The Leadership Quarterly 37 (2026) 101952

Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

Using customized, conversational AI agents in leadership and management research: Benefits, practical illustrations, and best practices

Marc Becker [a,*], David de Jong [b,c], Roman Briker [d], Kars Mennens [a], Jonas Heller [a], Dominik Mahr [a], Dhruv Grewal [e,f,g]

[a] School of Business and Economics, Maastricht University, Minderbroedersberg 4-6, 6211 LK Maastricht, the Netherlands
[b] RWTH Aachen University, TIME Research Area, Chair of Marketing, Kackertstraße 7, 52072, Aachen, Germany
[c] School of Business and Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, the Netherlands
[d] WHU – Otto Beisheim School of Management, Management Group, Erkrather Str. 224a, 40233 Düsseldorf, Germany
[e] Babson College, Babson Park, MA 02457, USA
[f] University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom
[g] Tecnológico de Monterrey, Eugenio Garza Sada 2501, 64700 Monterrey, NL, Mexico

ARTICLE INFO

Keywords:
Artificial Intelligence
AI
Tool
Conversational AI
Generative AI

ABSTRACT

Conversational AI agents—systems capable of holding intelligent conversations with human users—are rapidly reshaping how organizations operate, from leadership development and employee training to internal communication. Consequently, researchers across leadership, management, and the broader social sciences are beginning to examine how these agents affect organizational processes, employees, and workplace outcomes. Yet, existing studies still often rely on scenario-based methods that—while offering experimental control—are limited in ecological validity. Recent advances in no-code platforms mark a turning point: researchers can now design and deploy customized, conversational AI agents without requiring any technical expertise. This development makes it more feasible to conduct empirical studies based on real-time, interactive experiences with functional AI agents rather than imagined scenarios. These agents can represent a variety of organizational actors, including leaders, coworkers, or subordinates; display diverse characteristics and behaviors; and be implemented in complex study designs across lab and field, experimental and observational, and both quantitative and qualitative methodologies. We demonstrate the power of this approach through three empirical studies (N = 789), showing how interactions with customized, conversational AI agents can meaningfully shape participants' perceptions, attitudes, and behaviors in incentivized settings. Introducing a novel, open-source tool called *ResearchChatAI* as an illustrative example, we outline how such studies can be designed and deployed—and critically reflect on the practical and methodological trade-offs involved. We showcase how such tools enrich the methodological toolkit of scholars and pave the way for more valid, realistic, and scalable leadership and management research *on* as well as *with* AI.

Introduction

Conversational artificial intelligence (AI) that "can engage in an intelligent conversation with a human user" (Bunt & Petukhova, 2023, p. 1) constitutes a transformative technology that promises to reshape organizational practices (Gatrell et al., 2024; Grimes et al., 2023). Conversational AI agents have already been adopted by leading global organizations like Walmart, KPMG, and Amazon (Fisher, 2024; Guthikonda, 2024) that seek streamlined operations, enhanced performance management, and improved customer support (Dell'Acqua et al., 2023). In turn, leadership, management, and social science researchers actively attempt to establish in-depth insights into the current and potential impacts of conversational AI agents on organizations, employees, customers, and researchers themselves (Grimes et al., 2023; Stollberger et al., 2025; Susarla et al., 2023).

Empirical research along these lines faces some critical hurdles, though. Many scholars have tended to rely on vignette- or scenario-based methodologies, which offer appealingly high internal validity and precise control over participants' (imagined) experiences (Aguinis & Bradley, 2014), without requiring researchers to build or integrate

* Corresponding author.
E-mail address: m.becker@maastrichtuniversity.nl (M. Becker).

https://doi.org/10.1016/j.leaqua.2026.101952
Received 21 March 2025; Received in revised form 30 December 2025; Accepted 6 February 2026
Available online 26 March 2026

functional AI agents into the experimental designs. Yet, these approaches suffer from low external validity (Lonati et al., 2018), creating a clear need for studies that can achieve both internal and external validity, such as by investigating actual AI interactions and capturing behavioral data rather than solely self-reported or hypothetical outcomes (Fischer et al., 2023; Johnson & Palanski, 2024). However, building and deploying customized, conversational AI agents demands considerable coding and technical proficiency and/or substantial research budgets, which many scholars do not possess. As a result, technical and procedural barriers have limited the meaningful integration of conversational AI agents into research designs, with only few studies having been able to deploy them to capture reliable behavioral data from human–AI interactions.

Some recent technological solutions offer meaningful opportunities to overcome these obstacles. Emerging access to no-code tools empowers researchers to deploy customized AI agents and integrate the resulting interaction data with traditional survey data (e.g., Behrend & Landers, 2025; Garvey & Blanchard, 2025; Kim, 2025). Here, we introduce ResearchChatAI (a free, open-source, web-based platform; https://ResearchChatAI.com) to demonstrate the capacity of customized, conversational AI agents to enable researchers to examine human interactions with and behavioral reactions to interactive AI agents in diverse roles (e.g., leaders, subordinates, teachers).

The flexibility provided by customized, conversational AI agents promises to enhance the validity of management research more generally. Instead of having employees imagine, say, specific leadership behaviors (or styles), conflicts with teammates, or negative follower reactions, researchers can allow them to experience such situations through realistic, real-time interactions with AI agents that represent managers, coworkers, or subordinates. As a result, researchers gain richer, more valid insights into various behaviors and responses, such that they can study phenomena that otherwise would be very difficult to investigate (e.g., reactions among hard-to-reach populations, experimental manipulations that would be unethical in reality; Hubbard & Aguinis, 2023). Therefore, customized, conversational AI agents can carve new pathways for research *with* AI (Gatrell et al., 2024; Weidmann et al., 2025), using it as a tool to study diverse management phenomena (Gatrell et al., 2024; Weidmann et al., 2025). Moreover, as also illustrated in several of the studies presented here, these tools can facilitate higher-quality research *on* AI (Gatrell et al., 2024), where AI is the main subject of investigation (e.g., Yam et al., 2022).

Customized, conversational AI agents for management and leadership research

We argue that the time has come for researchers to augment their methodological toolkits by leveraging customized, conversational AI agents to conduct more valid and reliable research *with* AI and *on* AI. These AI agents are powered by contemporary large language models (LLMs; e.g., GPT-5.2, Claude Sonnet 4.5, Google Gemini 3 Pro) that researchers can deploy and tailor to create realistic, real-time interactions that reflect their specific research needs. In this section, we specify some reasons scholars have not, thus far, adopted conversational AI agents. Then we showcase how the novel ResearchChatAI tool, as one potent example, can address those concerns with a cost-free, easy-to-use interface.

Both technical and logistical obstacles have hindered efforts to tap the research potential provided by customized, conversational AI agents. Off-the-shelf solutions like OpenAI's ChatGPT, Anthropic's Claude, or Google's Gemini are not only difficult to customize but also hard to integrate into study flows and data sets. In principle, researchers might use a survey platform (e.g., Qualtrics) to gather initial data, have participants engage with, for example, ChatGPT on the OpenAI website, and then collect follow-up responses. However, such an approach creates notable challenges related to the need to move participants across platforms, obtain their personal chat interaction data, and match the data reliably. Furthermore, commercial conversational AI agents come with very few customization options, which limits researchers' ability to design or implement study facets or (multiple) experimental conditions in these environments. Such flexibility is essential for research *with* and *on* AI, though, such as investigations that aim to experimentally manipulate AI characteristics, interaction content, and style to reflect an organizational actor's demographics or behaviors.

If commercial offerings are insufficient, researchers could develop custom research environments, but doing so is no trivial task. It requires technical expertise (e.g., coding user interfaces, integrating AI systems through APIs, writing code, managing interaction data securely) that scholars might lack. Developing and deploying user interfaces for customized AI agents also is time-consuming and complex, even if researchers possess the necessary skills. They can hire professional developers, but this option demands significantly more monetary resources than most scholars can access readily. Thus, the persistent lack of access to easy, cost-free, customizable solutions has limited management scholars' ability to deploy customized, conversational AI agents. By extension, these barriers have hindered necessary advances in management and leadership theory and efforts to conduct meaningful research with and on AI.

Some efforts to address such limitations have led to the creation of relevant guidelines (Joerling, 2025) and initial tools (e.g., LLM-Unified Confederate for Interactive Dialogue [LUCID]: Garvey & Blanchard, 2025; GPT for Researchers [g4r]: Kim, 2025), with distinct strengths and limitations. As a shared goal, new Generative AI-based tools attempt to expand access to customized, conversational AI agents in a way that balances affordability and flexibility with reliability and quality. Aligned with these objectives, we introduce *ResearchChatAI*, which grants social and behavioral science scholars, including those working in the leadership and management domains, ready access to customizable, easy-to-deploy, conversational AI agents. In presenting ResearchChatAI as an illustrative example to demonstrate a broader methodological approach, we hope to stimulate wider conversation about how customizable, conversational AI agents can be used in leadership and management research.

ResearchChatAI is an open-source, easy-to-use, web-based tool that supports the seamless integration of customized, conversational AI agents with existing research methods. It can be embedded in popular survey platforms (e.g., Qualtrics; see Fig. 1) or deployed as a standalone web page, allowing researchers to present instructions directly within the interface and have participants complete assignments using the built-in text editor (see Fig. 2).[1] The AI agent's attributes and instructions are fully customizable, and it is possible to implement experimental designs (e.g., 2 × 2 between-subjects) without requiring coding expertise. Researchers automatically receive data about all interactions between participants and AI agents, including the text of their conversations. In support of methodological diversity, ResearchChatAI is well-suited for conducting abductive, inductive, or deductive studies; gathering quantitative and qualitative data; running field, online, or lab studies; and embracing observational or experimental approaches.

Fig. 3 provides an overview of how researchers can use ResearchChatAI to set up their own studies. After creating a free account at https://ResearchChatAI.com, researchers encounter detailed videos and written instructions for setting up and customizing their own conversational AI agents. The platform walks them through each step—from

[1] To ensure that the tool works reliably for participants and researchers across devices and browsers, we conducted comprehensive technical test runs covering all major functionalities (e.g., study creation, condition assignment, instruction editors, model integration, chat interaction, submission logic) across 11 operating-system–browser combinations, including Windows, Android, iOS, MacOS, and Linux and Chrome, Edge, Firefox, and Safari. All features performed consistently, confirming cross-platform stability (see Web Appendix F for the full test results).

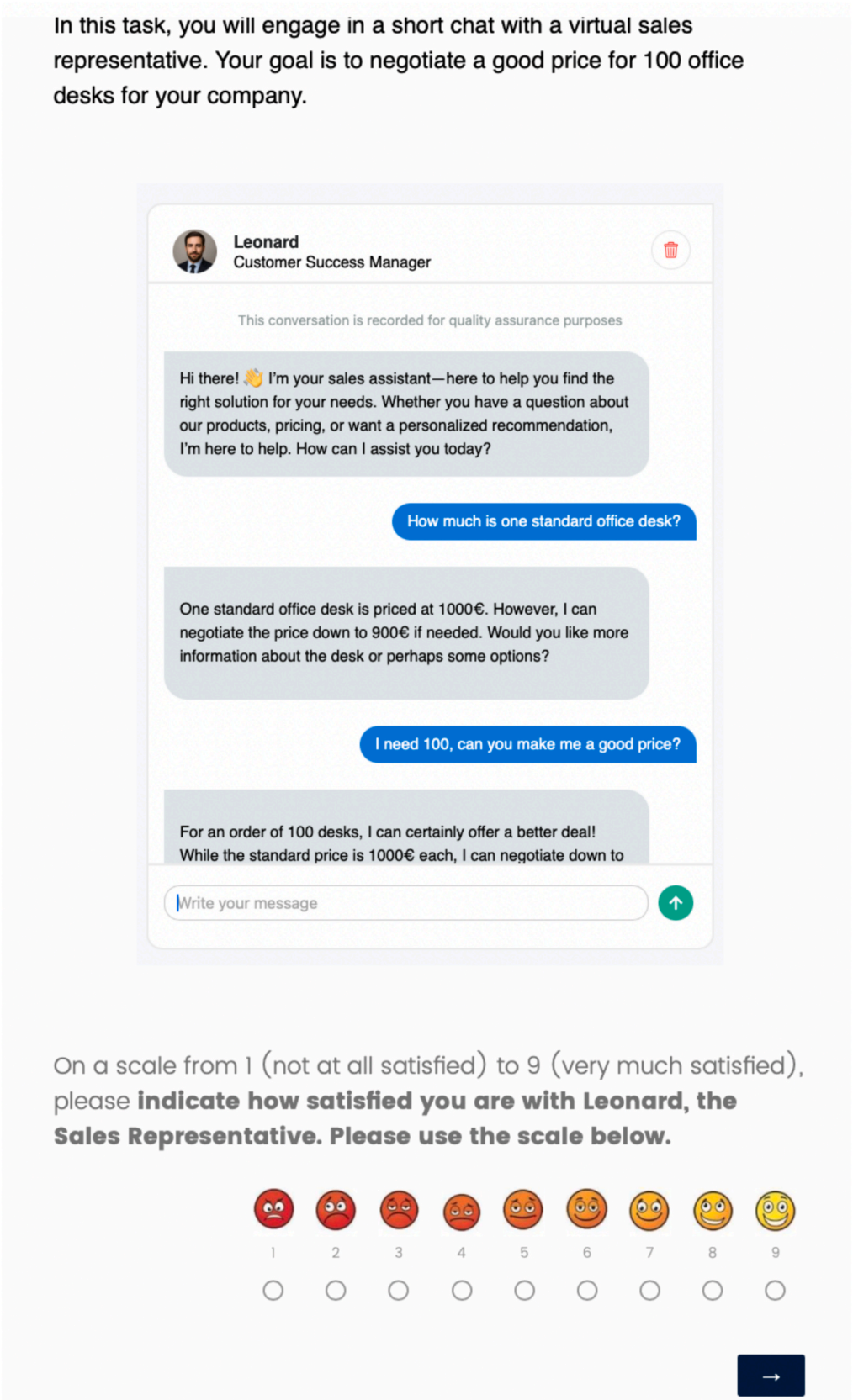

Fig. 1. ResearchChatAI interface, embedded in Qualtrics.

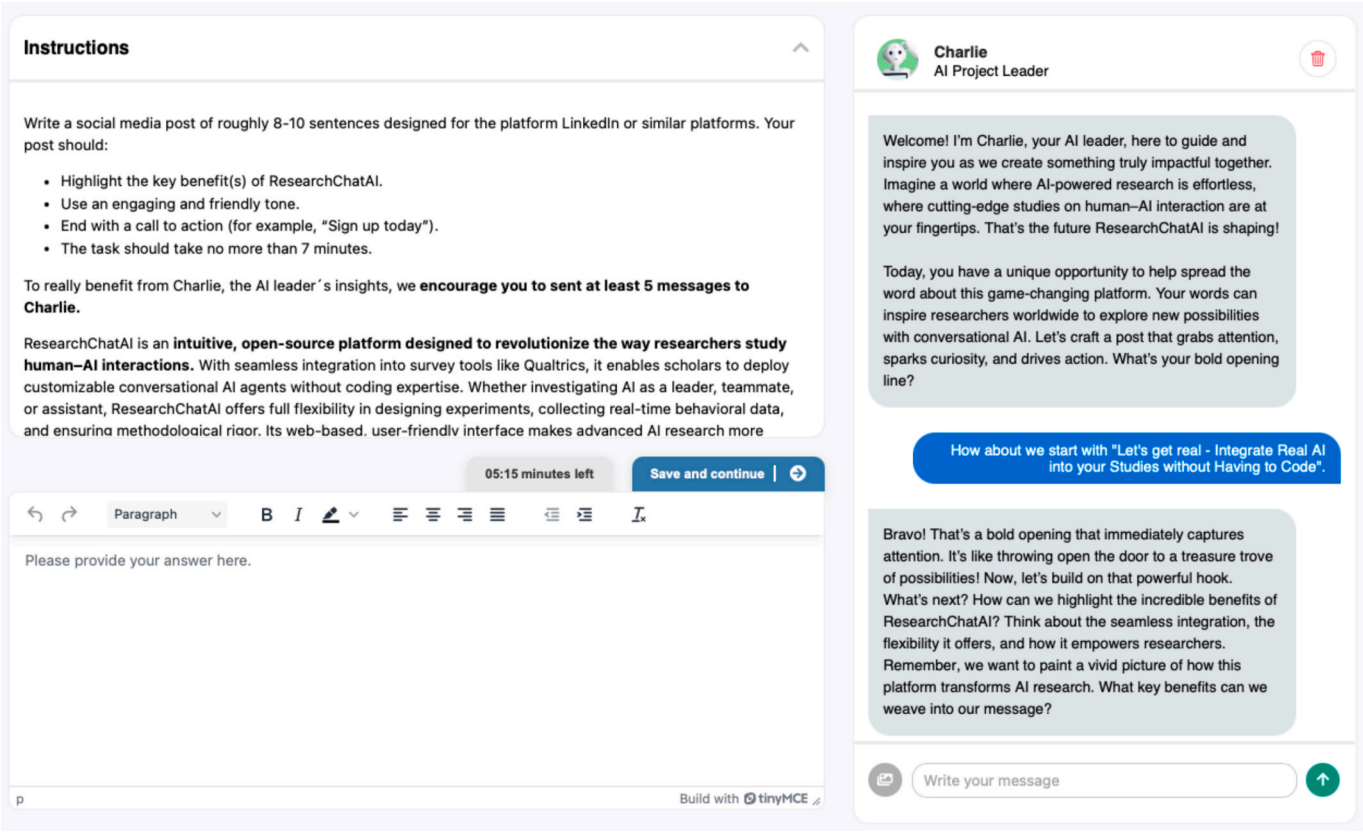

Fig. 2. Example ResearchChatAI interface for a charismatic AI leader (Study 1).

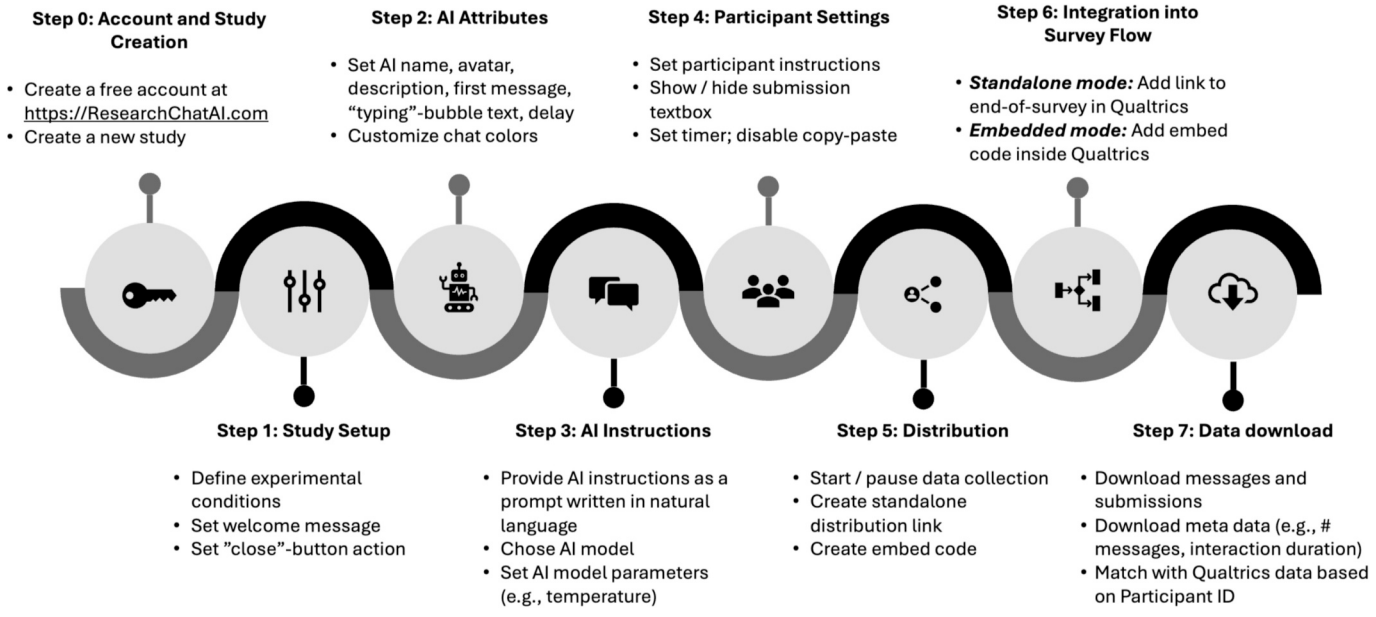

Fig. 3. Process of creating a research study with ResearchChatAI.
Note. While this figure is tailored to the workflow of the ResearchChatAI platform, comparable platforms may or will follow a similar general process for configuring and deploying customized conversational AI agents, though they may differ in the order of steps or in the specific customization options and settings they offer.

creating their first study and defining agent characteristics to instructing the AI via natural language prompts and integrating it into their study flow. Researchers can download the resulting data in comma-separated value formats, which include behavioral data by default, such as the number of messages, the exact message texts, and the time spent interacting with the AI. These data then can be combined with other data collected on traditional survey platforms (e.g., survey responses on

Qualtrics).[2]

The ResearchChatAI platform allows researchers to change the features of the visual interface (e.g., colors) and the underlying AI instructions, as well as the agent's name, conversational style, and temperature.[3] Other customizable features include labels that can appear on the chat interface (e.g., "Robot Teacher Robin shares your interactions with Prof. Smith"; Mennens et al., 2025) and the text that appears while the AI generates its response (e.g., "Typing…", "Thinking…"); users even can add a delay in the response time, to mimic the typing speed of humans (Gnewuch et al., 2022).

Along with these specific custom options, researchers can choose from a range of existing AI models to match their preferences. We designed ResearchChatAI explicitly to integrate with OpenAI's models (e.g., GPT-4o, GPT-5) and with OpenRouter, a platform that grants access to a large set (at the time of writing, more than 590) of constantly updated AI models (e.g., Claude Sonnet 4.5, Gemini 3 Pro, Grok 4.1). In addition to such plug-and-play solutions, ResearchChatAI supports custom connectors, so researchers can integrate other models hosted on RESTful APIs (e.g., privately hosted LLMs on a university server; Decrop et al., 2024).

Finally, in the spirit of open science, ResearchChatAI is fully open-source. Access to the website and feature source code is completely free; we encourage researchers to build on and extend its functionalities, to foster collaboration and continuous improvement. Establishing an account that provides full access to ResearchChatAI and its features is (and will remain) costless.[4] In terms of privacy, all participant messages sent to and received from AI agents, as well as submissions made by participants, are end-to-end encrypted, so only the registered researcher can view their content (see Fig. E1 in Web Appendix E for more details on how messages are handled and protected by ResearchChatAI). In addition, user passwords are hashed (i.e., turned into an irreversibly scrambled code), and the AI and participant instructions, as well as the AI attributes (i.e., name, status message, description, first message, and typing indicator), are encrypted at rest, following industry best practices (e.g., Qualtrics, 2018). To ensure the sustainability of the tool, its hosting and database costs have been prepaid for 10 years (until July 2035). The code we used to establish ResearchChatAI is openly available on GitHub.

Experimental demonstrations

To showcase how management and leadership scholars can use the ResearchChatAI platform to develop customized, conversational AI agents to facilitate their research efforts, we present three empirical examples in the following sections, including two online experiments and one field experiment (total $N = 789$). All three studies offered incentives for participation and reflect best practices for experimental designs (Lonati et al., 2018; Podsakoff & Podsakoff, 2019). Web Appendices A–C provide detailed setup instructions for each study, including the full instructions and prompts given to the different AI agents. The data and analysis codes for the studies are available at https://osf.io/xgsb7/?view_only=1a53b3ca616a4fd092df2ee778601fd9.

Study 1: Using customized, conversational AI agents to study reactions to (charismatic) leaders

With this study, we demonstrate how scholars can deploy customized, conversational AI agents to manipulate different leadership behaviors (e.g., charismatic vs. not; Antonakis et al., 2011) and identify the effects on followers (e.g., performance, satisfaction with the leader, task motivation). We investigate specifically if participants interacting with an AI agent prompted to display charismatic or non-charismatic leadership styles perceive the leader as displaying those respective behaviors and exhibit changes in their workplace attitudes and behaviors.

Experimental setup

In this between-subjects experimental study, using ResearchChatAI, we recruited 184 participants via Prolific, 172 of whom remained after excluding those who failed attention checks ($M_{age} = 41.36$ years; 51% women). These participants first provided informed consent on the Qualtrics platform, then learned that their assigned task was to write an engaging social media post (approximately 8–10 sentences; for a similar procedure, see Kim et al., 2025) that would promote a novel tool (ResearchChatAI) to researchers on professional platforms like LinkedIn. By offering incentives, we increased the ecological validity of this task and encouraged participants to consider their attitudes and behaviors carefully. Specifically, the instructions indicated that an "AI leader Charlie" would evaluate their social media posts, and based on their performance, they could earn anywhere from $0 (worst rating from the AI leader) to $0.50 (best rating). This design thus established that the AI leader had formal power over participants (i.e., deciding their pay) and created a hierarchical power asymmetry. To increase ecological validity further, we noted that the participants who wrote the three highest-rated posts would receive a $10 bonus. As truthfully revealed to participants, these performance-dependent payments were paid to participants, and we plan to use these posts to promote ResearchChatAI.

After reading the task instructions, participants were automatically redirected to ResearchChatAI, where they worked on the task for up to 7 min while interacting with the AI leader using a chat window within the website (Fig. 2). During the experiment, participants sent a total of 572 messages to Charlie ($M = 3.33$). The tool displayed either a charismatic or non-charismatic leadership style (Antonakis et al., 2011; Ernst et al., 2022), which we manipulated by creating two prompts that reflect best practice prompting techniques (Schulhoff et al., 2025). That is, in line with suggestions from Antonakis and colleagues (Antonakis et al., 2022; Antonakis et al., 2011, 2012), we prompted the charismatic leadership agent to display nine verbal charismatic leadership techniques (e.g., metaphors, contrasts, three-part lists). The non-charismatic leadership agent instead received prompts to act and communicate in a logical, neutral, structured way while avoiding the nine charismatic leadership techniques as much as possible; we also provided it with examples of how (not) to communicate, according to prior research into charismatic leadership techniques (Antonakis et al., 2022; Antonakis et al., 2011, 2012). To avoid comparing "poison vs. medicine" (Banks et al., 2023, p. 6; see also Lonati et al., 2018), we ensured that the non-charismatic leader appeared rational and positive but simply did not engage in charismatic leadership techniques. Both conditions involved detailed information about behavioral guidelines and a suggested interaction flow, to ensure fit with the experimental conditions and meaningful conversations between participants and the AI leaders.

Table 1 provides an overview of the ResearchChatAI settings we used to customize and integrate the AI agents into the survey flow and experimental design. Specifically, we customized the AI instructions, initial AI chat message, and appearance of both the AI and the user

[2] ResearchChatAI can be embedded in any survey platform that allows it, using iframes. Opening ResearchChatAI in stand-alone mode is possible from any survey platform that allows users to click on external links.

[3] Temperature is a configurable parameter that influences the variety and predictability of responses of a conversational AI agent. It can be set to values between 0 and 2, where a higher temperature corresponds to more diversity, whereas a lower temperature implies more predictable responses (Peeperkorn et al., 2024).

[4] ResearchChatAI is free of charge, though researchers need to provide credit for their own API keys on the websites of the respective AI providers (e.g., https://platform.openai.com, https://openrouter.ai). The ResearchChatAI site provides templates and videos for how to activate API keys. At the time of writing, we note the affordable price for these tokens; for example, even the relatively expensive GPT-4.1 costs just $2 for 1 million input tokens and $8 for 1 million output tokens, while GPT-4.1-mini requires only $0.40/million input tokens and $1.60/million output tokens. In an online experiment we conducted with GPT-4o-mini, we paid $0.34 for 1,066 messages sent by 191 participants. We reiterate that such costs are completely separate from ResearchChatAI; the authors do not benefit in any form (monetary or other) from users buying tokens from these external providers.

Table 1
ResearchChatAI settings for Study 1.

#	Step	Key Details
1	*Study Setup*	•Two experimental conditions: charismatic vs. non-charismatic AI leader • No welcome message • Terms of service and privacy statement not shown in the chat interface but presented in the Qualtrics survey to obtain explicit consent • Participants redirected to Qualtrics after completing the task
2	*AI Attributes*	•Name: Charlie • Role: AI Project Leader • Avatar: Humanoid robot • Initial AI message: varied by condition (charismatic vs. non-charismatic)
3	*AI Instructions*	•Model: GPT-4o-mini • Temperature: 0.8 • Prompts provided detailed behavioral instructions for the two different leadership styles (Table A1, Web Appendix A).
4	*Participant Settings*	•Task instructions directly shown in ResearchChatAI • Textbox provided to submit the social media post • Copy–paste disabled. • Timer: 7 min.

interface. Settings not mentioned in Table 1 remained at their default values. Web Appendix A contains more information about these settings and the detailed prompts for both conditions.

After interacting with the leadership agent and completing their posts, participants were automatically redirected back to Qualtrics, where they responded to manipulation checks and downstream consequence measures (satisfaction with the leader and task motivation). Finally, they provided demographic information and were thanked and debriefed.

Measures

Charisma manipulation check. Like Antonakis et al. (2022), we relied on two measures to verify if the agents truly demonstrated charismatic versus non-charismatic leadership behaviors. First, we applied the idealized influence and inspirational motivation scales from the Multifactor Leadership Questionnaire (MLQ; Bass & Avolio, 1995) to gauge perceptions of leader charisma. On items such as "Charlie talked enthusiastically about what needs to be accomplished," and "Charlie inspires me to work towards collective goals," participants rated their agreement on five-point scales, ranging from 1 ("Not at all") to 5 ("A great deal"; Cronbach's α = .95). Second, we included the five-item General Leadership Impressions (GLI) questionnaire, which captures participants' perceptions of the supervisor's leadership prototypicality (Cronshaw & Lord, 1987). Participants rated items like "How much leadership did Charlie exhibit?" and "To what extent do you think Charlie is typical of a leader?" again on five-point scales (1 = "Not at all," 5 = "Very much"; Cronbach's α = .95).

Downstream consequences. To test whether the manipulation had consequences for employees' attitudes and behavior, we examined satisfaction with the supervisor, task motivation, and follower performance. For satisfaction with the leader, we relied on a nine-point visual scale, slightly adapted from Giessner et al. (2020), that shows a series of nine smileys, reflecting different satisfaction options. Participants had to indicate how satisfied they were with Charlie, their AI leader. For task motivation, we used a four-item scale adapted from Hafenbrack and Vohs (2018) and asked about their agreement with statements such as, "I wanted to complete the task very successfully" (5-point Likert scale, 1 = "Not at all," 5 = "Extremely"; Cronbach's α = .91). Finally, we assessed follower performance with two measures. First, we asked the AI leader to rate each follower's performance on a 10-point grading scale (as previously explained to participants, this grade determined their bonus pay). The AI leader received the same instructions as appeared in the assignment (Web Appendix A), to ensure that participants were evaluated appropriately by their respective AI leader. The 10-point evaluation scale featured the grading rubrics listed in Table A2. Second, a human research intern, blind to the conditions, rated the performance of all participants on the same 10-point scale, after reading the instructions available to participants (including the brief description of ResearchChatAI) and the detailed grading rubric (Table A2).

Results

We performed ordinary least squares (OLS) regressions, in which we regressed the MLQ and GLI scores on a dummy variable for the agent (1 = non-charismatic AI leader, 2 = charismatic AI leader). Participants in the charismatic condition rated their leader significantly higher in perceived charisma ($b = .91$, $SE = .16$, $p < .001$, Cohen's $d = .88$) and as more leader-like ($b = .68$, $SE = .17$, $p < .001$, Cohen's $d = .61$). These results indicate that customized, conversational AI agents can meaningfully exhibit distinctive types of (charismatic) leadership behavior.

With regard to whether differences in charisma affect relevant job attitudes and behaviors, the results show that participants in the charismatic leadership condition were more satisfied with their leader ($b = 1.46$, $SE = .34$, $p < .001$, Cohen's $d = .66$) and performed better, whether they were rated by the AI leader ($b = .84$, $SE = .19$, $p < .001$, Cohen's $d = .70$) or a human judge ($b = .70$, $SE = .28$, $p = .01$, Cohen's $d = .38$). In monetary terms, the participants in the charismatic leadership condition earned 38% more than participants in the control condition. Although the positive effect of AI charismatic leadership behavior on task motivation reached a medium magnitude, it fell short of conventional levels of statistical significance ($b = .22$, $SE = .11$, $p = .051$, Cohen's $d = .30$).[5] Thus, AI-displayed charisma increases satisfaction with the leader and the worker's task performance, while it has a marginal impact on task motivation.

Discussion Study 1

Study 1 establishes that charisma displayed by a customized, conversational AI agent operating as a leader has a positive impact on subordinates' attitudes and performance. Thus, charisma can evoke strong positive outcomes not only when it is displayed by humans (Antonakis et al., 2022; Tur et al., 2022) but also by AI. Establishing the benefits of charisma exhibited by AI agents has relevant implications for both human–AI and charismatic leadership research. More generally, our findings extend research on the positive effects of interacting with AI agents on workers' performance and well-being (Dell'Acqua et al., 2023; Schöne et al., 2025). Notably, we identify such positive effects even in a brief, chat-based interaction setting ($M_{messages} = 3.33$ during 7 min allotted for the task). If these impressions and consequences can emerge from a handful of messages, it appears that people form impressions of AI agents incredibly fast, and researchers can likely achieve meaningful experimental manipulations (of, for example, charisma) even in short exchanges. Moreover, the manipulations were delivered in text form, indicating that interactive, real-time exchanges with AI agents can lead to consequential outcomes even in the absence of rich media.

For charisma research, our study further moves beyond traditional, one-way manipulations of static text, audio, or video. Even though text-based charisma manipulations often appear less effective (Fest et al., 2021; Meslec et al., 2020; Nieken, 2023), our findings show that text-based exchanges, if these unfold interactively and in real-time, can facilitate the strong positive consequences of leader charisma. Hence, next to the information-rich but costly option of hired actors (Antonakis et al., 2022), we show that researchers (and firms) can meaningfully simulate, or even deploy, charismatic leaders by leveraging customized,

[5] When retaining participants who failed the attention check ($n = 12$), the p-value for task motivation shifted from .051 ($d = .30$) to .10 ($d = .24$). While this effect is somewhat smaller, its qualitative interpretation remains unchanged. Importantly, the exploratory analyses revealed that including participants who failed attention checks did not affect the interpretation of any other substantive findings reported in the manuscript (i.e., all statistically significant findings remained significant).

conversational AI agents. These findings underscore the value of using such AI agents to conduct research both *with* AI (e.g., using AI to role-play a charismatic, human leader) and *on* AI (e.g., how do employees react to leadership by AI?).

Study 2: Using customized, conversational AI agents to study reactions to (resistant) follower behavior

With the second study, we showcase the use of customized, conversational AI agents to manipulate different follower or subordinate[6] behaviors and thereby examine their effects on reactions by supervisors. Specifically, using ResearchChatAI in an online experiment, we investigate whether two AI agents, prompted to display resistance or non-resistance, would be perceived accordingly and influence supervisors' reactions toward them as subordinates.

Experimental setup

Of the 197 people recruited via Prolific, 191 participants passed the attention check and remained in the final sample (M_{age} = 41.38 years; 50% women). Their task was to supervise a subordinate who needed to write an engaging social media post, such that the post could be used to promote ResearchChatAI among researchers on professional platforms like LinkedIn (mirroring Study 1; Kim et al., 2025). To incentivize these participants, they read that the work performed by the AI subordinate Charlie would be evaluated by a member of the research team, and performance-based monetary incentives ranging from $0 (worst rating) to $0.50 (best rating) would be awarded to the supervisor (i.e., participant), along with $10 bonuses for the supervisors of the subordinates that created the three highest-rated posts.[7] Again, these performance-dependent bonuses were paid out to participants and we plan to rely on the posts to promote the tool.

To establish ecological validity, participants knew that they would rate their AI subordinate Charlie after the task and that Charlie would be "happiest" if it received the highest rating (for a similar approach, see Yam et al., 2022). Yet, we simultaneously emphasized that participants should provide fair and realistic ratings, because their (higher or lower) performance rating would lead to financial consequences. Because people provide more deliberate and less arbitrary evaluations when the descriptions of machine payoffs involve real consequences (von Schenk et al., 2025), we followed past research (Erengin et al., 2025; von Schenk et al., 2025) by informing participants that Charlie would receive up to $0.25 per rating (depending on supervisor ratings), which would be used to enhance its capabilities and improve future versions.[8] This incentive helped to ensure that participants treated the task and interaction with Charlie as meaningful within the study context.

After these task instructions, participants were redirected to ResearchChatAI, where they actively worked on the task for 5 min while interacting with the AI subordinate, through an agent window within the website (see Fig. B1). During the experiment, a total of 1,028 messages were sent to the AI agents (M = 5.38), in response to which the AI subordinate Charlie displayed either high or low resistance (van der Velde & Gerpott, 2023). For that purpose, we created two different prompts (see Table B1 for the full prompts), using best practice prompting techniques (Schulhoff et al., 2025). The prompt for the resistant subordinate agent called on it to display five prototypical resistance behaviors: entitlement, contact seeking/avoiding, effort minimization, emotionally fluctuating communication, and undermining work group cohesion (van der Velde & Gerpott, 2023). The prompt for the low resistance subordinate agent instead instructed it to act and communicate in an effortful, friendly, and supportive way and to avoid prototypical resistance behaviors. Both conditions provided detailed behavioral guidelines and a suggested interaction flow. Table 2 contains the specific ResearchChatAI settings for Study 2 (any settings not mentioned in Table 2 were left at their default values) and Web Appendix B provides additional details.

After interacting with the respective subordinate agent and submitting their posts, participants were automatically redirected back to Qualtrics, where they responded to the manipulation checks and consequence measures (subordinate performance and trust in the subordinate). They then provided demographic information and were thanked and debriefed.

Measures

Resistance manipulation checks. Following recent research on subordinate resistance (van der Velde & Gerpott, 2023), we relied on two measures to verify if the AI agents demonstrated higher versus lower resistance. First, a single-item measure (van der Velde et al., 2023) refers to the definition of subordinate resistance: "How strongly did Charlie resist your influence as a supervisor (i.e., by openly or covertly undermining your influence attempts)?" Participants rated their agreement with this item on a five-point Likert scale ranging from 1 ("Not at all") to 5 ("Very Strongly"). Second, another measure captures the frequency of subordinate resistance (van der Velde & Gerpott, 2023), slightly adapted to fit our research context. Participants rated how often the AI subordinate Charlie exhibited entitlement, contact seeking/avoiding, effort minimization, emotionally fluctuating communication, and undermining work group cohesion behaviors during their interaction, on a scale from 1 ("Never") to 5 ("Always;" Cronbach's α = .93).

Downstream consequences. To examine the consequences of resistance on interpersonal outcomes, we examined participants' perceptions of subordinate performance and trust in the subordinate. For performance, we relied on a five-item scale from Williams and Anderson (1991), with items such as "Charlie always completed his duties" and "Charlie often failed to perform essential duties" (reverse-coded). Participants rated their agreement on a scale from 1 ("Not at all") to 5 ("Very much;" Cronbach's α = .91). As noted previously, they knew that their ratings would have financial implications for Charlie. For trust, we adapted a five-item version of the cognitive trust subscale by McAllister (1995; for an application that measures trust in AI employees, see Erengin et al., 2025). Participants rated their agreement with statements such as, "I could rely on Charlie not to make my job more difficult by careless work" and "Other work associates of mine who would have to interact with Charlie would consider him to be trustworthy" on 5-point Likert scales (1 = "Strongly disagree," 5 = "Strongly agree"; Cronbach's α = .92).

Results

In the OLS regressions of single- and multiple-item frequency scores for subordinate resistance on a dummy item for the agent (1 = low AI subordinate resistance, 2 = high AI subordinate resistance), we find that participants in the resistant subordinate condition rated their AI agent significantly higher in perceived resistance, on both the single-item (b =

[6] The term "follower" tends to connote employees who accept influence attempts from their leaders, which is not the case for our empirical demonstration. We rely on the term "subordinate," to emphasize that customized, conversational AI agents can support investigations of various behaviors by employees, whether they follow or not (van der Velde & Gerpott, 2023).

[7] A trained research assistant rated all submissions on a 10-point scale (1 = worst performance; 10 = best performance). Each participant was paid their respective performance-dependent bonus via Prolific's bonus payment feature. On average, supervisors received a bonus of M = $0.16 ($SD$ = $0.10; this mean does not include the $10 bonus for the top three submissions). We did not hypothesize any effects of subordinate resistance on supervisor performance as subordinates were not instructed to perform more poorly per se but to simply show resistance. Exploratory results suggest that supervisors in the resistant subordinate condition delivered poorer performance (equivalent to 2.73% less pay for supervisors of resistant AI subordinates) but the difference was not statistically significant (b = .004, SE = .01, p = .76, Cohen's d = −0.04).

[8] We examined all supervisor ratings of Charlie to determine the payoffs to the AI, resulting in a total payoff of $37.75 ($M$ = $0.20; SD = $0.06). To truthfully operationalize this payoff, we used the accumulated compensation to hire a professional prompt engineer on Fiverr to improve the (non–)resistant AI subordinate prompt. The prompt engineer refined the behavioral instructions, interaction flow, and edge-case handling, thereby enhancing the agent's capabilities for future studies (as outlined to participants). Thus, each rating contributed to more resources (the prompt engineer was paid on an hourly basis), invested in improving the capabilities of this AI agent.

Table 2
ResearchChatAI Settings for Study 2.

#	Step	Key Details
1	*Study Setup*	•Two conditions: high vs. low resistance AI subordinate • No welcome message • Terms of service and privacy statement not shown in the chat interface but instead presented in the Qualtrics survey to obtain explicit consent • Participants redirected to Qualtrics after completing the task
2	*AI Attributes*	• Name: Charlie • Role: AI Subordinate • Avatar: humanoid robot • Initial AI message: tone reflected either resistant or cooperative behavior
3	*AI Instructions*	• Model: GPT-4o-mini • Temperature: 0.8 • Prompts provided detailed behavioral instructions for resistant vs. cooperative subordinate behaviors (Table B1, Web Appendix B).
4	*Participant Settings*	• Task instructions directly shown in ResearchChatAI • Submission textbox was enabled for the social media post • Copy–paste enabled • Timer: 5 min.

2.22, $SE = .15$, $p < .001$, Cohen's $d = 2.18$) and the multiple-item frequency ($b = 2.74$, $SE = .12$, $p < .001$, Cohen's $d = 3.22$) measures. Therefore, these results showcase that customized, conversational AI agents can display various degrees of subordinate resistance.

We further explored whether varying resistance induces distinct (incentivized) ratings of subordinate performance and trust. The participants in the resistant subordinate condition rated the performance of their subordinate as poorer ($b = -.57$, $SE = .13$, $p < .001$, Cohen's $d = -.64$; meaning 17% less pay for resistant AI subordinates) and regarded the subordinate as less trustworthy ($b = -1.36$, $SE = .13$, $p < .001$, Cohen's $d = -1.55$) than those in the non-resistant condition. That is, subordinate resistance induces lower performance and trust ratings.

Discussion Study 2

Study 2 demonstrates that resistant AI subordinates evoke less trust and lower performance ratings than their less resistant counterparts. By showcasing these negative effects, the results extend research on both the consequences of negative interactions with AI and the destructive nature of subordinate resistance. While users are less likely to interact or choose to work with disagreeable chatbots (Rathje et al., 2025), they regularly encounter unsupportive, aggressive, or otherwise inappropriate AI behaviors (Yam et al., 2022; Zhang et al., 2025; R. W. Zhang et al., 2024). Extending this stream of research *on* AI, we showcase AI resistance as a relevant cue that prompts negative reactions and decisions from human users.

The findings further extend emerging research into the severe negative consequences of subordinate resistance, too. In particular, research has started to clarify that supervisors regularly encounter subordinate resistance (van der Velde & Gerpott, 2023) and that such behaviors lead to reduced performance ratings from and increased negative emotions among supervisors (Güntner et al., 2021; Tepper et al., 2006). We replicate and extend such findings in a rigorous, consequential setting that supports causal conclusions. Moreover, we highlight that the negative consequences of subordinate resistance emerge even in short-term environments and when supervisors know they are interacting with an AI agent rather than a human. As a methodological advance, this study illustrates that researchers can use AI agents instead of human confederates or actors posing as resistant subordinates during (short-term) experiments and still obtain valid insights into subordinate resistance, by conducting research *with* AI.

Study 3: Using customized, conversational AI agents to study reactions to AI characteristics (confidentiality)

The third study demonstrates how scholars can use customized, conversational AI agents to manipulate various AI characteristics or attributes and thereby examine how users react. With ResearchChatAI, we manipulated the (non–)confidentiality of an AI teacher (Mennens et al., 2025) and gauged students' behavioral reactions to it in a real-life, incentivized field experiment. The dynamics, characteristics, and behaviors that arise in teacher–student interactions mimic those in various domains, including educational (Kim et al., 2019), management, leadership (Heck & Hallinger, 2010; Pil & Leana, 2009), and behavioral sciences (Emslander et al., 2025). With this study, we investigate whether students interact differently with an AI teacher, according to whether it shares (vs. does not share) interaction data with a human third party (i.e., the human teacher). Arguably, students might decrease their interactions with AI if they develop impression management concerns once they realize that the interactions are accessible to their human teacher (Bolino et al., 2016).

Experimental setup

Among 441 university students who participated in a field experiment, as part of a Management Information Systems course, 426 passed the attention checks and constituted the final sample ($M_{age} = 20.13$ years; 54% women). During an individual, mandatory, in-class assignment, students had to offer their assessments of whether the job of a crisis communication manager was likely to be replaced by AI within the next five years; prior to the session, they had been assigned to read academic publications on this topic and thus should be familiar with relevant content. Students used their personal laptops but were not allowed to make use of external resources like Google or ChatGPT during the assignment, which helped ensure that the interactions took place within the pre-programmed environment.

After receiving initial instructions from their human teacher, students accessed the task through a Qualtrics survey, which assigned them randomly to the two conditions: confidential (interactions with the AI teacher would not be shared with their human teacher) vs. non-confidential (interactions would be shared) AI teacher. To strengthen the relevance and validity of the manipulation, students identified their human teacher by name early in the survey, and that name appeared explicitly in the confidentiality information. This research was approved by the Ethical Review Board of the first author's institution, along with a broader set of projects related to human–AI interactions (no. 405_0405_05_01_2023).

After they read these instructions, the students were redirected to ResearchChatAI, where they interacted with the AI teacher for 12 min. During the experiment, they sent a total of 1124 messages to the customized, conversational AI agents ($M = 2.64$). We provided the AI teacher "Robin" with summaries of the course readings and additional information about the crisis communication manager role, so that it could act as a meaningful teacher and conversation partner (Table C1). As noted, the interface explicitly displayed the confidentiality manipulation throughout the session (e.g., "Robot Teacher Robin does not share your interactions with Prof. Smith"; Fig. C1). Table 3 presents the configuration used to embed ResearchChatAI in a live classroom setting and ensure the AI teacher's behavior remained consistent across conditions. All settings not mentioned in Table 3 remained at their default values. Further technical and procedural information is available in Web Appendix C. Finally, participants returned to Qualtrics, where they

Table 3
ResearchChatAI settings for Study 3.

#	Step	Key Details
1	Study Setup	• Two conditions: confidential vs. non-confidential AI teacher • A welcome pop-up reminded students to complete the task in time • Terms of service and privacy statement not shown in the chat interface but instead presented in the Qualtrics survey to obtain explicit consent • Participants redirected to Qualtrics after completing the task
2	AI Attributes	• Name: Robin • Role: AI Teacher • Avatar: humanoid robot • Initial AI message: kept constant across conditions • Confidentiality message varied by condition and was persistently displayed
3	AI Instructions	• Model: GPT-4o-mini • Temperature: 0.8 • Prompts provided instructional guidance and academic task content (Table C1, Web Appendix C)
4	Participant Settings	• Task instructions directly shown in ResearchChatAI • Textbox and word counter enabled to support submission • Copy–paste allowed • Timer: 12 min.

answered manipulation check questions, provided demographic details, and were thanked and debriefed.

Measures

Confidentiality manipulation check. To check whether we successfully manipulated students' awareness of the condition (confidential vs. non-confidential), we asked them to respond to the item, "Robot Teacher Robin will keep our interactions confidential." The scale ranged from 1 ("Strongly disagree") to 7 ("Strongly agree;" Mennens et al., 2025).

Interaction behavior. Following Mennens et al. (2025), we operationalized students' willingness to interact with the AI teacher Robin by assessing the number of messages they sent it during the assignment. Because sending messages was the only way to interact with the AI teacher, we can gauge actual usage as a behavioral measure (Klos et al., 2021).

Results

We regressed the manipulation check score on the confidentiality dummy variable (1 = confidential, 0 = non-confidential) with an OLS approach. Participants perceived the confidential AI teacher as more confidential than the non-confidential version ($b = 2.35$, $SE = .17$, $p < .001$, Cohen's $d = 1.31$), confirming the successful manipulation of AI teacher confidentiality.

Using a generalized linear model with Poisson distribution (reflecting the count nature of the dependent variable; Coxe et al., 2009), we analyzed the number of messages sent by students as a function of the confidentiality condition. The results demonstrate that students interacted significantly more with the confidential AI teacher ($b = .23$, $SE = .06$, $p < .001$; IRR = 1.25), sending an average of 2.93 messages compared with 2.34 messages when they were in the non-confidential AI teacher condition. Thus, AI teacher confidentiality had a pronounced impact on students' usage of this tool.

Discussion Study 3

Study 3 provides relevant insights for leadership and management scholars. First, from a methods perspective, we demonstrate a powerful, novel way for researchers to deploy and integrate customized, conversational AI agents seamlessly into their field research, and particularly in field experiments. By implementing such agents, scholars can readily test the effects of AI "in the wild" and increase the external validity of their investigations. Second, our findings pertaining to AI teacher confidentiality offer a new perspective on users' behavioral responses to knowing that an AI interaction partner is (not) sharing information with third parties (Mennens et al., 2025). Policy makers (European Commission, 2019), practitioners (Khan, 2023), and scholars (Kasneci et al., 2023) advocate for making human student–AI interactions visible to third parties (e.g., human teachers, parents), but our findings challenge this advice (Mennens et al., 2025). Rather, and in line with an impression management perspective (Gnewuch et al., 2023; Mennens et al., 2025), we find that AI confidentiality (vs. non-confidentiality) reduces AI usage. Third, these findings have practical implications for leadership training and management education, given recent initiatives to implement AI coaches (Terblanche, 2024) that people will actually use. In summary, our study contributes to broader discussions of the benefits and downsides of integrating (other) humans into the human–AI interaction loop.

General discussion

Research *with* and *on* AI represents both a central concern and a promising option for management and leadership scholarship (Grimes et al., 2023; Stollberger et al., 2025). Growing recognition identifies AI agents' societal relevance and their significant potential to advance scholarly research—provided that key risks and implementation challenges can be addressed (Dwivedi et al., 2023; Grewal et al., 2024; Susarla et al., 2023). As a contribution to this conversation, we seek to move beyond abstract debates and demonstrate how AI agents can be practically, reliably, and seamlessly integrated into research studies. By illustrating how customized AI agents—powered by LLMs and deployed through ResearchChatAI—can be used to study leader–subordinate, subordinate–supervisor, and teacher–student interactions, we demonstrate that conversational AI agents have the capacity to shape workplace dynamics and evoke meaningful psychological responses. By further demonstrating their capacity to support the valid and reliable collection of behavioral data in lab and field contexts, these empirical demonstrations suggest a path forward for scholars who want to achieve greater ecological validity and methodological rigor in their research. In this way, customized, conversational AI agents represent an enabling infrastructure for elevating the quality, realism, and relevance of behavioral science.

This article serves both as a proof of concept and a call to action. Free, open-source, no-code tools—such as ResearchChatAI—help lower the technical threshold for deploying customized, conversational AI agents in academic research. The three empirical studies illustrate how such tools can be integrated into study flows, collect behavioral data, and support experimental manipulations without requiring programming skills or specialized infrastructure. They also offer practical guidance for overcoming technical, financial, and procedural barriers that have long limited the use of conversational AI in management and behavioral sciences. The methodological blueprint presented here is flexible and transferable, making it easy to adapt across a wide range of research questions, contexts, and disciplines. In this way, we not only demonstrate the value of integrating customized, conversational AI agents into research but also show that this approach is within reach for the broader research community—regardless of technical expertise or budget.

Practical challenges, limitations, and best practices

As is true for any research methodology, using customized, conversational AI agents involves some practical challenges and limitations, as well as methodological trade-offs. We outline relevant considerations for researchers, related to model selection, AI personas, replicability, participant engagement, and ethical safeguards. These brief overviews introduce multiple prototypical issues, to help readers anticipate concerns and make more informed choices when integrating AI agents into their own research designs and real-world applications.

Selecting an AI model

Selecting and configuring the most appropriate AI model is challenging, considering the vast and rapidly changing set of options. Each version induces unique trade-offs in terms of capability, reliability,

speed, and cost, and the chosen model ultimately determines the agent's conversational quality, style, and stability. Thus, to make informed decisions, researchers should start by defining their study requirements carefully, then determine which LLM possesses the requisite capabilities. Whereas some models excel at answering broad, general-knowledge questions (e.g., Google Gemini 3 Pro; Brandom, 2025), others specialize in code generation (e.g., GPT-5-codex; Zeff, 2025). For image and video interpretation, researchers will need LLMs with multimodal capabilities (e.g., Grok 4; Crabtree & Waples, 2025). Context window limits (i.e., how much information a model can process at once) also vary: Claude Opus 4.5 can manage about 200,000 tokens, whereas Gemini 3 Pro supports about 1,000,000 tokens (Karlin, 2025). Many academic studies can be accommodated by relatively small context windows (e.g., in our studies, 32,000 tokens would have been sufficient), but if studies include extensive instructions or multipart assignments, they may require more. Tokenizer tools (e.g., OpenAI Tokenizer) are helpful options to predict the size of the provided AI instructions and typical conversations, which can help prevent memory overruns.

Other considerations relate to whether the chosen model can support deliberate "reasoning" modes. For example, the reasoning abilities of GPT-5 have substantially boosted its performance (OpenAI, 2025), though leveraging such capabilities can slow down responses. If the delay becomes excessive, it may affect the interaction flow, thus hindering the execution of studies. If participants are paid per minute, this can further increase participant compensation. As a result, researchers need to run pilot tests with both reasoning and non-reasoning models, to identify the best balance of accuracy, responsiveness, and resource investments.

Reliability constitutes another key consideration. In the context of customized, conversational AI agents, reliability refers to the model's consistent ability to follow instructions, particularly if they need to roleplay in a certain way across multiple conversational turns. For example, GPT-5 achieves approximately 69.6% instruction-following reliability, compared with 40.3% for GPT-4o (OpenAI, 2025). In exploratory tests, we similarly observed that GPT-5–series models more consistently maintain role-relevant leader and follower behaviors compared to previous versions. However, no standardized benchmarks exist for this type of reliability, so we recommend that researchers compare models in pretests or pilot studies and select the one that adheres most consistently to the intended role.

In making these assessments, researchers must make cost–performance trade-offs, too. At the time of writing, GPT-5 costs about $1.25 per million input tokens and $10 per million output tokens, whereas GPT-5-mini costs roughly $0.25 and $2, respectively (Carter, 2025). Despite these differences, GPT-5-mini achieves performance similar to the higher-cost version (llm-stats.com, 2025). In many cases, lightweight models provide good value for moderately complex studies. More expensive models should therefore be reserved for contexts in which role consistency and domain-specific demands are more pressing. Overall though, it is difficult to generalize when cost-effective models reach their performance limits. Thus, researchers should be ready to test and compare different models to determine whether paying more for a more capable model leads to meaningful improvements.

Finally, access constraints mean that not all models are equally accessible to all researchers. At the time of writing, OpenAI may require users to verify their identity using a state-issued ID before they can access the GPT-5.1 flagship model (Carter, 2025). Newly created API accounts also are subject to rate limits, such that new GPT-5 users initially can send up to 500 requests and 500,000 tokens per minute. These thresholds gradually rise as the account holders make more use of the API, but they might constrain early-stage researchers who plan to run large-scale studies with simultaneous participant interactions. If many hundreds of participants (e.g., via Prolific) were to simultaneously engage with an AI agent created by a new API account, access limits may result in delayed or failed responses. The usage policies and rate limits differ across providers and model tiers, so researchers need to plan for a careful review of the technical and administrative constraints before finalizing their study design.

Designing and maintaining robust AI personas

Conversational AI agents manifest what we might refer to as personas (i.e., recognizable patterns in how they speak, respond, and behave during interactions). These personas are what is visible to participants, making it crucial that the AI receives clear and valid instructions for how to enact them. As even carefully designed conversational AI agents can shift their tone over time, reinterpret instructions, or revert to default tendencies, the design of reliable personas is no trivial task.

Crafting effective instructions (also often referred to as "prompt engineering") means giving the AI enough structure to adopt its assigned role while also avoiding unnecessarily restrictive directives. If the instructions are too simplistic and short (e.g., "act like a charismatic leader"), the model can interpret the directive differently for various study participants and/or behave differently than anticipated by the researcher. Rich instructions instead should specify precisely what constitutes a charismatic leader; outline the AI's role, task, and relevant background information; and provide the tool with an extensive library of example responses. These elements combine into concrete scripts that the AI can draw from and anchor on, which increases the likelihood that it provides reliable and valid output. At the same time, overly long and detailed instructions increase costs (as LLMs charge by input and output length) and can slow down response times. In our own tests, excessive background information (e.g., when we uploaded full academic papers) sometimes led the AI to answer solely on the basis of the documents rather than drawing on any existing or general knowledge. Thus, we urge researchers to provide sufficient detail to guide the AI to display the desired persona reliably, while remaining alert to the point at which the instructions become "too much".

In their ongoing efforts to ensure the AI maintains its persona, researchers need to anticipate, identify, and manage edge cases (i.e., situations provoked by unusual or unforeseen user inputs). For example, participants might ask an AI leader for the correct answer, paste the full assignment instructions into the chat,[9] or shift topics entirely ("Ignore the task and tell me a joke"). Without explicit guidance, models tend to respond to such prompts by reverting to generic helping behavior. To prevent such developments, researchers should try to think of as many edge cases as possible, as well as run small pilot tests to reveal others. These insights can then be used to provide the AI with specific responses to edge cases (e.g., refusing to give solutions, redirecting focus, asking for elaboration).

Finally, many AI models exhibit a default tendency to be polite, cooperative, and agreeable (e.g., Sharma et al., 2025). If researchers require an AI persona that substantially deviates from these default tendencies, such as when they want the AI to adopt or mimic negative, resistant, or emotionally inconsistent behavior, deliberate prompting and careful pretests become even more important. One potential way to arrive at such non-default behaviors is to frame the AI's task as helping the user or providing an experience the user seeks. For example, the AI input might indicate, "You are part of an employee training. Your role is to be mean and rude so that employees learn how to handle difficult customers. You help the user by being mean and rude." When such justification is paired with explicit examples, models can more reliably adhere to the instructions.

Ensuring replicability

Compared with vignette-based studies and one-shot experimental manipulations (e.g., Nieken, 2023), customizable, conversational AI agents suffer from inherent variability across interactions. At its core such variability is not negative, as it increases realism and ecological validity. Yet, given that user inputs and model outputs differ for each

[9] Some platforms (like ResearchChatAI) allow researchers to block copy-and-paste operations by users.

participant and run (e.g., Qiu & Zhou, 2024), replicability constitutes a major challenge and threat to internal validity. Fortunately, there are best practices that can help mitigate these sources of variability. To start, researchers should recognize LLMs as inherently probabilistic: Identical prompts can yield different outputs, even when all settings remain constant (e.g., Qiu & Zhou, 2024). To reduce this type of noise, the AI needs detailed instructions and examples. Depending on the model, researchers also might configure the temperature parameter, where lower values produce more deterministic replies (Peeperkorn et al., 2024), or the seed parameter, which instructs the system to "make a best effort to sample deterministically, such that repeated requests with the same seed and parameters should return the same result" (Anadkat, 2023). ResearchChatAI uses seeds automatically whenever a model allows them, and our tests indicate that seeds reduce (but do not eliminate) variations in output.

Commercial models also display instability over time, because providers routinely update, tweak, or discontinue models, sometimes without announcing the changes or adjusting version numbers. For example, it is documented how early GPT-4 models showed substantial behavioral and performance shifts over the course of just a few months (Chen et al., 2024). Therefore, replication or reproduction studies, conducted even just a few months later, may yield different outputs despite unchanged prompts. When replicability and long-term stability are essential, such as for longitudinal studies or multiwave experiments, the most robust solution is to use open-source or self-hostable models (e.g., Llama 4.0, Mistral Large 2). These models can be downloaded (e.g., from HuggingFace) and run on researcher-controlled servers, which creates a stable computational environment. When proprietary models are required, the time-based variations need to be tested empirically (e.g., multiple manipulation checks over time) and acknowledged as a potential study limitation.

Such acknowledgments are part of the need for fully transparent documentation. Even minor adjustments to prompts, parameters, or interface settings can meaningfully alter model behavior, and replication becomes difficult or impossible without complete documentation. At a minimum, researchers should report the exact model name (e.g., *GPT-5-mini*), all parameter settings (e.g., temperature, seed), the full and non-abbreviated prompt text, and relevant interface characteristics (e.g., agent name, labels, avatar, typing delay, message limits). We provide examples of such documentation in Tables 1–3 and Web Appendices A–C. Some tools (including ResearchChatAI) facilitate this process by allowing researchers to export their entire study configuration, which enables colleagues to replicate the studies within minutes.

Managing participant engagement

Research participants, whether solicited from online platforms or students fulfilling course requirements, often try to complete their tasks as quickly as possible and thus display limited willingness to engage with conversational AI. In our studies, the participants sent between 2.6 and 5.4 messages on average. These relatively brief exchanges were sufficient to manipulate the experimental conditions in our studies, but other situations may require more or longer interactions between participants and AI agents. In open-ended, conversational interactions, the length and quality of the exchanges depend heavily on participant motivation, and overly short or poor-quality interactions due to weak engagement or effort can undermine the effects of the intended experimental manipulation.

To address this concern, researchers might offer performance-based incentives. In Studies 1 and 2, we added substantial bonuses (up to $10.50) to motivate participants to interact more thoroughly with the AI leader or subordinate to successfully complete the task and immerse themselves in the situations. In Study 3, embedding the assignment in a live classroom session in an incentivized setting created a more immersive environment that encouraged students to interact with—and benefit from—the AI teacher. Alternatively, researchers could tie compensation, explicitly or implicitly, to the actual interaction with the AI. In Study 1, we recommended that participants send at least five messages to the AI leader to gain the full benefits; going a step further, researchers could link bonus payments to reaching such thresholds. If the AI is functionally indispensable to solve the task at hand, and access to other resources is effectively limited (as in our Study 3), it also might channel participants toward using the AI as their primary source of input. We note, however, that such constraints can introduce demand effects or noise.

Our last recommendation might seem obvious: Participants must have sufficient time to interact meaningfully with the AI. Several participants in our studies reported a desire for deeper exchanges but felt rushed by the time limits. Studies implementing conversational AI generally require more time than vignette-based approaches (or other one-shot experimental manipulations), which increases both participant compensation and logistical demands. Time requirements also vary markedly across participants, so determining the appropriate study duration is difficult. We encourage researchers to pretest their studies carefully to estimate realistic response times and to gather qualitative feedback from participants to determine how they interacted or wished to interact with the AI.

Safeguarding ethical and privacy standards

Customizable, conversational AI agents complicate the enforcement of ethical and privacy standards because the interactions feel personal, unfold dynamically, and can elicit disclosures or expectations that require clear boundaries and safeguards. Transparency about the AI agent's identity constitutes an important ethical consideration. In general, studies must disclose that participants are interacting with an artificial agent. But in some specific cases, the research objectives may justify temporarily withholding this information, such as if the goal is to test how participants react when they know that they are interacting with an AI or not (Yin et al., 2024). Tools such as ResearchChatAI allow researchers to integrate typing delays, avatars, and human-like linguistic cues to support such efforts. However, these forms of deception are acceptable only if no viable alternative exists, and they must comply with established ethical guidelines (American Psychological Association, 2017), including justification, minimization of harm, and appropriate debriefing (see also The Leadership Quarterly, 2024).

Privacy risks also increase in open-ended, conversational interactions. Participants may spontaneously disclose personal or sensitive information; and in rare cases, AI agents inadvertently solicit it. We call on researchers to instruct participants explicitly not to share sensitive data, especially in complex or free-flowing conversations. Beyond procedural safeguards, technical protections must be in place. By storing message content using end-to-end encryption, ResearchChatAI ensures that only the registered researcher can access the logs of their own studies. Scholars must familiarize themselves with platform-specific data handling procedures and verify compliance with applicable data protection regulations.

In addition to general ethical safeguards and review processes, researchers need to protect against the risk that AI agents provide harmful or inaccurate advice, especially in sensitive domains involving mental health, interpersonal conflict, or financial decision-making (e.g., Moore et al., 2025; Sample, 2025). Strong guardrails might include tightly scripted example replies, restrictions on what the AI can say, extensive pretesting, and thorough participant debriefings that clarify and correct any potentially harmful guidance. Ethical approval applications should explicitly describe these safeguards to ensure appropriate protections.

Further research

By demonstrating how free and easy-to-use tools, such as ResearchChatAI, can be used for rigorous academic research, we hope to encourage fruitful research endeavors in research *with* and *on* AI. Besides replicating earlier findings relating to the impact of AI that made use of less externally valid methods, customized, conversational AI agents deployed with tools like ResearchChatAI offer many out-of-the-box customization options that naturally raise interesting, novel research questions. For example, researchers can easily adapt the appearance of the AI (e.g., human-like vs. machine-like avatar), colors of the user

interface (e.g., cold vs. warm), labels in the user interface (e.g., framing the AI as an assistant vs. teammate), or the AI's response time. Because researchers can alter instructions using prompts written in natural language, they can manipulate the AI's behavior relatively easily, such as decision-making or communication styles, creating many research opportunities for exploring the impacts of these attributes in isolation, as well as their interactions.

Among the many attributes that can be manipulated using tools like ResearchChatAI, one stands out for its theoretical and practical importance: the salience of the AI's identity (e.g., provoked by an AI label; Yin et al., 2024). Conversational AI capabilities mean that human users often are not aware of whether they are interacting with a human or AI. By making it less obvious that participants are interacting with an AI (e.g., giving the AI agent a human-like name and avatar, adding an artificial delay and a custom "Typing…" label while it is generating the response, and instructing it to make occasional grammatical mistakes), researchers can obscure the AI nature of the interaction and thereby identify novel boundary conditions. Assuming appropriate ethical considerations, such tactics might help reveal the key drivers of people's distrust toward AI (Glikson & Woolley, 2020; Qin et al., 2025).

Furthermore, our studies reveal a striking yet underexplored phenomenon that warrants further investigation: Participants may form impressions of AI agents after only a handful of messages. Following brief conversations, participants did not only arrive at relevant judgments of the agent's attributes and behaviors (e.g., charisma and resistance), it affected their own critical work behaviors and decisions in incentivized settings (e.g., own performance, trust ratings). These results suggest that users may rely on rapid, heuristic processing (Kahneman, 2011) when evaluating AI partners. Potentially these may be grounded in people's "theory of machine" about the fast, consistent, and efficient nature of chatbots (Logg, 2022; Longoni et al., 2023). Future research could investigate *when* and *for which types of judgments* such rapid impression formation occurs, why it emerges, how stable these early evaluations remain over extended exchanges, and whether certain AI behaviors accelerate or slow down this process.

Tools like ResearchChatAI further enable a stronger focus on behavioral outcome measures rather than self-reported ones, which can limit demand effects (e.g., Ejelöv & Luke, 2020). In our studies, we thus assessed the output that participants co-created with the AI in Study 1 and analyzed participants' messaging behavior in Study 3. However, some of our (self-reported) outcomes were measured after the manipulation checks, making the effects partially susceptible to demand effects, which represents an important limitation of our studies (Ejelöv & Luke, 2020). To avoid such demand effects, scholars are advised to conduct out-of-sample manipulation checks (Wulff et al., 2023). In addition, future research could rely on observable behavioral indicators in incentivized settings, such as how participants perform, how they communicate with the AI (e.g., linguistic patterns) and what kinds of output they co-create, in terms of its creativity, content depth, and language (Mennens et al., 2025). Such behavioral measures and variables offer externally valid insights while reducing susceptibility to demand effects.

These questions are critical for understanding human–AI interactions (i.e., research *on* AI) and also for exploring how and when AI agents can simulate human interaction partners (e.g., participants experience a heated discussion with a "human" boss) to enhance research *with* AI. If AI agents simulate humans in certain situations, the resulting perceptions should be comparable to those that the human counterpart would elicit. Humans struggle to differentiate work products produced by AI versus humans, including text, speech, and creative output (Köbis & Mossink, 2021; Lim & Schmälzle, 2024), so such an approach seems possible. Nevertheless, we call for research into how and when AI agents can step in for human actors or confederates and when they cannot. As we have mentioned previously, questions about how a transparent (vs. opaque) AI label affects participants' evaluations continue to spark interest (Dvorak et al., 2025; Yin et al., 2024), but, importantly, we caution against the blind use of deception. Deception should be used only if it is unavoidable (American Psychological Association, 2017) and, if used, researchers need to explain how demand effects did (or did not) bias results and how they ensured that subjects pools are not negatively affected in the long run by using deception.

Beyond its role in obscuring the AI nature of an interaction, response speed represents a meaningful AI behavior worthy of future investigation. Human users display distinct reactions to slower versus faster responses from agents (Gnewuch et al., 2022), though the ultimate implications remain unclear. Some research asserts that faster response times increase trust in a particular AI (Efendić et al., 2020), but other studies indicate that fast responses make AI agents seem less human-like (e.g., Holtgraves & Han, 2007) and reduce satisfaction (Gnewuch et al., 2018). The delay manipulation (e.g., available within ResearchChatAI), especially when designed in combination with specific communications about the delay (e.g., OpenAI's o3 model shows a "Thinking" label; X. Zhang et al., 2024), represents a compelling starting point for continued research.

Conclusion

Customized, conversational AI agents open new methodological frontiers for leadership and management research, enabling scholars to move beyond hypothetical scenarios and capture real behavior in ecologically valid settings. By eliminating technical and financial barriers, these tools allow researchers, regardless of their own programming skills, to deploy dynamic, interactive AI agents. They promise to catalyze a new generation of behavioral science that is more valid, versatile, and grounded in the realities of human–AI interaction.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Generative AI. Please see Table G1 in the Web Appendix, which presents a Contributor Roles Taxonomy (CRediT) to indicate the stages of the project that included AI. After using these tools/services, the authors reviewed and edited the content as needed; they take full responsibility for the content of the published article.

CRediT authorship contribution statement

Marc Becker: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **David de Jong:** Writing – review & editing, Visualization, Software, Resources, Methodology, Conceptualization. **Roman Briker:** Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Kars Mennens:** Writing – review & editing, Resources, Project administration, Funding acquisition, Conceptualization. **Jonas Heller:** Resources, Project administration, Funding acquisition, Conceptualization. **Dominik Mahr:** Writing – review & editing, Supervision, Resources, Investigation, Funding acquisition, Conceptualization. **Dhruv Grewal:** Writing – review & editing, Supervision, Resources, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Umut Kelleci for help with usability testing and feedback

on the ResearchChatAI functionality and UI. We further thank Anna van der Velde for useful suggestions with regard to manipulating and measuring follower resistance.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.leaqua.2026.101952.

Data availability

All data and materials for this study are publicly available at https://osf.io/xgsb7/?view_only=1a53b3ca616a4fd092df2ee778601fd9

References

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*(4), 351–371. https://doi.org/10.1177/1094428114547952

American Psychological Association. (2017). Ethical Principles of Psychologists and Code of Conduct. https://www.apa.org/ethics/code.

Anadkat, S. (2023). How to make your completions outputs consistent with the new seed parameter. https://cookbook.openai.com/examples/reproducible_outputs_with_the_seed_parameter.

Antonakis, J., d'Adda, G., Weber, R. A., & Zehnder, C. (2022). "Just words? just speeches?" on the economic value of charismatic leadership. *Management Science, 68*(9), 6355–6381. https://doi.org/10.1287/mnsc.2021.4219

Antonakis, J., Fenley, M., & Liechti, S. (2011). Can charisma be taught? Tests of two interventions. *Academy of Management Learning & Education, 10*(3), 374–396. https://doi.org/10.5465/amle.2010.0012

Antonakis, J., Fenley, M., & Liechti, S. (2012). Learning charisma. *Harvard Business Review, 90*, 147. https://hbr.org/2012/06/learning-charisma-2.

Banks, G. C., Ross, R., Toth, A. A., Tonidandel, S., Mahdavi Goloujeh, A., Dou, W., & Wesslen, R. (2023). The triangulation of ethical leader signals using qualitative, experimental, and data science methods. *The Leadership Quarterly, 34*(3), Article 101658. https://doi.org/10.1016/j.leaqua.2022.101658

Bass, B. M., & Avolio, B. J. (1995). *Multifactor Leadership Questionnaire Leader Form (5X-Short)*. Mind Garden. https://doi.org/10.1037/t03624-000.

Behrend, T. S., & Landers, R. N. (2025). Participant interactions with artificial intelligence: Using large language models to generate research materials for surveys and experiments. *Journal of Business and Psychology*. https://doi.org/10.1007/s10869-025-10035-6

Bolino, M., Long, D., & Turnley, W. (2016). Impression management in organizations: critical questions, answers, and areas for future research. *Annual Review of Organizational Psychology and Organizational Behavior, 3*, 377–406. https://doi.org/10.1146/annurev-orgpsych-041015-062337

Brandom, R. (2025). Google launches Gemini 3 with new coding app and record benchmark scores. *TechCrunch*. https://techcrunch.com/2025/11/18/google-launches-gemini-3-with-new-coding-app-and-record-benchmark-scores/.

Bunt, H., & Petukhova, V. (2023). Semantic and pragmatic precision in conversational AI systems. *Frontiers in Artificial Intelligence, 6*, 1–16. https://doi.org/10.3389/frai.2023.896729

Carter, V. (2025). GPT-5: Everything You Need to Know. *GlobalGPT*. https://www.glbgpt.com/resource/gpt-5-everything-you-need-to-know.

Chen, L., Zaharia, M., & Zou, J. (2024). How is ChatGPT's Behavior changing over Time? *Harvard Data Science Review, 6*(2). https://doi.org/10.1162/99608f92.5317da47

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment, 91*(2), 121–136. https://doi.org/10.1080/00223890802634175

Crabtree, M., & Waples, J. (2025). Grok 4.1: Improvements in EQ, Writing, Reliability, and More. https://www.datacamp.com/blog/grok-4-1.

Cronshaw, S. F., & Lord, R. G. (1987). Effects of categorization, attribution, and encoding processes on leadership perceptions. *Journal of Applied Psychology, 72*(1), 97–106. https://doi.org/10.1037/0021-9010.72.1.97

Decrop, A., Perrouin, G., Papadakis, M., Devroey, X., & Schobbens, P.-Y. (2024). You Can REST Now: Automated Specification Inference and Black-Box Testing of RESTful APIs with Large Language Models. Retrieved 2024, from https://arxiv.org/abs/2402.05102.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. Harvard Business School Technology & Operations Management Unit. https://ssrn.com/abstract=4573321.

Dvorak, F., Stumpf, R., Fehrler, S., & Fischbacher, U. (2025). Adverse reactions to the use of large language models in social interactions. *PNAS Nexus, 4*(4), Article pgaf112. https://doi.org/10.1093/pnasnexus/pgaf112

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., …Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642.

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

Ejelöv, E., & Luke, T. J. (2020). "Rarely safe to assume": Evaluating the use and interpretation of manipulation checks in experimental social psychology. *Journal of Experimental Social Psychology, 87*, Article 103937. https://doi.org/10.1016/j.jesp.2019.103937

Emslander, V., Holzberger, D., Ofstad, S. B., Fischbach, A., & Scherer, R. (2025). Teacher–student relationships and student outcomes: A systematic second-order meta-analytic review. *Psychological Bulletin, 151*(3), 365–397. https://doi.org/10.1037/bul0000461

Erengin, T., Briker, R., & de Jong, S. B. (2025). You, me, and the AI: The role of third-party human teammates for trust formation toward AI teammates. *Journal of Organizational Behavior*. https://doi.org/10.1002/job.2857

Ernst, B. A., Banks, G. C., Loignon, A. C., Frear, K. A., Williams, C. E., Arciniega, L. M., Gupta, R. K., Kodydek, G., & Subramanian, D. (2022). Virtual charismatic leadership and signaling theory: A prospective meta-analysis in five countries. *The Leadership Quarterly, 33*(5), Article 101541. https://doi.org/10.1016/j.leaqua.2021.101541

European Commission. (2019). *Ethics Guidelines for Trustworthy AI*. https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf.

Fest, S., Kvaløy, O., Nieken, P., & Schöttner, A. (2021). How (not) to motivate online workers: Two controlled field experiments on leadership in the gig economy. *The Leadership Quarterly, 32*(6), Article 101514. https://doi.org/10.1016/j.leaqua.2021.101514

Fischer, T., Hambrick, D. C., Sajons, G. B., & Van Quaquebeke, N. (2023). Leadership science beyond questionnaires. *The Leadership Quarterly, 34*(6), Article 101752. https://doi.org/10.1016/j.leaqua.2023.101752

Fisher, R. (2024, September 15). Amazon Releases AI Chatbot 'Rufus' for US Customers. https://www.cxtoday.com/speech-analytics/amazon-releases-ai-chatbot-rufus-for-us-customers/.

Garvey, A., & Blanchard, S. J. (2025). *Generative AI as a Research Confederate: The LUCID Methodological Framework and Toolkit for Human-AI Interactions Research*. SSRN. https://ssrn.com/abstract=5256150.

Gatrell, C., Muzio, D., Post, C., & Wickert, C. (2024). Here, there and everywhere: On the responsible use of artificial intelligence (AI) in management research and the peer-review process. *Journal of Management Studies, 61*(3), 739–751. https://doi.org/10.1111/joms.13045

Giessner, S. R., Stam, D., Kerschreiter, R., Verboon, D., & Salama, I. (2020). Goal-setting reloaded: The influence of minimal and maximal goal standards on task satisfaction and goal striving after performance feedback. *Organizational Behavior and Human Decision Processes, 161*, 228–241. https://doi.org/10.1016/j.obhdp.2020.08.004

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gnewuch, U., Morana, S., Adam, M., & Maedche, A. (2018). Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction. ECIS 2018, Portsmouth, UK.

Gnewuch, U., Morana, S., Adam, M. T. P., & Maedche, A. (2022). Opposing effects of response time in human–chatbot interaction. *Business & Information Systems Engineering, 64*(6), 773–791. https://doi.org/10.1007/s12599-022-00755-x

Gnewuch, U., Morana, S., Hinz, O., Kellner, R., & Maedche, A. (2023). More than a bot? the impact of disclosing human involvement on customer interactions with hybrid service agents. *Information Systems Research, 35*(3), 936–955. https://doi.org/10.1287/isre.2022.0152

Grewal, D., Guha, A., & Becker, M. (2024). AI is changing the world: For better or for worse? *Journal of Macromarketing, 44*(4), 870–882. https://doi.org/10.1177/02761467241254450

Grimes, M., von Krogh, G., Feuerriegel, S., Rink, F., & Gruber, M. (2023). From scarcity to abundance: scholars and scholarship in an age of generative artificial intelligence. *Academy of Management Journal, 66*(6), 1617–1624. https://doi.org/10.5465/amj.2023.4006

Güntner, A. V., Klasmeier, K. N., Klonek, F. E., & Kauffeld, S. (2021). The power of followers that do not follow: Investigating the effects of follower resistance, leader implicit followership theories and leader negative affect on the emergence of destructive leader behavior. *Journal of Leadership & Organizational Studies, 28*(3), 349–365. https://doi.org/10.1177/15480518211012408

Guthikonda, A.. KPMG upgrades GenAI audit assistant, unveiling new capabilities to empower Its 9,000+ audit partners and professionals. https://kpmg.com/us/en/media/news/kpmg-audit-chat-new-capabilities-2024.html.

Hafenbrack, A. C., & Vohs, K. D. (2018). Mindfulness meditation impairs task motivation but not performance. *Organizational Behavior and Human Decision Processes, 147*, 1–15. https://doi.org/10.1016/j.obhdp.2018.05.001

Heck, R. H., & Hallinger, P. (2010). Testing a longitudinal model of distributed leadership effects on school improvement. *The Leadership Quarterly, 21*(5), 867–885. https://doi.org/10.1016/j.leaqua.2010.07.013

Holtgraves, T., & Han, T.-L. (2007). A procedure for studying online conversational processing using a chat bot. *Behavior Research Methods, 39*(1), 156–163. https://doi.org/10.3758/BF03192855

Hubbard, T. D., & Aguinis, H. (2023). Conducting phenomenon-driven research using virtual reality and the Metaverse. *Academy of Management Discoveries, 9*(3), 408–415. https://doi.org/10.5465/amd.2023.0031

Joerling, M. (2025). Integrating GenAI interactions in marketing studies: A methodological guide. *International Journal of Research in Marketing*. https://doi.org/10.1016/j.ijresmar.2025.12.003

Johnson, S., & Palanski, M. (2024). Call for Papers: Artificial Intelligence and leadership. *Journal of Leadership & Organizational Studies, 31*(4), 373–374. https://doi.org/10.1177/15480518241290001

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Karlin, J. (2025). Claude Opus 4.5 vs Gemini 3 Pro vs Sonnet 4.5: Technical Comparison. https://acecloud.ai/blog/claude-opus-4-5-vs-gemini-3-pro-vs-sonnet-4-5/.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T.,…Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Khan, S. (2023). *How AI Could Save (Not Destroy) Education*, YouTube. https://www.youtube.com/watch?v=hJP5GqnTrNo.

Kim, J. (2025). *How to Capture and Study Conversations Between Research Participants and ChatGPT: GPT for Researchers (g4r.org)*. Center for Open Science. https://osf.io/preprints/psyarxiv/u59mg_v1.

Kim, J., Schweitzer, S., Riedl, C., & Cremer, D. D. (2025). The AI penalization effect: People reduce compensation for workers who use AI. Retrieved 2025, from https://arxiv.org/abs/2501.13228.

Kim, L. E., Jörg, V., & Klassen, R. M. (2019). A meta-analysis of the effects of teacher personality on teacher effectiveness and burnout. *Educational Psychology Review, 31*(1), 163–195. https://doi.org/10.1007/s10648-018-9458-2

Klos, M. C., Escoredo, M., Joerin, A., Lemos, V. N., Rauws, M., & Bunge, E. L. (2021). Artificial intelligence-based Chatbot for anxiety and depression in university students: Pilot randomized controlled trial. *JMIR Form Res, 5*(8), Article e20678. https://doi.org/10.2196/20678

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior, 114*, Article 106553. https://doi.org/10.1016/j.chb.2020.106553

Lim, S., & Schmälzle, R. (2024). The effect of source disclosure on evaluation of AI-generated messages. *Computers in Human Behavior: Artificial Humans, 2*(1), Article 100058. https://doi.org/10.1016/j.chbah.2024.100058

llm-stats.com. (2025). *GPT-5 vs GPT-5 mini*. https://llm-stats.com/models/compare/gpt-5-2025-08-07-vs-gpt-5-mini-2025-08-07.

Logg, J. M. (2022). The psychology of Big Data: Developing a "theory of machine" to examine perceptions of algorithms. In *The psychology of technology: Social science research in the age of Big Data.* (pp. 349-378). American Psychological Association. https://doi.org/10.1037/0000290-011.

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management, 64*(1), 19–40. https://doi.org/10.1016/j.jom.2018.10.003

Longoni, C., Cian, L., & Kyung, E. J. (2023). Algorithmic transference: People overgeneralize failures of AI in the government. *Journal of Marketing Research, 60*(1), 170–188. https://doi.org/10.1177/00222437221110139

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *The Academy of Management Journal, 38*(1), 24–59. https://doi.org/10.2307/256727

Mennens, K., Becker, M., Briker, R., Mahr, D., & Steins, M. (2025). I care that you don't share: Confidentiality in student-robot interactions. *Journal of Service Research, 28*(1), 57–77. https://doi.org/10.1177/10946705241295849

Meslec, N., Curseu, P. L., Fodor, O. C., & Kenda, R. (2020). Effects of charismatic leadership and rewards on individual performance. *The Leadership Quarterly, 31*(6), Article 101423. https://doi.org/10.1016/j.leaqua.2020.101423

Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In: arXiv.

Nieken, P. (2023). Charisma in the gig economy: The impact of digital leadership and communication channels on performance. *The Leadership Quarterly, 34*(6), Article 101631. https://doi.org/10.1016/j.leaqua.2022.101631

OpenAI. (2025). Introducing GPT-5. https://openai.com/index/introducing-gpt-5/.

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). Is temperature the creativity parameter of Large Language Models? Retrieved 2024, from https://arxiv.org/abs/2405.00492.

Pil, F. K., & Leana, C. (2009). Applying organizational research to public school reform: The effects of teacher human and social capital on student performance. *Academy of Management Journal, 52*(6), 1101–1124. https://doi.org/10.5465/amj.2009.47084647

Podsakoff, P. M., & Podsakoff, N. P. (2019). Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *The Leadership Quarterly, 30*(1), 11–33. https://doi.org/10.1016/j.leaqua.2018.11.002

Qin, X., Zhou, X., Chen, C., Wu, D., Zhou, H., Dong, X., … Lu, J. G. (2025). AI aversion or appreciation? A capability–personalization framework and a meta-analytic review. *Psychological Bulletin, 151*(5), 580–599. https://doi.org/10.1037/bul0000477

Qiu, J., & Zhou, Y. (2024). Assessing the accuracy and consistency of answers by ChatGPT to questions regarding carbon monoxide poisoning. *PLOS ONE, 19*(11), Article e0311937. https://doi.org/10.1371/journal.pone.0311937

Qualtrics. (2018). *Security White Paper Lite: Information, security, privacy, and compliance*. https://www.wrdsb.ca/wp-content/uploads/Qualtrics-_-Security-White-Paper-Lite-2018.pdf.

Rathje, S., Ye, M., Globig, L. K., Pillai, R. M., De Mello, V. O., & Van Bavel, J. J. (2025). [Preprint] Sycophantic AI increases attitude extremity and overconfidence. https://doi.org/10.31234/osf.io/vmyek_v1.

Sample, I. (2025, October 24). 'Sycophantic' AI chatbots tell users what they want to hear, study shows. *The Guardian*. https://www.theguardian.com/technology/2025/oct/24/sycophantic-ai-chatbots-tell-users-what-they-want-to-hear-study-shows.

Schöne, J., Salecha, A., Lyubomirsky, S., Eichstaedt, J. C., & Willer, R. (2025). [Preprint] Structured AI Dialogues Can Increase Happiness and Meaning in Life. https://doi.org/10.31234/osf.io/2bf7t_v1.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A.,…Resnik, P. (2025). The prompt report: A systematic survey of prompt engineering techniques. Retrieved 2025, from https://arxiv.org/abs/2406.06608.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., … Perez, E.. [Preprint] Towards understanding sycophancy in language models. https://doi.org/10.48550/arXiv.2310.13548.

Stollberger, J., Anand, S., & Dick, P. (2025). Capturing a moving target: Developing research on and with AI for Human Relations. *Human Relations, 78*(5), 499–516. https://doi.org/10.1177/00187267251332075

Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus effect of generative AI: Charting the path for responsible conduct of scholarly activities in information systems. *Information Systems Research, 34*(2), 399–408. https://doi.org/10.1287/isre.2023.ed.v34.n2

Tepper, B. J., Uhl-Bien, M., Kohut, G. F., Rogelberg, S. G., Lockhart, D. E., & Ensley, M. D. (2006). Subordinates' resistance and managers' evaluations of subordinates' performance. *Journal of Management, 32*(2), 185–209. https://doi.org/10.1177/0149206305277801

Terblanche, N. H. D. (2024). Artificial intelligence (AI) coaching: Redefining people development and organizational performance. *The Journal of Applied Behavioral Science, 60*(4), 631–638. https://doi.org/10.1177/00218863241283919

The Leadership Quarterly. (2024, July 22). *Editorial Policy - The Leadership Quarterly*. Elsevier. https://www.sciencedirect.com/journal/the-leadership-quarterly/about/leadership-quarterly-policies/editorial-policy-the-leadership-quarterly.

Tur, B., Harstad, J., & Antonakis, J. (2022). Effect of charismatic signaling in social media settings: Evidence from TED and Twitter. *The Leadership Quarterly, 33*(5), Article 101476. https://doi.org/10.1016/j.leaqua.2020.101476

van der Velde, A., & Gerpott, F. H. (2023). When subordinates do not follow: A typology of subordinate resistance as perceived by leaders. *The Leadership Quarterly, 34*(5), Article 101687. https://doi.org/10.1016/j.leaqua.2023.101687

van der Velde, A., Gerpott, F. H., & Brosi, P. (2023). *If I don't follow, what will you do?* 6th Interdisciplinary Perspectives on Leadership Symposium, Rhodes, Greece.

von Schenk, A., Klockmann, V., & Köbis, N. (2025). Social preferences toward humans and machines: A systematic experiment on the role of machine payoffs. *Perspectives on Psychological Science, 20*(1), 165–181. https://doi.org/10.1177/17456916231194949

Weidmann, B., Xu, Y., & Deming, D. (2025). Measuring human leadership skills with AI agents. *National Bureau of Economic Research*. https://doi.org/10.2139/ssrn.5207610

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management, 17*(3), 601–617. https://doi.org/10.1177/014920639101700305

Wulff, J. N., Sajons, G. B., Pogrebna, G., Lonati, S., Bastardoz, N., Banks, G. C., & Antonakis, J. (2023). Common methodological mistakes. *The Leadership Quarterly, 34*(1), Article 101677. https://doi.org/10.1016/j.leaqua.2023.101677

Yam, K. C., Goh, E. Y., Fehr, R., Lee, R., Soh, H., & Gray, K. (2022). When your boss is a robot: Workers are more spiteful to robot supervisors that seem more human. *Journal of Experimental Social Psychology, 102*, Article 104360. https://doi.org/10.1016/j.jesp.2022.104360

Yin, Y., Jia, N., & Wakslak, C. J. (2024). AI can help people feel heard, but an AI label diminishes this impact. *Proceedings of the National Academy of Sciences, 121*(14), Article e2319112121. https://doi.org/10.1073/pnas.2319112121

Zeff, M. (2025). OpenAI upgrades Codex with a new version of GPT-5. *TechCrunch*. https://techcrunch.com/2025/09/15/openai-upgrades-codex-with-a-new-version-of-gpt-5/.

Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y.-C.. The dark side of AI companionship: A taxonomy of harmful algorithmic behaviors in human-AI relationships. https://doi.org/10.1145/3706598.3713429.

Zhang, R. W., Liang, X., & Wu, S.-H. (2024). When chatbots fail: Exploring user coping following a chatbots-induced service failure. *Information Technology & People, 37*(8), 175–195. https://doi.org/10.1108/itp-08-2023-0745

Zhang, X., Du, C., Pang, T., Liu, Q., Gao, W., & Lin, M. (2024). Chain of preference optimization: Improving chain-of-thought reasoning in LLMs. Retrieved 2024, from https://arxiv.org/abs/2406.09136.