern was observed only partially. For simple questions, the mean belief rate was higher in the short-delay condition (71.9%) than in the long-delay condition (64.5%), consistent with *H2.1*. For complex questions, however, belief rates were nearly identical between short- and long-delay conditions (68.8% vs 68.7%), providing no support for the expected pattern.

Median Splits

Median splits tested if participant subgroups showed stronger patterns.

We performed, as in Study 1, three median-split analyses to investigate if the expected interaction would be more pronounced in certain participant subgroups. We divided them into equally sized low and high groups ($n = 17$ each) based on their number of timeouts, mean confidence rating, and mean reaction time. For each measure, participants were split into two groups at the median value across the sample ($Mdn = 1$ timeout, $Mdn = 4.45$, and $Mdn = 11,809$ ms, respectively). Descriptive belief rates for all six groups are provided in Table A.2, and corresponding interaction plots are shown in Figure A.2 in the Appendix.

Some subgroups showed a general short-delay preference, but not the predicted interaction.

None of the three mixed ANOVAs revealed a significant three-way interaction between group, complexity, and response latency. However, a consistent pattern was observed in the subgroup with more timeouts, higher mean confidence, or slower mean reaction time. Belief rates were higher in the short-delay condition than in the long-delay condition for both simple and complex questions, whereas this difference was absent or minimally visible in the re-

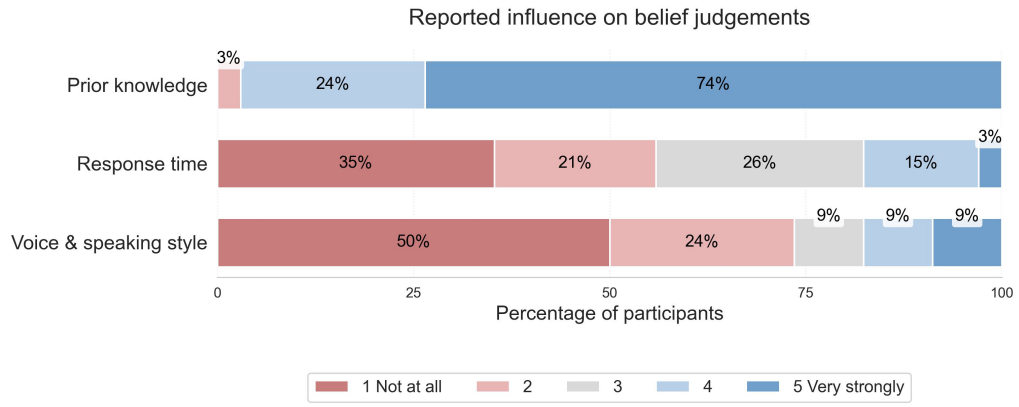

Figure 4.8: Distribution of responses to the self-report items about perceived influences on belief judgements in Study 2. The scale ranged from 1 (not at all) to 5 (very strongly).

spective other subgroup. This main effect of latency was significant for the higher-confidence group, $F(1, 16) = 5.35$, $p = 0.034$, $\eta_p^2 = 0.250$, and for the slower-reaction-time group, $F(1, 16) = 6.40$, $p = 0.022$, $\eta_p^2 = 0.286$, but not for their counterparts ($p = 0.787$ and $p = 0.742$, respectively). For the timeout median split, the mixed ANOVA showed significant group × complexity, $F(1, 32) = 5.19$, $p = 0.030$, $\eta_p^2 = 0.140$, and group × latency interactions, $F(1, 32) = 4.56$, $p = 0.040$, $\eta_p^2 = 0.125$. Thus, both complexity and latency influenced belief rate in the higher-timeout group but had little effect in the lower-timeout group. None of these patterns was observed in Study 1.

4.4.5 Self-Report Measures

Interaction Perception

Participants again reported relying mostly on prior knowledge.

Figure 4.8 shows the response distribution for three items: prior knowledge, response time, and the AI's voice. Similarly to Study 1, prior knowledge received the highest ratings ($M = 4.68$, $SD = 0.64$), with 97% rating it as a strong or very strong influence. Response time received, in contrast,

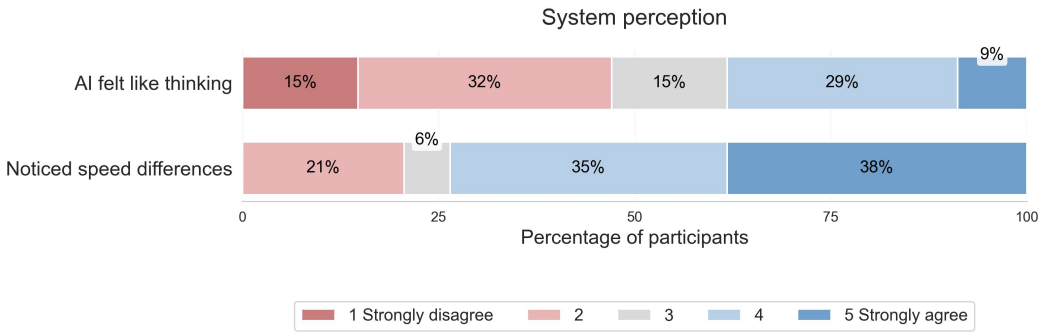

Figure 4.9: Distribution of responses to the self-report items about system response behaviour in Study 2. The scale ranged from 1 (strongly disagree) to 5 (strongly agree).

low ratings ($M = 2.29, SD = 1.19$). More than half of the participants rated it as not influential or only slightly influential, and just 6 of 34 rated it as strong or very strong influence. The AI's voice received similarly low ratings ($M = 2.03, SD = 1.34$), with 73.5% of participants rating it as not influential or only slightly influential.

Most participants noticed different delays, but not all interpreted the system as thinking.

As shown in Figure 4.9, 25 of 34 participants (73.5%) agreed or strongly agreed that the system responded at different speeds ($M = 3.91, SD = 1.14$). Thus, the latency manipulation was generally noticed, however, less strongly than in Study 1. Ratings for whether the system seemed to be thinking before responding were similarly spread across the scale as in Study 1 ($M = 2.85, SD = 1.26$), see Figure 4.9.

Voice naturalness was mixed; the interaction was rarely perceived as a real conversation.

Figure 4.10 shows responses to four items addressing the impressions about naturalness of the voice, interaction, and animation. Voice ratings were mixed ($M = 3.15, SD = 1.40$), with similar proportions agreeing (44.1%) and disagreeing (41.2%) that the voice sounded natural and human-like. Most participants did not feel that the interaction was similar to a real conversation ($M = 2.09, SD = 1.31$), with 73.5% disagreed or strongly disagreed. As can be seen, ratings of whether the AI animation during speaking felt lively were mixed ($M = 2.88, SD = 1.32$). However, the animation was not perceived as distracting ($M = 1.68, SD = 0.91$).

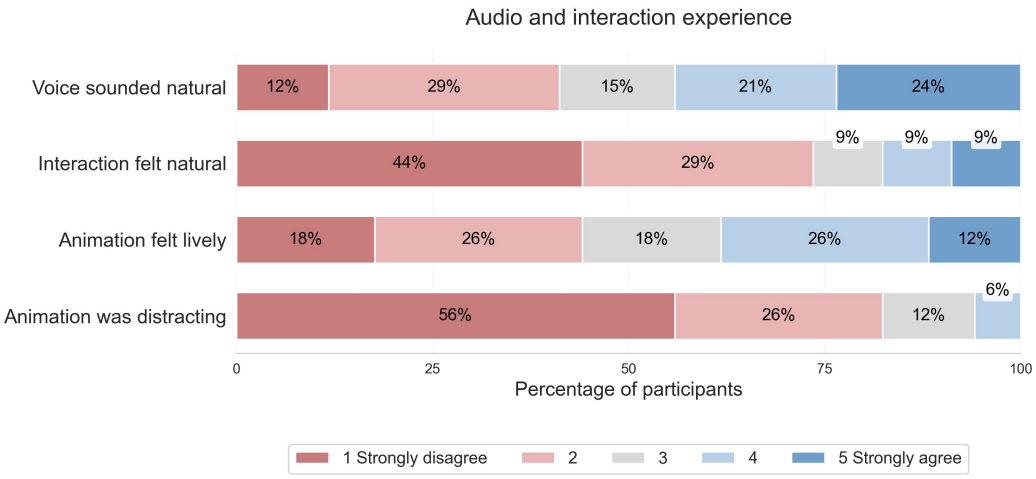

Figure 4.10: Distribution of responses to the self-report items about voice and animation of AI in Study 2. The scale ranged from 1 (strongly disagree) to 5 (strongly agree).

Qualitative Feedback

As in Study 1, open questions provided additional context to the quantitative results.

Prior knowledge. As in Study 1, most participants described relying on their existing knowledge when available, otherwise falling back on plausibility or intuition. Several participants also noted that their familiarity with a question's topic shaped how confident they felt in their judgement.

> *"If I knew the answer for sure: obvious choice. Otherwise: plausibility check. Does the answer make sense in the context of the stuff I do know about the topic? If I had no idea, I selected low confidence."* (P99)

Answer length. Participants again disagreed on how answer length related to credibility. Some associated shorter, more direct answers with higher plausibility:

> *"If the answer was presented in shorter form and was straight to the point, the answer seemed more likely than a long answer with the same amount of information."* (P96)

Others reported the opposite, becoming more sceptical of longer answers: *"Longer answers were sounding more made up"* (P94).

Time pressure. The time limit for a decision was again perceived as too short by some participants. In Study 2, there was an additional challenge of processing the spoken response while keeping their eyes on the written question: *"[I was] listening to the audio while simultaneously reading the question again. It also was harder to get the answer only by listening without reading it too."* (P111)

The AI's voice. Unlike in Study 1, several participants commented on the voice as a factor shaping their trust in the response: *"The AI voice sounded trustworthy, and it was easy to listen to."* (P113). For some, the voice's confidence appeared to compensate for their own uncertainty about the answer's content:

> *"A confident AI voice made me evaluate answers as correct rather than incorrect, even though I did not know an exact answer."* (P98)

Chapter 5

General Discussion

This thesis investigated whether response latency influences users' beliefs about AI-generated content and whether this effect depends on question complexity. In Study 1, the interaction was presented in a text-based chat interface. In Study 2, the same question-answer pairs and delay manipulations were replicated in a voice-based setting, where participants read the questions aloud and received spoken answers with a synthesised human-like voice. In both online studies, participants evaluated answers from a simulated AI system after either a short or a long delay by deciding whether they believed each answer and how confident they were in that judgement. In this chapter, we bring together the findings from both experiments and discuss what they indicate about the role of response time in Human-AI communication.

5.1 Overview of Findings

We expected latency effects to depend on question complexity.

We hypothesised that the meaning of response latency would depend on the question complexity. Short delays were expected to be more beneficial for simple questions, whereas longer delays were expected to be more beneficial for complex questions.

Study 1 showed the predicted pattern descriptively, but not significantly.

In both studies, this predicted interaction was not statistically confirmed for belief rates and confidence ratings. However, the descriptive patterns differed between the two studies. In Study 1, the belief rate means followed the hypothesised direction. Answers to simple questions were believed slightly more often after a short delay, whereas answers to complex questions were believed slightly more often after a long delay than after a short one. Therefore, the belief rate trend was descriptively consistent with H1.1 and H1.2, although the overall interaction was not significant, and the confidence ratings did not show the same pattern.

Study 2 showed a more general preference for short delays.

In Study 2, this pattern did not replicate. Instead, short delays produced higher belief rates for both simple and complex questions. This means that the pattern for simple trials remained partly consistent with H2.1, whereas the pattern for complex trials did not support H2.2. The analysis demonstrated a more general preference for shorter response latency in the voice-based setting. Therefore, H2 was not supported.

Long delays may signal effort.

One possible explanation for the differences in the hypothesised pattern in Study 1 and Study 2 is that response latency can be interpreted at least in two ways. In the text-based chat interface, a long delay after a complex question may have appeared more appropriate, given that the system was processing a difficult request. This interpretation is consistent with the effort heuristic, according to which people may evaluate outcomes more positively when they believe that more time and effort were invested [Efendić et al., 2020]. It also relates to Kim et al. findings, that the advantage of fast ChatGPT responses depended on information complexity, and slower responses could be interpreted as signs of effortful processing [Kim et al., 2025]. Similarly, Tan et al. showed that longer LLM response latencies could be associated with higher thoughtfulness and usefulness [Tan et al., 2026].

Long delays can violate expectations of fast system responses.

In contrast, a long delay after a simple question may have violated users' expectations of fast computational processing. Prior HCI work shows that longer system delays can disrupt interaction flow [Nielsen, 1993; Shneiderman, 1984]. Recent chatbot work similarly suggests that

fast responses are especially beneficial for routine confirmation requests, and longer delays need additional context to be interpreted positively [Kim et al., 2025, 2026]. Thus, a long delay may not automatically signal effort, it can also be read as unnecessary waiting or reduced system efficiency.

Voice interaction might make long delays feel more like interruptions, while short delays feel more fluent.

In Study 2, the voice mode may have shifted the meaning of latency. A long delay of six seconds may have been experienced as less than a reasonable processing time and more as an interruption in the communication flow. Even though the study did not involve a natural conversation with multiple turns, the spoken interaction could still have created stronger expectations of typical rapid turn-taking [Kendrick and Torreira, 2015; Stivers et al., 2009]. Therefore, short delays may have made the system feel more fluent for both simple and complex trials.

Confidence depended more on prior knowledge and plausibility than on latency.

Response latency did not have a strong effect on the confidence ratings. In both studies, confidence was relatively stable across short and long delays. Participants often made quick decisions to accept or reject an answer, but their confidence seemed to depend more on factors such as prior knowledge and the wording of the answer. The self-report measures support this: in both experiments, prior knowledge was rated as the strongest influence on belief judgements, while response time was rated as much less influential. Participants' comments showed the same trend. Many reported relying primarily on what they already knew or heard and, when they lacked prior knowledge, on intuition or perceived plausibility.

Delay differences were more pronounced in the text-based study than in the voice-based study.

At the same time, participants did notice the latency manipulation. In both studies, most participants agreed that the system responded at different speeds. This was especially pronounced in Study 1, where the mean rating was higher (+0.7) than in Study 2. In the text-based interface, participants waited for the written answer to appear. In Study 2, the delay was followed by a spoken answer and an animated voice interface, which may have drawn attention more strongly to the voice, the audio content that cannot be reread, and the AI agent's representation.

Latency & complexity mattered most when answers were difficult to judge from content alone.

The exploratory believability bin analyses help clarify under which conditions response latency may have played a role. By grouping trials by their overall belief rate across all participants, we examined whether the predicted pattern was more visible for answers that were neither clearly accepted nor rejected by the sample. In Study 1, the expected pattern was most visible in the medium-believability bin, with a belief rate range of 55.6% - 76.5%. This suggests that timing cues, such as response latency, had little room to affect participants' decisions when they already knew the answer or found it clearly implausible. Thus, latency can act as a relevant cue in conditions of uncertainty when content-based judgement is less straightforward. The median-split analyses in Study 1 showed a comparable descriptive tendency. The predicted pattern was clearer within participants with lower confidence ratings and slower reaction times. This suggests that the latency cue was more important for participants who needed more time to make a decision and were less certain about their judgements.

Study 2 suggested limited exploratory support for the medium-bin pattern.

In Study 2, however, the medium bin showed the expected direction only for simple questions, while belief rates for complex questions were almost identical across the two delay conditions. Thus, the exploratory support for H3 was limited. At the same time, the medium bin was still the only bin in Study 2 in which part of the predicted pattern appeared. Similarly, the median-split analyses suggested a general short-delay advantage in some subgroups, particularly within participants with higher confidence or slower reaction times.

Exploratory findings should be interpreted descriptively rather than confirmatory

These exploratory findings should be interpreted descriptively rather than confirmatory. A medium belief rate does not necessarily mean that each participant was uncertain about that trial. It only indicated that the trial produced less agreement across the sample. Additionally, for the median-split analyses, the subgroup patterns should also be interpreted more as descriptive tendencies. The relevant group × complexity × latency interactions were not significant, and the predicted complexity × latency pattern did not reach significance. Nevertheless, the exploratory analyses suggest that latency effects may appear in conditions

of uncertainty, but these effects may differ between the text-based and voice-based settings.

Response latency affected participants' decision-making more than their final judgements.

The time-related measures suggest that response latency affected participants' evaluation process, even though it did not influence their final belief judgements. In Study 1, participants took slightly longer to make a belief decision and to rate their confidence after long delays. In Study 2, these main delay effects were not significant. However, post-audio decision times were descriptively slightly longer after long delays. This suggests that the system's timing may have changed participants' own response behaviour. A related perspective comes from work by Bogon et al. on the cognitive integration of system delays, which shows that users can adapt their own actions to the temporal properties of a system response [Bogon et al., 2025].

The high belief rates point to a general tendency toward overtrust in AI answers.

Apart from latency manipulation, both studies showed a general tendency to believe AI-generated answers. Even though participants were informed before the experiment that half of the answers would be incorrect. Nevertheless, the overall belief rate was 64.7% in Study 1 and 68.2% in Study 2. The warning about the system's accuracy did not fully affect participants' belief judgements. Instead, participants were more likely to accept the AI's answers rather than reject them. This fits with earlier research showing that AI-generated responses can seem fluent and convincing even when they contain incorrect information [Tamanna et al., 2024]. It also connects to concerns about overtrust in AI, as users may rely on AI-generated information regardless of whether the content is incomplete, misleading or difficult to verify [Jung et al., 2024; Klingbeil et al., 2024]. The slightly higher belief rate in Study 2 may also relate to the voice mode. Previous research found that adding voice to interfaces with LLM can increase anthropomorphism and perceived information accuracy [Cohn et al., 2024]. In our study, some participants also reported that the AI voice sounded trustworthy and confident. However, this difference should be viewed with caution, as the two studies used independent samples.

Another consistent pattern in both studies was that complex question-answer pairs tended to receive higher belief

Higher belief and confidence in complex answers may reflect answer/audio length.

rates than simple ones. Confidence ratings were also a bit higher for complex than for simple trials in both studies. This should not be interpreted as an effect of question complexity alone. In our study, complex questions required longer and more detailed answers, which might have made them seem more informative or convincing. Qualitative feedback from participants supports this idea. Some participants reported that longer explanations made answers seem more reliable. At the same time, other participants described the opposite, and became more sceptical when answers were too detailed. Thus, the length and detail of the answers may have influenced how plausible they appeared. As a result, the higher belief rates for complex trials could be caused by a combination of question complexity, answer length, and perceived informativeness.

5.2 Implications

Response latency should be understood as a context-dependent interaction cue.

Overall, the findings show that response latency does not have a single, fixed meaning. Its impact can change depending on the interaction context, the type of question, and the user's ability to evaluate the answer quality. Thus, designers of AI systems should not view response time only as a technical performance metric that needs to be minimised. Response latency can affect how an AI system is perceived, even when the content of the answer stays unchanged.

The effort heuristic may be weaker when delays are experienced directly.

Most earlier research on the effort heuristic used in their studies written descriptions of response times [Efendić et al., 2020] or pre-recorded videos [Kim et al., 2025; Zhang et al., 2024], rather than the response latencies people actually experienced. By testing delays that participants actually experienced during the interaction, this thesis suggests that the effort heuristic may be weaker and more context-dependent than this earlier work implies.

Delay may act as a fallback cue under uncertainty.

The exploratory analyses discussed above also suggest that response latency seems to work as a fallback cue, used when prior knowledge and plausibility of the answer are

not enough to decide whether an answer is correct [Jung et al., 2024].

Voice interaction may strengthen expectations for fast conversational timing.

The difference in belief rates between text and voice settings suggests that voice interaction has stronger expectations for conversational timing [Reeves and Nass, 1996; Stivers et al., 2009], which may explain the preference for short delays in Study 2. This also supports the idea that people apply human conversation rules to computers, especially when the interaction feels more human-like, as with voice [Reeves and Nass, 1996].

Interface cues and answer length can shape trust beyond answer accuracy.

The findings also point out the need to be aware of interaction cues that can influence trust in AI-generated content. Participants showed a general tendency to believe AI-generated answers despite knowing half were wrong. Simply telling people an AI can make mistakes does not seem to be enough [Tamanna et al., 2024]. Interfaces should integrate more obvious signals to reveal uncertainty directly, instead of relying on users' critical thinking. Also, the wording and length of an answer may influence how credible it seems more than response timing does. However, participants disagreed on whether longer answers were more or less believable.

5.3 Limitations

Several limitations should be considered when interpreting these results.

Simulated AI limits ecological validity.

First, we used a simulated AI with fixed, pre-written answers instead of a live language model to keep content and length constant across conditions and participants. Results may not fully apply to real systems, where response time varies. Real LLM interactions involve multiple follow-up questions, prompt reformulation, and review of previous responses. In our studies, each trial consisted of a single question-answer pair, which reduces ecological validity. Moreover, information about the accuracy distribution may also have influenced the participants' strategies. Some participants reported keeping this distribution in mind.

Online participation reduced experimental control.

Second, the online setting limited our control over participants' behaviour during the experiment. We cannot control and verify, for example, for Study 2, if participants read each question aloud and consistently followed the procedure. In addition, participants completed the studies using their own devices and browsers, which may have introduced small differences in timing, audio playback, or the visual presentation of animation.

Samples were small and homogeneous.

Third, both samples were relatively small ($n = 36, n = 34$), recruited through personal networks, and mostly young, well-educated, with a majority of native German speakers completing tasks in English.

The studies may have been underpowered.

Fourth, the studies showed no significant interactions between response latency and complexity, and several subgroup analyses used small groups ($n = 17 - 18$). Therefore, the studies may have been too small to detect a real effect.

Time limits may have changed participants' strategies.

Fifth, the 13-second time limit in Study 1 and the 7-second time limit in Study 2 made many participants feel rushed, especially for complex, long answers. This may have changed how much they relied on latency or other cues compared to a more relaxed interaction setting.

Complexity was confounded with answer length.

Sixth, question complexity was confounded with answer length. Complex answers were about twice as long as simple ones. Thus, the effects of complexity may depend not only on the question complexity alone, but also on answer length, or the amount of details provided.

Stimulus selection involved subjective decisions.

The question-answer pairs were created and refined using LLMs, then selected and reviewed by the researchers. Therefore, they may not fully mirror how real AI systems would answer in open interactions. The classification of questions as simple or complex, and the plausibility of answers, were also subjective decisions that may have influenced the stimulus set.

Only two delays were tested.

Finally, we tested only 400 ms and 6000 ms delay durations. It is therefore possible that other latencies would have produced different effects.

Chapter 6

Summary and Future Work

6.1 Summary and Contributions

The thesis tested response latency and question complexity in text- and voice-based AI interactions.

In summary, this thesis explored whether response latency, the delay between a participant's question and an AI system's answer, influences how much people believe AI-generated content, and whether this depends on question complexity. We conducted two online experiments to test it using a simulated AI system with fixed, pre-written answers, half correct and half incorrect. Study 1 presented the interaction as a text-based chat system. In contrast, Study 2 replicated the same question-answer pairs and delay manipulation in a voice-based setting, with spoken answers using a generated human-like voice. In both studies, participants decided whether they believed each answer after a short 400 ms or a long 6000 ms delay, and rated their confidence in that judgement.

No significant interaction, but latency mattered descriptively under uncertainty.

Both studies did not find a statistically significant interaction between latency and complexity. Still, in Study 1, the predicted pattern's direction matched the hypothesis: short delays led to slightly higher belief rates for simple questions, while long delays led to slightly higher belief rates for complex questions. In Study 2, participants generally were

more likely to believe answers when they were preceded by a short delay, regardless of question complexity, suggesting a stronger timing expectation in interaction using voice mode. Both studies also showed a high belief rate in AI-generated answers (64.7% and 68.2%), despite knowing half of them were wrong. Confidence ratings were unaffected by latency in both studies. However, our exploratory bin analyses showed that latency, together with complexity, may be most important when participants were uncertain about the correctness of the answers. Additionally, latency also changed how long participants took to make their decisions.

The thesis contributes to understanding how response latency, uncertainty, and modality can shape belief and trust in AI.

This thesis makes several contributions to research in Human-AI interaction and trust. First, we investigated response latency as a cue for belief judgements in factual AI-generated answers, rather than only as a factor influencing users' satisfaction or perceived usefulness. This thesis adds evidence to existing concerns about overtrust in AI [Bhat, 2025; Jung et al., 2024; Klingbeil et al., 2024]. Second, participants experienced delays in a live simulated AI chat-based interaction, rather than delays described in text or watching pre-recorded videos [Efendić et al., 2020; Kim et al., 2025; Zhang et al., 2024]. Third, the believability bin analyses helped make the latency-complexity pattern more visible by focusing on trials whose correctness was harder to judge. Finally, by comparing text-based and voice-based interactions, this thesis shows that the meaning of response latency is not fixed but depends on context, interaction modality, question complexity, and users' ability to evaluate answer quality.

6.2 Future Work

Future work should test more delay durations and task contexts.

Future studies should test a wider range of response delay durations. Testing more gradual variations, similar to Tan et al., could help to examine when delays are experienced as effortful, fluent, or disruptive. Another important direction would be to study latency across a broader range of AI interactions and investigate whether it depends on the task context.

More natural multi-turn AI interaction could change the interpretation of delays.

Future research could also explore latency in more natural multi-turn interactions with real AI systems. Repeated exchanges, follow-up questions, and longer interactions may change how users interpret response times and whether it affects users' trust and reliance over time.

Future work should investigate how voice characteristics shape the meaning of pauses.

For voice-based interaction, future studies could investigate how different voices, speaking styles, or perceived confidence may change how users interpret pauses before an answer. Additionally, future work could measure perceived difficulty and uncertainty more directly. This would help to understand when users rely on timing cues the most.

Larger and more diverse samples could reveal individual and cultural differences.

Larger and more diverse samples would further make it possible to investigate individual differences, such as gender, language background, and cultural expectations about timing. Since Wang and Lo found that younger and older adults may prefer different response times [Wang and Lo, 2025], age may be especially relevant. Future studies could also be conducted in participants' native languages to reduce language-related uncertainty.

Appendix A

Exploratory Median-Split Analyses

The following tables and figures report exploratory median-split analyses for belief rates in both studies.

A.1 Study 1

Split	Group	Simple-Short	Simple-Long	Complex-Short	Complex-Long
Confidence	Low	57.92	56.94	63.22	68.28
	High	67.02	63.21	70.66	70.61
Timeout	Low	63.24	60.67	64.83	70.34
	High	61.69	59.48	69.05	68.55
Reaction time	Fast	61.60	62.49	65.76	68.24
	Slow	63.33	57.66	68.12	70.65

Table A.1: Mean belief rates (%) for each median-split group and experimental condition in Study 1.

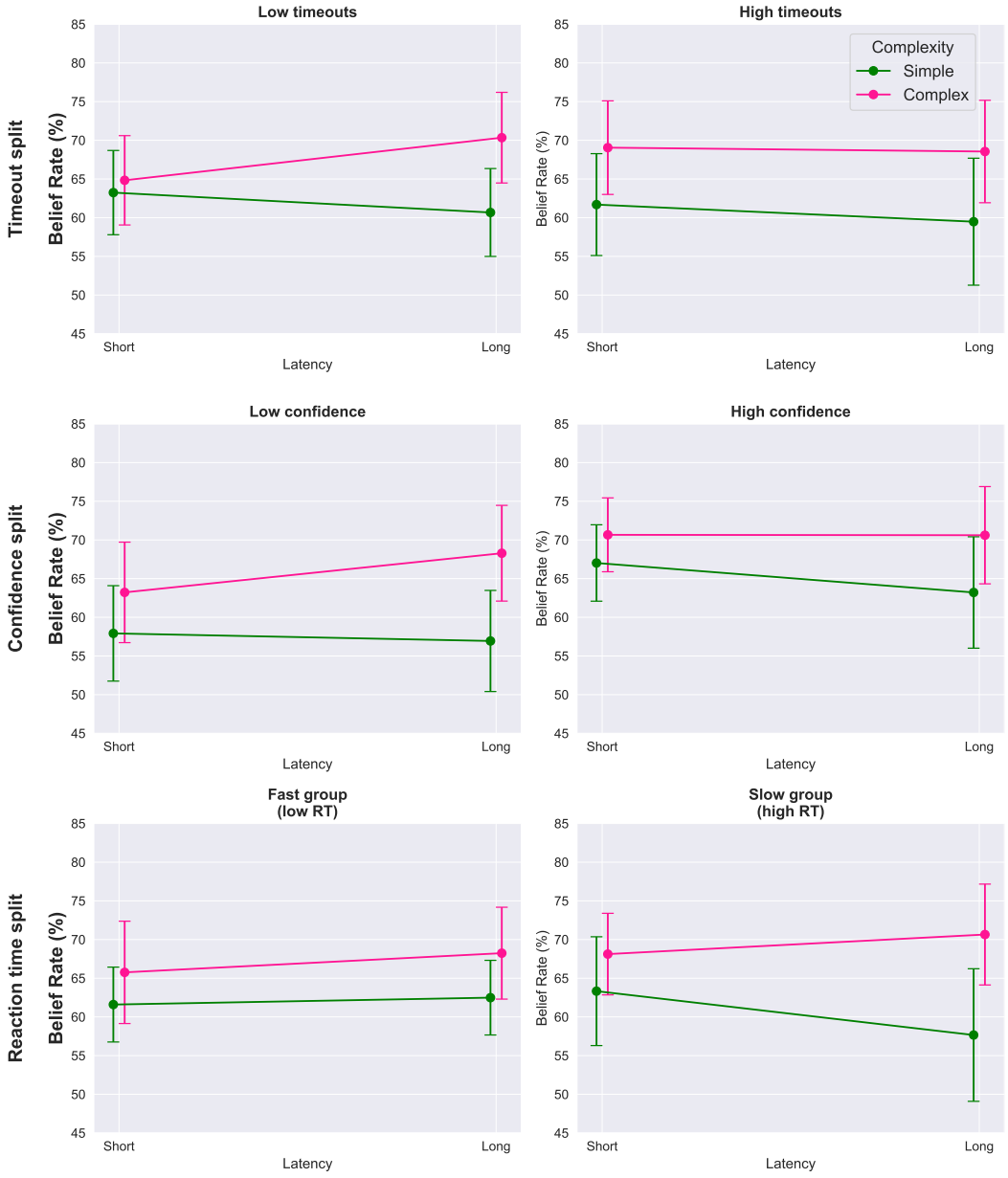

Figure A.2: Interaction plots of belief rate by latency condition and complexity for three median split analyses in Study 1 (timeout, confidence, reaction time). Left column shows the low group, right column shows the high group.

A.2 Study 2

Split	Group	Simple-Short	Simple-Long	Complex-Short	Complex-Long
Confidence	Low	64.77	64.06	68.21	67.16
	High	69.20	66.67	75.20	70.22
Timeout	Low	67.62	69.61	67.76	68.42
	High	66.36	61.12	75.64	68.95
Reaction time	Fast	68.78	70.58	72.12	72.12
	Slow	65.20	60.15	71.29	65.25

Table A.3: Mean belief rates (%) for each median-split group and experimental condition in Study 2.

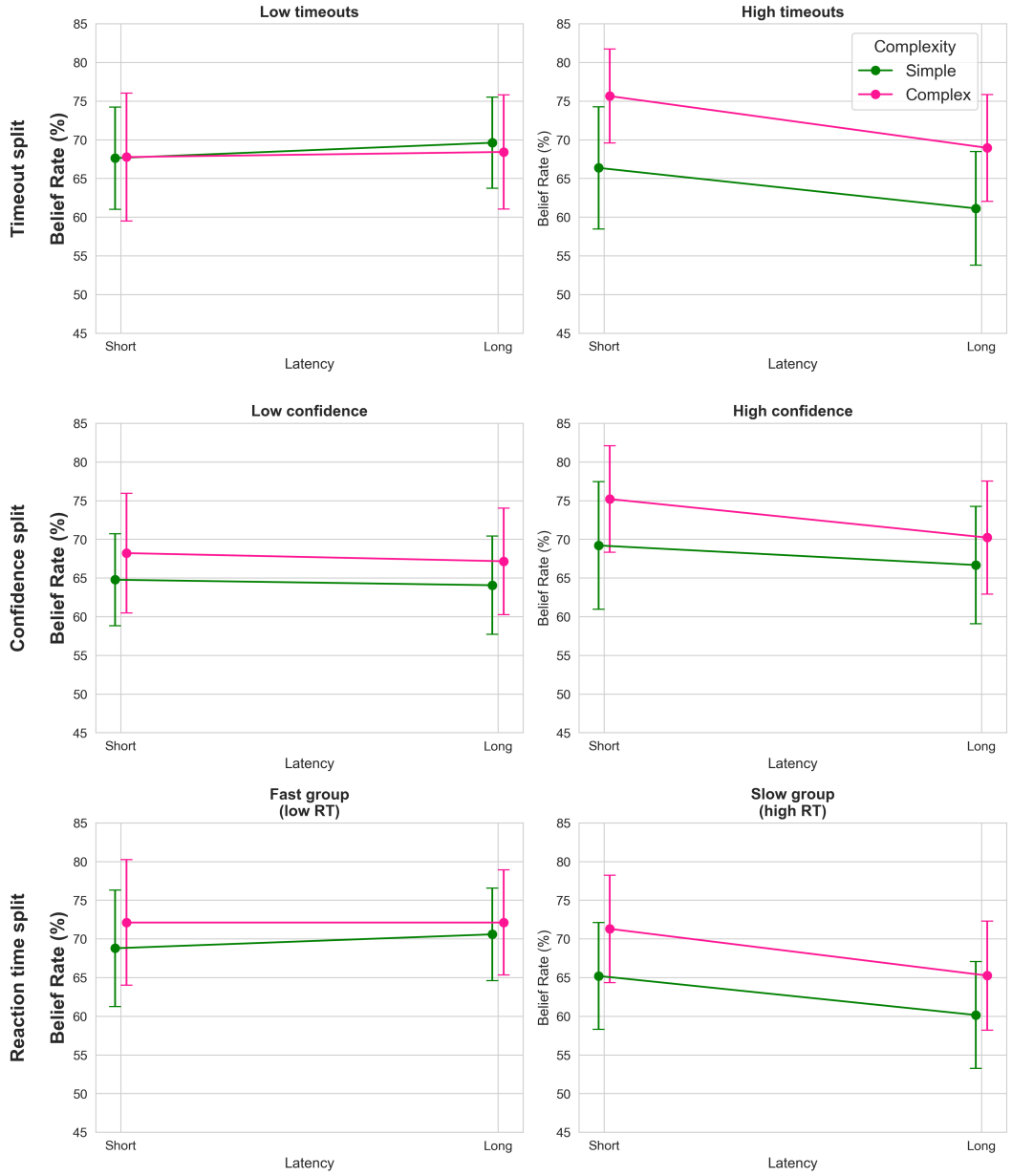

Figure A.4: Interaction plots of belief rate by latency condition and complexity for three median split analyses in Study 2 (timeout, confidence, reaction time). Left column shows the low group, right column shows the high group.

Bibliography

[1] Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024.

[2] Maalvika Bhat. How Dynamic vs. Static Presentation Shapes User Perception and Emotional Connection to Text-Based AI. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 846–860, New York, NY, USA, 2025. Association for Computing Machinery. doi.org/10.1145/3708359.3712131.

[3] Johanna Bogon, Sabrina Hößl, Christian Wolff, Niels Henze, and David Halbhuber. Cognitive Integration of Delays: Anticipated System Delays Slow Down User Actions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. doi.org/10.1145/3706598.3713475.

[4] Michelle Cohn, Mahima Pushkarna, Gbolahan O. Olanubi, Joseph M. Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery. doi.org/10.1145/3613905.3650818.

[5] Jim Dabrowski and Ethan V Munson. 40 years of searching for the best computer system response time. *Interacting with Computers*, 23(5):555–564, 2011.

[6] Andreea Danielescu, Sharone A Horowit-Hendler, Alexandria Pabst, Kenneth Michael Stewart, Eric M Gallo, and Matthew Peter Aylett. Creating Inclusive Voices for the 21st Century: A Non-Binary Text-to-Speech for Conversational Assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. doi.org/10.1145/3544548.3581281.

[7] Mark Dingemanse and Andreas Liesenfeld. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology.

In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5614–5633, 2022.

[8] Emir Efendić, Philippe P.F.M. Van de Calseyde, and Anthony M. Evans. Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, 157:103–114, 2020. doi.org/10.1016/j.obhdp.2020.01.008.

[9] Romina Etezadi and Mehrnoush Shamsfard. The state of the art in open domain complex question answering: a survey. *Applied Intelligence*, 53(4):4124–4144, 2023.

[10] Ulrich Gnewuch, Stefan Morana, Marc T. P. Adam, and Alexander Maedche. Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. In *26th European Conference on Information Systems: Beyond Digitization - Facets of Socio-Technical Change, ECIS 2018, Portsmouth, UK, June 23-28, 2018. Ed.: U. Frank*, page Code 143975, 2018.

[11] Kashyap Haresamudram, Ilaria Torre, Magnus Behling, Christoph Wagner, and Stefan Larsson. Talking body: the effect of body and voice anthropomorphism on perception of social agents. *Frontiers in Robotics and AI*, 11:1456613, 2024.

[12] T.M. Holtgraves, S.J. Ross, C.R. Weywadt, and T.L. Han. Perceiving artificial social agents. *Computers in Human Behavior*, 23(5):2163–2174, 2007. doi.org/10.1016/j.chb.2006.02.017.

[13] Martin Huschens, Martin Briesch, Dominik Sobania, and Franz Rothlauf. Do you trust ChatGPT?–perceived credibility of human and AI-generated content. *arXiv preprint arXiv:2309.02524*, 2023.

[14] Joseph Jaffe, Beatrice Beebe, Stanley Feldstein, Cynthia L Crown, Michael D Jasnow, Philippe Rochat, and Daniel N Stern. Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the society for research in child development*, pages i–149, 2001.

[15] Yongnam Jung, Cheng Chen, Eunchae Jang, and S. Shyam Sundar. Do We Trust ChatGPT as much as Google Search and Wikipedia? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery. doi.org/10.1145/3613905.3650862.

[16] Kobin H Kendrick and Francisco Torreira. The timing and construction of preference: A quantitative study. *Discourse Processes*, 52(4):255–289, 2015.

[17] Jungkeun Kim, Jeong Hyun Kim, Seongseop (Sam) Kim, Chulmo Koo, and Namho Chung. Is a shorter reaction time always better? Empirical investigation of the impact of response speed on ChatGPT recommendations. *International Journal of Hospitality Management*, 130:104239, 2025. doi.org/10.1016/j.ijhm.2025.104239.

[18] Kaeun Kim, Ghazal Shams, and Kawon Kim. From seconds to sentiments: differential effects of chatbot response latency on customer evaluations. *International Journal of Human–Computer Interaction*, 42(1):597–612, 2026.

[19] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report RBR-8-75, 1975.

[20] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 160:108352, 2024. doi.org/10.1016/j.chb.2024.108352.

[21] Justin Kruger, Derrick Wirtz, Leaf Van Boven, and T.William Altermatt. The effort heuristic. *Journal of Experimental Social Psychology*, 40(1):91–98, 2004. doi.org/10.1016/S0022-1031(03)00065-9.

[22] Siru Liu, Aileen P Wright, Barron L Patterson, Jonathan P Wanderer, Robert W Turer, Scott D Nelson, Allison B McCoy, Dean F Sittig, and Adam Wright. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *Journal of the American Medical Informatics Association*, 30(7):1237–1245, 2023.

[23] Maria Madsen and Shirley Gregor. Measuring human-computer trust. In *11th australasian conference on information systems*, volume 53, pages 6–8. Citeseer, 2000.

[24] Mykola Maslych, Mohammadreza Katebi, Christopher Lee, Yahya Hmaiti, Amirpouya Ghasemaghaei, Christian Pumarada, Janneese Palmer, Esteban Segarra Martinez, Marco Emporio, Warren Snipes, and others. Mitigating response delays in free-form conversations with LLM-powered intelligent virtual agents. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pages 1–15, 2025.

[25] Theresa Matzinger, Michael Pleyer, and Przemysław Żywiczyński. Pause length and differences in cognitive state attribution in native and non-native speakers. *Languages*, 8(1):26, 2023.

[26] Shira Michel, Sufi Kaur, Sarah Elizabeth Gillespie, Jeffrey Gleason, Christo Wilson, and Avijit Ghosh. "It's not a representation of me": Examining Accent

Bias and Digital Exclusion in Synthetic AI Voice Services. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 228–245, New York, NY, USA, 2025. Association for Computing Machinery. doi.org/10.1145/3715275.3732018.

[27] Robert B Miller. Response time in man-computer conversational transactions. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, pages 267–277, 1968.

[28] Youngme Moon. The Effects of Physical Distance and Response Latency on Persuasion in Computer-Mediated Communication and Human–Computer Communication. *Journal of Experimental Psychology: Applied*, 5(4):379–392, 1999.

[29] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, 1994.

[30] Jakob Nielsen. Response times: the three important limits. *Usability Engineering*, 1993.

[31] Arie Wahyu Prananta, Nugroho Susanto, Jean Henry Raule, and others. Transforming education and learning through Chat GPT: A systematic literature review. *Jurnal Penelitian Pendidikan IPA*, 9(11):1031–1037, 2023.

[32] Byron Reeves and Clifford Nass. The media equation: How people treat computers, television, and new media like real people. *Bibliovault OAI Repository, the University of Chicago Press*, 10(10):19–36, 1996.

[33] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735, 1974.

[34] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. Effects of gesture on the perception of psychological anthropomorphism: a case study with a humanoid robot. In *International conference on social robotics*, pages 31–41. Springer, 2011.

[35] Emanuel A Schegloff and Harvey Sacks. Opening up closings. *Semiotica*, 8(4): 289–327, 1973.

[36] Lawrence M Schleifer and Benjamin C Amick III. System response time and method of pay: Stress effects in computer-based tasks. *International Journal of Human-Computer Interaction*, 1(1):23–39, 1989.

[37] Eric Schurman and Jake Brutlag. Performance related changes and their user impact. In *velocity web performance and operations conference*, 2009.

[38] William Seymour and Max Van Kleek. Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), October 2021. doi.org/10.1145/3479515.

[39] Yike Shi, Qing Xiao, Qing Hu, Hong Shen, and Hua Shen. The Siren Song of LLMs: How Users Perceive and Respond to Dark Patterns in Large Language Models. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. doi.org/10.1145/3772318.3791149.

[40] Ben Shneiderman. Response time and display rate in human performance with computers. *ACM Comput. Surv.*, 16(3):265–285, September 1984. doi.org/10.1145/2514.2517.

[41] Maria Grazia Sindoni. The femininization of AI-powered voice assistants: Personification, anthropomorphism and discourse ideologies. *Discourse, Context Media*, 62:100833, 2024. doi.org/10.1016/j.dcm.2024.100833.

[42] Laura Spillner, Johanna Rockstroh, Paola Raquel Peña, Nina Wenig, Nima Zargham, and Benjamin R. Cowan. Beyond the Illusion: LLMs and the Case for Pragmatic Cues in Conversation. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, CUI '25, New York, NY, USA, 2025. Association for Computing Machinery. doi.org/10.1145/3719160.3737614.

[43] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, and others. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009.

[44] Salma Begum Tamanna, Gias Uddin, Song Wang, Lan Xia, and Longyu Zhang. ChatGPT Inaccuracy Mitigation during Technical Report Understanding: Are We There Yet? *arXiv preprint arXiv:2411.07360*, 2024.

[45] Felicia Fang-Yi Tan, Moritz Alexander Messerschmidt, Wen Yin, and Oded Nov. The Impact of Response Latency and Task Type on Human-LLM Interaction and Perception. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. doi.org/10.1145/3772318.3790716.

[46] Emma M Templeton, Luke J Chang, Elizabeth A Reynolds, Marie D Cone LeBeaumont, and Thalia Wheatley. Long gaps between turns are awkward for strangers but not for friends. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1875):20210471, 2023.

[47] Torrey Trust, Jeromie Whalen, and Chrystalla Mouza. Editorial: ChatGPT: challenges, opportunities, and implications for teacher education. Contemp. Issues Technol. Teacher Educ. 23 (1), 1–23.: Society for Information Technology & Teacher Education, Waynesville, NC USA. *Contemporary Issues in Technology and Teacher Education*, 23(1):1–23, March 2023.

[48] Nick von Felten, Luisa Ella Müller, and Johannes Schöning. AI Washing Inflates Expected Performance but Not Interaction Outcomes: An AI Placebo Study Using Fitts' Law. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, page 1120–1138, New York, NY, USA, 2026. Association for Computing Machinery. doi.org/10.1145/3805689.3812249.

[49] Francesco Walker, Matteo Favetta, Linde Hasker, and Richard Walker. They Prefer Humans! Experimental Measurement of Student Trust in ChatGPT. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery. doi.org/10.1145/3613905.3650955.

[50] Joseph B Walther. Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication research*, 19(1):52–90, 1992.

[51] Ya-Ling Wang and Chi-Wen Lo. The Effects of Response Time on Older and Young Adults' Interaction Experience with Chatbot. *BMC Psychology*, 13:1–12, 2025.

[52] Ruiyun Xu, Yue Feng, and Hailiang Chen. Chatgpt vs. google: A comparative study of search performance and user experience. *arXiv preprint arXiv:2307.01135*, 2023.

[53] Mohammad Yani and Adila Alfa Krisnadhi. Challenges, Techniques, and Trends of Simple Knowledge Graph Question Answering: A Survey. *Information*, 12(7), 2021. doi.org/10.3390/info12070271.

[54] Zhengquan Zhang, Konstantinos Tsiakas, and Christina Schneegass. Explaining the Wait: How Justifying Chatbot Response Delays Impact User Trust. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24, New York, NY, USA, 2024. Association for Computing Machinery. doi.org/10.1145/3640794.3665550.

[55] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. Anthropomorphism: opportunities and challenges in human–robot interaction. *International journal of social robotics*, 7(3):347–360, 2015.

Index

Typeset July 5, 2026