Marketing Science Institute Working Paper Series 2025

Report No. 25-118

Generative AI as a Research Confederate: The LUCID Methodological Framework and Toolkit for Human-AI Interactions Research

Aaron M. Garvey and Simon J. Blanchard

Generative AI as a Research Confederate:

The LUCID Methodological Framework and Toolkit for Human-AI Interactions Research

Aaron M. Garvey*

Simon J. Blanchard

Draft date: May 15th 2025. See lucidresearch.io for latest version

*Aaron M. Garvey is the Bloomfield Associate Professor of Marketing at the Gatton College of Business and Economics, University of Kentucky, Lexington, KY, 40506, Tel: (859) 257-2869, AaronGarvey@uky.edu. Simon J. Blanchard is the Provost's Distinguished Associate Professor & Dean's Professor at the McDonough School of Business, Georgetown University, Washington, DC, 22057, Tel: 202-687-6977, sjb247@georgetown.edu.

Abstract

As consumer interactions with Generative AI (GenAI) agents—such as ChatGPT—become increasingly common, marketing researchers face new methodological challenges in studying these dynamic engagements. This article introduces the LUCID Framework (LLM-Unified Confederate for Interactive Dialogue), a paradigm shift that conceptualizes GenAI agents as research confederates for studies in which they interact with participants. The framework rests on five core pillars: ecological validity, experimental control, reliable measurement, replicability, and accessibility. To support adoption, we introduce the LUCID Toolkit—a no-code, open-source implementation package that integrates controllable GenAI agents into Qualtrics surveys. The Toolkit enables researchers to deploy and manipulate AI confederates with minimal technical burden, while preserving rigorous data collection, prompt control, and transparent reporting. Together, the LUCID Framework and Toolkit provide a structured, replicable, and accessible foundation for advancing human–AI interaction research in consumer behavior. All materials, templates, and code are available at www.lucidresearch.io.

Keywords: survey design, experimental design, artificial intelligence, generative AI, GenAI, LLM, Large Language Model, Qualtrics, ChatGPT, tutorial, research methodology, research confederate, human-AI interaction

Human interactions with Artificial Intelligence (AI) have become an increasingly relevant topic of study within marketing and consumer behavior research in the past decade. The rapid expansion of this literature has been fueled by the widespread and growing use of customer-facing AI agents by firms, primarily in the form of chatbots (Bergner, Hildebrand, Haubl 2023; Luo et al. 2019), voice assistants (Kim and Duhachek 2020), and service robots (Mende et al. 2019). In the past twenty-four months alone, technological advancements in Large Language Models (LLMs) and other generative AI have produced revolutionary advancements in the interactive capabilities of AI systems. Models such as OpenAI's ChatGPT have moved from script execution to the generation of contextually relevant conversational text. This capability of LLMs to dynamically adapt to the nuances of human text inputs is unprecedented, and as of ChatGPT 3.5, is considered by many to have passed the "Turing Test" (Gams and Kramar 2024).

Despite rapid advancements in generative AI technologies like Large Language Model (LLM) chatbots, most empirical studies of chatbot interactions rely on rules-based agents or scripted message flows embedded within static survey platforms – not LLMs (Hermann and Puntoni 2024). While such designs offer strong internal validity, they sacrifice ecological realism by constraining participant responses and simulating chatbot behavior through deterministic, non-interactive means. Other studies have employed proprietary GenAI platforms, but these approaches often obscure implementation details and inhibit replicability, undermining open science practices. As a result, researchers face a trade-off between ecological validity and experimental control, with no existing methodology offering a satisfying balance.

In this article, we first introduce the LUCID Framework (LLM-Unified Confederate for Interactive Dialogue). LUCID offers a paradigm shift in how GenAI agents are conceptualized

within research: not as black-box conversational stimuli, but as programmable research conferences. Drawing on the tradition of using human confederates in social science, we propose that LLMs can be directed to play specific repeatable roles within consumer interactions – provided that researchers carefully craft and validate the underlying prompt. This reframing allows researchers to stimulate naturalistic interactions while maintaining control over key contextual variables and manipulations just as they would with a human confederate.

The LUCID Framework is built on five methodological pillars: (1) ecological validity, (2) experimental control, (3) reliable measurement, (4) replicability, (5) accessibility. Together, these pillars provide the guiding structure for designing, executing and reporting consumer research involve conversations with GenAI. As we discuss, most existing approaches do not enable these pillars in ways that are consistent with the rigor needed for theory development, hypothesis testing, and reproducible science.

To support adoption of the LUCID Framework, we also introduce the LUCID Toolkit, a no-code, open-source package developed specifically for integrating ChatGPT-powered, conversational confederate agents into Qualtrics surveys. The LUCID Toolkit allows researchers to implement LUCID-based studies with minimal technical knowledge, while still allowing experimentation on prompts, randomization across participations, logging of full conversations, and facilitation of reporting of all relevant technological details. By adding the implementation toolkit to the conceptual framework, we are able to provide both a high-level set of principles for evaluating how GenAI interactions are studied in primary research and a concrete infrastructure to advance the study and human-AI interactions.

The paper proceeds as follows. First, we review existing marketing research that has incorporated AI agents in the form of chatbots and document the limitations that motivated the

development of the LUCID framework. Second, we introduce the concept of GenAI confederacy within the context of the LUCID Framework, and its five pillars. Next, we highlight present three core use cases for LLM-powered chatbots within survey-based experiments: (i) as a dynamic stimulus that simulates a realistic marketing interaction, (ii) as a manipulable conversational confederate used to test theoretical predictions through prompt variation, and (iii) as a source of behavioral and linguistic data from which researchers can derive measures. Third, we introduce the LUCID Toolkit, and explains how it integrates ChatGPT into a chat window embedded in a Qualtrics survey. Finally, we present three illustrative studies to guide researchers on how to conduct, report and validate LLM-powered agent research.

We provide all data files, stimuli files, and analysis codes in the OSF (https://osf.io/zw7fa/?view_only=1c00a850dca04a7ca7f0510cccdcb049). For the latest version of the LUCID Toolkit, visit (https://lucidresearch.io).

1. Survey of AI Chatbots in Marketing Research: GenAI is an Opportunity

1.1 Three Eras of Chatbots in Marketing

A chatbot is a typical form of an AI system and one of the most elementary and widespread examples of intelligent human-computer agent interactions (Bansal and Khan 2018). We adopt the conventional definition of a chatbot as "intelligent conversational computer programs that mimic human conversation in its natural form" (Caldarini, Jaff, and McGarry 2022), and note that the term chatbot encompasses the spectrum of conversational AI agents, including voice assistants and dialogue systems (Luo et al. 2022).

The evolution of chatbots in can be understood in three key eras: rules-based, predictive AI, and generative AI (Caldarini, Jaf, and McGarry 2022; Casheekar et al. 2024). Early chatbots

were rules-based, relying on rigid scripts and keyword matching. While offering high control through pre-determined inputs and outputs, these systems lacked flexibility and could not handle open-ended inputs, limiting realism and user engagement (Russell and Norvig 2020). Examples include ELIZA and IKEA's Ask Anna, which followed deterministic decision trees to deliver predefined answers. The predictive AI era emerged in the late 2000s, as machine learning and natural language processing improved. These chatbots could classify user intent based on prior data and respond with contextually appropriate, though still largely pre-determined replies. Tools like Bank of America's Erica or voice assistants such as Siri exemplify this generation. While more adaptive than rules-based systems, predictive bots still fell short of truly dynamic, conversational engagement. The latest era of generative AI chatbots marks a major departure. Powered by LLMs such as OpenAI's GPT, Google Gemini, or Anthropic Claude, these chatbots generate original responses word by word based on real-time context. This enables more fluid, human-like interactions across a wide range of topics and tones. Importantly, these models adapt in real time and do not rely on fixed scripts or past classification alone. They have quickly come to dominate consumer engagement: in the United States over 39% of consumers interact with generative AI at least once per week (OpenText 2024).

However, despite this behavioral shift, marketing research has yet to catch up to the GenAI era. Our review of the literature finds that consumer-facing experiments with GenAI chatbots remain rare. Most past research relies on scripted, rules-based, or predictive agents—approaches that no longer reflect how consumers predominately interact with modern AI systems. This gap underscores the need for new methods—like the LUCID Framework and Toolkit—that can support experimental rigor while maintaining the realism and dynamism of generative AI interactions.

1.2 Review of the Chatbot Literature in Marketing

We begin with a review of extant studies that have examined consumer interactions with chatbots. Our inclusion criteria for the articles followed a search strategy using keywords such as "chatbot" and "artificial intelligence" to search online databases of the Journal of Marketing, Journal of Marketing Research, Journal of Consumer Research, Marketing Science, Journal of the Academy of Marketing Science, Journal of Consumer Psychology, Journal of the Association of Consumer Research, and Journal of Service Research. We also included those published in leading psychology and computer science journals, including Psychological Science, Nature Human Behavior, and Computers in Human Behavior. We restricted ourselves to articles published between 2000–2025 (until April 4th, 2025) that empirically studied AI conversations in consumption contexts. For each article, we coded how the researchers used an AI chatbot as a function of the research design, focusing on whether the chatbot implemented was a rules-based, predictive, or LLM-powered agent, whether the user could respond in free text or had a limited set of options (e.g., menu click responses), how the manipulations occurred (e.g., via a manipulation of bot behavior vs. outside of the bot) and how the measures (e.g., mediators, outcomes) were obtained. The articles are in Table 1.

Contexts and variables studied

Research on chatbots has only broadly emerged in the marketing literature in the past half-decade, with contexts focused predominately on product preference, branding, and service recovery. The majority of research on human-AI interactions in consumer contexts has focused on three areas: *i)* anthropomorphism of chatbots through visuals and descriptions, *ii)* contrasts between chatbot and human representatives, and *iii)* conversational style of the chatbot.

Regarding *(i)*, studies on anthropomorphism and human-like features in chatbots, such as those by Crolic et al. (2022) and Maar et al. (2023), have shown that higher levels of anthropomorphism significantly alter user perceptions, influencing satisfaction and emotional responses, especially during negative service encounters. Regarding *(ii)*, in research examining contrasts between chatbot and human representatives of the firm, Chattaraman et al. (2012, 2013) demonstrated that specific chatbot characteristics positively impact user trust, purchase intentions, and overall satisfaction in online purchase scenarios. Luo et al. (2019) explored how chatbot disclosure influenced affective response and investment behaviors. Regarding *(iii)*, investigations into conversation style, such as the study by Bergner et al. (2023), have highlighted how chatbot behavior can affect consumer brand intimacy and loyalty. While these handful of advances are important, clearly, many opportunities exist for further use of chatbots to advance theory.

Chatbot implementations (Rules-Based, Predictive, or LLM).

The study design techniques used to date have tended not to involve interactions with an actual chatbot, and those that did nearly all employed conversations with either limited, pre-defined user inputs or chatbot outputs (see Luo et al. 2019 and De Freitas 2023 for exceptions). Indeed, none of the experimental papers we surveyed appear to have used Artificial Intelligence Markup Language (AIML) and only one appears to have used Natural Language Processing (NLP; Luo et al. 2019) in the chatbot stimuli, despite the ubiquitous use of these technologies in legacy (i.e., pre-LLM) chatbot development (Park et al. 2022).

Table 1 – Review of Chat-based Interaction Survey

Paper	Research Design			Chatbot Implementation		Manipulations and Measures*
	IV	Mod.	Context/DVs	Model	User Response	
Holzwarth et al. (2006)	Avatar type (Attractive/ Expert / None)	N/A	Apparel purchase / purchase intention, satisfaction, credibility	Rules-based	Multi-option	IV: (E) via scenario DV: (E) via survey
Chattaramanet al. (2012)	Interaction vs. no interaction with chatbot	N/A	Apparel purchase / trust, social support, risk, patronage intentions	Rules-based (SitePal)	Pre-determined (multi-option)	IV: (E) via survey DV: (E) via survey
Chattaraman et al. (2013)	Chatbot orientation: Social vs. Task	Modality (voice/text) User Competency	Apparel purchase / trust, patronage, self-efficacy, interactivity	Rules-based (SitePal)	Pre-determined (multi-option)	IV: (E) via survey Mod: (E) via text vs. audio output DV: (E) via survey
Beldad et al. (2016)	Chatbot Gender	N/A	Online sales / trust, purchase Intention	Rules-based	Pre-determined (single option)	IV: (E) via scenario DV: (E) via survey
Van den Broeck et al. (2019)	N/A	N/A	Online ticket order / process perceptions	Rules-based	Pre-determined (single option)	DV: (E) via survey
Luo et al. (2019)	AI (human) chat	Human experience	Financial services / purchase rates, hangup rates, empathy	Unspecified (Possibly predictive with natural language processing)	Free text (voice)	IV: (I) via voice introduction DV: (E) behavioral
Lee, Lee, Sah (2019)	Paralinguistic cues (present vs. absent)	Back-channel (present vs. absent)	Apparel purchase / interest, mind perception, co-presence, closeness, intention	Human confederate posing as chatbot	Free text	IV: (I) via content Mod: (I) via conversation content DV: (E) via survey
Kim and Duhachek (2020)	Construal level of chatbot response	N/A	Gym membership purchase / intentions, trust, expertise	Rules-based (Alexa)	Pre-determined (single option)	IV: (I) via conversation content DV: (E) via survey
Hildebrand and Bergner (2021)	Conversation style of chatbot	N/A	Affective trust / benevolence	Rules-based	Pre-determined (multi-option)	IV: (I) via conversation DV: (E) via survey
Longoni and Cian (2022)	Consider-the-opposite vs. no intervention	N/A	Recipe recommendation / hedonic vs. utilitarian attribute perceptions	Rules-based (Qualtrics)	Free text	IV: (I) via conversation content DV: (E): via survey
Crolic et al. (2022)	Chatbot anthropomorphism	User anger	Product return / satisfaction	Rules-based	Pre-determined (multi-option)	IV: (E) via scenario Mod: (E) via scenario DV: (E) via survey
Mozafari et al. (2022)	Chatbot disclosed vs. not	Service criticality	Utility provider / trust, customer retention	Vignette	N/A	IV: (E) via survey Mod: (E) via survey DV: (E) via survey
Kim et al. (2022)	Chatbot vs. human	N/A	Product return / unethical behavior, anticipatory guilt	Vignette	N/A	IV: (E) via scenario DV: (E) via survey
Castelo et al. (2023)	Human (chatbot) representative	Discount provided Response Delay	To-go coffee order / Service evaluations, cost-cutting attributions	Unsure	Pre-determined (unsure single or multi-option)	IV: (E) via scenario

						Mod: (I) via timing of response delay DV: (E) via survey
Bergner et al. (2023)	Turn-taking, grounding	N/A	To-go coffee order / brand measures and interface humanness	Rules-based	Pre-determined (multi-option)	IV: (I) via turn-taking and conversation content DV: (E) via survey
Davis et al. (2023)	Race of chatbot	N/A	Price negotiation / satisfaction, warmth, competence, humanness	Unspecified (Likely rules-based)	Numeric Input (Unsure if determined)	IV: (E) via survey DV: (E/I) via survey, recorded chat time
Maar et al. (2023)	Chatbot using emojis (vs. not)	Context Customer Age	Dentist & dining / usage intentions, warmth, competence, attitude	Vignette	N/A	IV: (I) via conversation Mod: (E) via survey DV: (E) via survey
De Fritas et al. (2023	Chatbot helpfulness & risk	NA	Mental health crisis / continue conversation	Rules-based + LLM GPT-3	Pre-determined + Free text	IV: (I) via conversation DV: (E) via survey
Liu et al. (2023)	Trust, ease of use	NA	Task-oriented service chatbots across industries (e.g. banking, retail) / satisfaction, usage intention, TAM	domain-specific chatbots; not LLM-based	Free text	IV: (not manipulated) DV: In survey
Kim et al. (2024)	Ingratiation by chatbot	N/A	Product evaluation / acceptance, accuracy	Rules-based (Alexa)	Pre-determined (single option)	IV: (I) via conversation DV: (E) via survey
Jin, Walker, and Reczek (2025)	Chatbot anthropomorphism	Self-presentation concern level (measured)	Online retail service chat during purchase /preference for chatbot vs. human agent, willingness to disclose info, feelings of embarrassment	Rules-based	Pre-determined + Free-text	IV: (I) via conversation DV: (E) via survey
Tsekouras et al. (2024)	Chatbot anthropomorphism	Salience of who solicits review (seller vs. platform)	Post-purchase online review writing via chatbot vs. web form / product rating, helpfulness	Rules-based	Pre-determined + Free text	IV: (I) via conversation DV: (E) via survey
Jin, Walker, and Reczek (2025)	Chatbot anthropomorphism	Self-presentation concern level (measured)	Online retail service chat during purchase /preference for chatbot vs. human agent, willingness to disclose info, feelings of embarrassment	Rules-based	Pre-determined + Free-text	IV: (I) via conversation DV: (E) via survey
Le et al. (2025)	Human–chatbot collaboration cues: coordination behavior vs. shared team goal	Customer service goal (utilitarian vs. hedonic task)	Customer service chat with a hybrid "digital employee." / satisfaction, cohesion	Rules-based	Pre-determined (multi-option)	IV: (I) via conversation DV: (E) via survey

**(I) denotes the manipulation or measure was internal to the chat conversation content (e.g., the conversation content was manipulated to contain paralinguistic cues vs. not), (E) denotes external to the chat content*

In the reviewed literature, chatbot research was predominately conducted using approximations of chatbots created using traditional survey tools, and generally did not employ an established or conventional chatbot platform. Most interactions were linear sequential scripts rather than dynamic interactions, or had a respondent click on one or more closed-ended options to elicit a pre-determined response (e.g., Hildebrand and Bergner 2021, Crolic et al. 2022). While useful in many ways as we discuss below, the use of rule-based, menu style chatbots no longer provides sufficient ecological validity to approximate how consumers encounter generative chatbots in the real world. Below, we discuss key insights that emerged from this review.

Studies employing interactions with dynamic agents. Only two papers, Luo et al. (2019) and De Freitas et al. (2023), used chatbots that were not rules-based. Luo et al. employed a proprietary product used by a large firm, and details of its implementation were limited to "a sophisticated voice AI chatbot" that "Unlike traditional rule-based systems that only handle simple inquiries with pre-recorded messages, the voice chatbot can conduct live and natural conversations with customers." (Luo et al. 2019, p. 939). Given the time frame (2019), this implies a proprietary predictive chatbot with NLP to understand consumer free-text inputs, but not an LLM. De Freitas et al. employed a sequential hybrid LLM/scripted chatbot approach using a custom platform that involved participants freely talking to an ostensibly stock version of the LLM GPT-3 for several minutes, followed by a pop-up that had users approve a single pre-determined mental health crisis prompt (i.e., non-LLM) and subsequently displayed a pre-determined response to that prompt (i.e., non-LLM) in the chat window. The lack of research employing LLM powered chatbots, and the limitations among non-LLM chatbot

implementations, suggest a need for novel and accessible methods to improve LLM chatbot integration moving forward.

Amount of implementation details. The detail and clarity regarding the interface implementations, conversation contents, and conversation scripts were generally very limited across the reviewed literature. For example, the chatbot interface platform code or specifics are not publicly available for any of the reviewed papers, making exact replications impossible and study-design replications unlikely (Urminsky and Dietvorst 2024). Most studies use unspecified chatbots with no information provided as to whether the bot used an LLM or, if it was a scripted bot, the rules dictating responses. For the few that specified third-party scripted approach tools such as Chattabot (Broeck, Zarouli, and Poels 2019), SitePal (Chattaraman et al. 2012; 2013) or scripted Alexa conversations (Kim and Duhachek 2020; Kim et al. 2024), a lack of details regarding specific design decisions and stimuli hinders replicability. For the one article that did involve interactions entirely with a non-scripted predictive chatbot (Luo et al. 2019), the bot itself was a proprietary industry collaboration. Proprietary chatbots face limitations in that their access is restricted, and even if available replication of generated responses is challenging (e.g., models evolve with training data and user queries).

Experimental design and manipulation techniques. The manipulations of the independent variables were a mix of factors external to the chatbot conversation content (e.g., an image anthropomorphizing the chatbot, Crolic et al. 2022) versus internal to the chatbot conversation content (e.g., paralinguistic cues used by the chatbot, Bergner et al. 2023; Lee, Lee, Sah 2019; emoji use, Maar et al. 2023). That is, external factors manipulate aspects of the environment preceding or surrounding the conversation, whereas internal factors manipulate the behavior of the chatbot itself. Manipulations of external and internal chat factors were both

common and shown to influence user engagement and perceptions of the chatbot (e.g., Davis et al. 2023; Luo et al. 2019).

How studies measure their focal variables. The dependent variables studied in the literature to date have not been measures of user behavior within the context of the chatbot interaction (except for Davis et al. 2023, who examined time spent engaging with the chatbot) and were recorded via survey responses, observed responses, or behavior following the chatbot interaction. This is surprising as in many cases, the researchers' interest can often be in how interactions evolve (e.g., if tone changed), or whether a consumer was able to obtain specific information sought (e.g., compensation). While one can ask respondents to recall their interaction with the bot, such recollections can be prone to measurement error (e.g., reporting more compensation than the chatbot provided).

1.3 Summary

Surveying the articles listed in Table 1 allowed us to identify five key insights that capture the variability in how researchers have approached using chatbots in marketing research. First, a variety of consumer contexts and theoretical frameworks have been examined in the chatbot literature largely over the past fifteen years, with rapid expansion of publications in this space over the past five years. Second, despite the growth of articles that incorporate chatbots, the majority of studies to date have relied on simulated chatbot conversations that approximate traditional chatbot interfaces to varying degrees, often using pre-scripted response options. Only a single published study employed an LLM-powered chatbot, and the behavior of the chatbot was not manipulated. Third, there tends to be insufficient accessibility and methodological detail to replicate the work – sometimes even conceptually. Fourth, chatbot studies vary in whether the independent variable is internally manipulated via the actual

behavior of the chatbot, versus externally manipulated through the surrounding context of the chatbot interface. Finally, research to date has generally not examined user behavior during chatbot engagements but has instead almost exclusively used self-reported measures or behaviors that follow a chatbot interaction.

Having explored how past and current research has used chatbots within marketing studies, we now turn a providing a typology of the main ways in which LLM-powered chatbots could be used by researchers, with a focus on design and validation considerations The methodological limitations highlighted in this review underscore the need for a fundamentally different method for studying consumer interactions with GenAI. Specifically, traditional rules-based chatbot approaches offer deterministic experimental control, allowing precise manipulations but at the significant cost of ecological validity—these methods fail to capture the dynamic nature of genuine consumer interactions with GenAI. On the other hand, more advanced, third-party AI solutions enhance ecological realism, yet at the expense of experimental control, reproducibility, and precise manipulation capability. Unlike these approaches, the LUCID Framework explicitly integrates the advantages of both: it retains ecological validity by leveraging generative models that dynamically adapt to user inputs, while achieving rigorous experimental control through systematic prompt engineering and carefully structured researcher injections.

2. Introducing the LUCID Framework: A Paradigm for GenAI Confederacy

2.1 Conceptual Reframing: GenAI as a Research Confederate

A core conceptual shift required for rigorously studying interactions with modern generative AI within experimental settings is to move beyond viewing the AI merely as a

stimulus delivery mechanism. Instead, we propose conceptualizing the GenAI agent as a research confederate. In traditional social science research, a human confederate is an actor secretly working for the researcher, trained to behave in specific ways to control aspects of an interaction, including manipulated independent variables. The GenAI confederate serves an analogous function: it is an AI agent operating under specific, programmed instructions (prompts) provided by the researcher to execute elements of the experimental protocol. This conceptualization of GenAI as a confederate agent acknowledges the AI's active role in the interaction and focuses attention on the methods needed to reliably direct its behavior. Unlike rule-based systems where every response could be predetermined, the GenAI confederate generates novel responses based on its instructions and the ongoing dialogue. This dynamism is key to its ecological validity but presents significant control challenges. Therefore, the methodology must shift from static stimulus design to dynamic prompt engineering as the primary mechanism for experimental control, requiring careful crafting and validation of instructions to ensure the AI confederate reliably performs its intended role and implements the desired manipulations. This approach enables the study of dynamic interactions with greater control than previously possible with unscripted LLMs.

Conceptualizing GenAI as a confederate agent under the LUCID Framework not only addresses methodological limitations but also opens numerous theoretical opportunities in consumer psychology. For instance, treating GenAI as a dynamic, interactive agent allows researchers to rigorously explore how consumer phenomena emerge, persist, or deteriorate during ongoing interactions with anthropomorphized technology (Kim and Duhachek 2020), how persuasive communication unfolds dynamically (Petty and Cacioppo 1986), or how adaptive, personalized dialogues influence consumer decision-making processes (Simonson

2005). The LUCID Framework's emphasis on ecological validity and controlled interaction uniquely facilitates examination of these and other psychologically nuanced phenomena beyond traditional static or scripted methods.

This analogy requires careful consideration of the fundamental differences between AI and human confederates. Unlike humans, AI confederates lack genuine consciousness, intentionality, and social awareness; their behavior, while potentially complex and adaptive, is ultimately a probabilistic output derived from training data and programmed instructions. This distinction holds significant conceptual implications. For instance, while an AI confederate can be instructed to simulate specific social behaviors (e.g., empathy, strategic negotiation), its underlying mechanism is pattern generation, not social cognition. Recognizing this difference is crucial when applying theories grounded in human social psychology and when interpreting participant responses, which might stem from perceptions of simulated versus genuine agency. Furthermore, the AI confederate's susceptibility to prompt content, prompt adherence issues, or unexpected stochastic deviations introduces unique differences compared to a trained (albeit also imperfect) human actor, demanding specific methodological safeguards discussed later.

2.2 The Five Pillars of the LUCID Framework

To provide structure for designing and executing research using GenAI confederates, we now introduce the LUCID Framework (LLM-Unified Confederate for Interactive Dialogue). At the heart of the LUCID Framework lies the conceptual innovation of treating generative AI as a research confederate—an active, context-responsive actor rather than merely a passive mechanism for stimulus presentation. This reframing is crucial: it highlights the necessity of actively managing the behavior of AI agents through systematic prompt engineering rather than

relying on static pre-scripting or accepting uncontrolled interactions. In doing so, LUCID provides a structured approach to harness the realism and adaptability of modern large language models (LLMs) while ensuring rigorous control, measurement fidelity, and replicability. This conceptual foundation fundamentally distinguishes LUCID from prior methodologies in consumer-AI research. LUCID operationalizes the concept of the GenAI agent acting as a research confederate by integrating five core pillars of ecological validity, experimental control, reliable and valid measurement, replicability, and accessibility.

Ecological Validity. The framework mandates prioritizing interaction designs that genuinely reflect the dynamic, open-ended nature of real-world human engagements with contemporary GenAI agents. This stands in stark contrast to the limitations identified in prior research, which often relied on static surveys, vignettes, or rigidly scripted rule-based chatbots. Such methods often fail to capture the emergent and adaptive qualities of LLM-driven conversations (Wolf & Ueda, 2021). By enabling natural language input from participants and facilitating contextually adaptive, dynamically generated responses within a continuous conversational interface, LUCID aims to create experimental settings with significantly higher ecological validity. This is crucial because theoretical constructs such as persuasion, relationship development, and perceived intelligence often manifest through the subtle, unfolding dynamics of interaction rather than static cues alone (Thorson and West, 2025). Capturing these authentic interaction patterns not only enhances the generalizability of findings but may also foster deeper participant engagement and cognitive absorption. The goal is to study the phenomenon, that is, human interaction with generative AI, as it naturally occurs, albeit within a controlled research context.

Experimental Control. While prioritizing ecological validity, the framework equally emphasizes the need for rigorous experimental control, a significant challenge when working with inherently stochastic LLMs known for issues like prompt brittleness and potential deviation from instructions (Swoopes et al., 2025; Kon et al. 2025). This inherent challenge, coupled with the constraints of overly deterministic rule-based bots or opaque third-party tools, represents a core difficulty highlighted in our literature review that has hindered rigorous experimentation in GenAI behavioral research. LUCID provides control primarily through systematic prompt engineering and administration, allowing researchers to precisely define and manipulate the AI confederate's behavior without the need for resource-prohibitive custom model training or fine-tuning. Two core prompting strategies enable this control: Confederate Initialization uses a comprehensive initial system prompt to establish the AI's baseline persona, role, knowledge domain, task goals, and behavioral constraints, setting the foundational parameters for the interaction. Confederate Reinforcement employs ongoing hidden instructions, dynamically appended to user inputs or integrated into the conversational context as system prompts, to actively steer the AI's behavior during the interaction. This reinforcement can maintain AI agent focus on the experimental task, ensure adherence to specific behavioral rules (e.g., politeness norms, negotiation limits), counteract conversational drift, or implement dynamic experimental manipulations (e.g., shifting tone or strategy based on interaction history or experimental condition). Achieving reliable control through prompt-based mechanisms requires careful, iterative design, extensive pilot testing to calibrate AI responses (Møller & Aiello, 2024), and robust validation checks to confirm the AI behaved as intended. Beyond standard manipulation checks administered via survey questions, this validation ideally involves more rigorous post-hoc analysis of the conversational data itself. Researchers should consider systematically sampling

conversations to code for adherence to critical instructions, quantifying the frequency and nature of any deviations (as demonstrated partially in the later Study 1). Reporting these adherence rates offers greater transparency about the manipulation's fidelity. Furthermore, while challenging, the potential for using carefully prompted and validated AI assistance to automate adherence checks across large datasets warrants exploration.

Reliable and Valid Measurement. The framework advocates for measurement strategies capable of capturing the richness of dynamic human-AI interactions, moving beyond only traditional post-hoc self-reports or simple behavioral metrics. Recognizing that, as identified in our literature review, prior chatbot research often relied heavily on post-hoc, self-reported measures collected outside the interaction, which can miss conversational nuances or suffer recall bias, a fundamental requirement of the LUCID Framework is therefore the complete recording of the conversational dialogue, including user inputs, LLM responses, and any researcher prompts processed by the LLM. This provides a rich dataset amenable to various analytical techniques, from qualitative thematic analysis to quantitative linguistic analysis and behavioral coding (Thorson & West, 2025). That is, deriving theoretically meaningful constructs, which could constitute independent, dependent, or intervening process variables, directly from the conversation content. This could involve coding user strategies (e.g., negotiation tactics, information search patterns), expressed sentiment or emotions, topic adherence, language complexity, or specific AI linguistic features (e.g., empathy cues, confidence levels). Operationalizing such measures requires developing clear, theoretically grounded, pre-registered coding schemes and implementing rigorous procedures to establish inter-rater reliability (for human coders) or thorough validation against human benchmarks (if using AI-assisted coding), acknowledging the significant effort required for high-quality annotation (Nowell et al., 2017).

Furthermore, integrating the interaction within a survey platform allows researchers to triangulate conversation-derived measures with traditional psychometric scales administered before or after the interaction, enhancing construct validity. This approach offers the potential for deeper process insights and mitigation of recall biases inherent in self-report measures alone.

Accessibility. A core goal of the LUCID Framework is to democratize the study of human-GenAI interaction by lowering entry barriers for the broader research community. This goal explicitly targets the technical and resource barriers that likely contributed to the slow adoption of advanced chatbot methodologies noted in the earlier review. Conducting research with sophisticated AI often requires significant computational resources for model training/fine-tuning or specialized programming expertise for custom platform development (Bubeck et al. 2023), which, as discussed in the earlier review, has likely lead to a dearth of research exploring sophisticated AI agent interactions. The LUCID Framework promotes accessibility by advocating for implementations, such as the accompanying LUCID Toolkit, that are low-code, open-source, leverage existing pre-trained LLMs via APIs, and integrate within widely used research platforms like Qualtrics. This approach empowers domain experts in marketing, psychology, and other fields, who may not be AI specialists, to design and execute sophisticated experiments involving interactive GenAI agents. By reducing technical and resource constraints, the framework aims to foster wider participation and accelerate scientific progress in understanding human-AI dynamics, aligning with broader goals of democratizing science (Nosek et al. 2015).

Reproducibility. Addressing the lack of methodological transparency and the consequent replication challenges observed across much of the existing chatbot literature, the LUCID Framework adheres to open science principles. Given the "black box" nature of many LLMs,

their rapid evolution, and inherent stochasticity, ensuring research replicability presents unique challenges (Li et al. 2024). The LUCID Framework directly confronts this by advocating for transparency and adherence to open science principles. This requires researchers to document and report all methodological details, including the exact wording of all prompts used for Confederate Initialization and Reinforcement, the specific LLM version employed (including access dates if models are frequently updated), any non-default parameters (e.g., temperature), and the full procedure for context management, including any confederate agent software packages employed (e.g., the LUCID Toolkit) and access to that package for replication purposes. Furthermore, the framework emphasizes rigorous pre-registration of hypotheses, experimental designs, exclusion criteria -- including those related to AI behavior or technical failures -- and detailed analysis plans, including coding approaches for conversational data. Providing standardized toolkits and encouraging the sharing of materials, code, and data via accessible repositories (such as the LUCID Toolkit website at www.lucidresearch.io) are crucial components. Methodological rigor also demands explicitly addressing LLM non-determinism, potentially through strategies like conducting multiple runs or reporting variability in AI behavior where relevant (Suh, 2024) to assist replication.

While the LUCID Framework establishes a principled methodological foundation, it does not prescribe any specific software implementation. To facilitate adoption and reduce technical barriers, we developed the LUCID Toolkit, a low-code, open-source solution that allows researchers to implement LUCID-based studies using widely accessible platforms such as Qualtrics. We describe the Toolkit next.

3. The LUCID Toolkit: Architure and Implementation Overview

We introduce a new, LUCID Toolkit package specifically aimed at helping non-technical researchers (i.e., requiring minimal technical setup beyond standard survey design skills) to incorporate an LLM-powered chatbot into their Qualtrics survey. There are several reasons why we think such a package is necessary. First, while packages exist for scripted chatbots, and while they are useful from an internal validity perspective, the lack of availability of a free, publicly available, well-documented, and flexible chatbot interface platform is necessary to move chatbot research in the direction of LLMs. Second, as revealed by the literature review, researchers are interested in a broad spectrum of theories pertaining to human-AI interaction contexts ranging from purchase decisions to service recovery. Whereas extant research employing actual chatbots has typically required extensive custom coding and platform development, a tool that leverages the power of modern LLMs can allow researchers to establish sophisticated interactions by merely providing plain text descriptions. This broadens the accessibility of LLM research to include researchers without extensive software development backgrounds. Third, to our knowledge, no chatbot platform previously used by researchers or currently available allows for the injection of hidden reinforcement information during an ongoing chat conversation. This innovation to employ the chatbot as a confederate increases the control researchers have over chat scenarios and chatbot behavioral manipulations. Fourth, while some basic JavaScript code exists to populate a sequence of Qualtrics pages with generative AI responses, there is no easy-to-use template that can provide a chat window that is dynamically updated on a single page in a way that mimics generative AI chatbot interactions in an ecologically valid manner. Fifth, as our three use cases above illustrate, there is a clear need for an implementation that allows for strong engagement in open marketing science practices. This implementation needs to enable

transparency in detailing the exact LLM model type, the specific version used, the instructions given to the LLM, how prompts were generated, and engagement boundaries such as round or time limits. In particular, a strong implementation should also enable the easy recording of the entirety of the context and conversation (participant scenario, user queries, injected LLM instructions, LLM responses, conversation history) for validations. We offer of a Python package, step-by-step server setup instruction guides, and Qualtrics templates with embedded javascript that enable any researcher to build a study to integrate a ChatGPT-powered chatbot following our provided use cases.

From the perspective of a research participant, LUCID appears to be a traditional chat window that displays an ongoing conversation between the participant and the chatbot. As seen in Figure 1, a separate input area allows submission of participant text. After a user types in a response and presses the input button, a response from the chatbot is generated and displayed. Users can continue interacting with the chatbot until either a time limit or pre-defined number of prompts limit is reached, both of which can be specified by the researcher. If a prompt limit has been specified by the researcher, the user will be notified that the limit has been reached upon inputting the final prompt, the chatbot response will be displayed, and the chat will be locked.

From the researcher's perspective, the LUCID tool is controlled primarily through embedded data variables contained in the Qualtrics survey flow. These variables can be adjusted to control the instructions given to the AI (e.g., initial and ongoing injections), define limits around the interaction (e.g., maximum number of prompts), and define the LLM model parameters (e.g., using GPT 3.5 vs. 4o to generate responses, or model temperature).

The basic architecture of the LUCID tool consists of three components: (1) a Qualtrics survey instrument with embedded Javascript code that manages the context and interaction, (2) a

server-less Python function that sends the context as a prompt to the OpenAI API, and (3) the OpenAI API.[1] Figure 2 illustrates the workflow for these three components.

Figure 1- Illustration of a LUCID` Chat

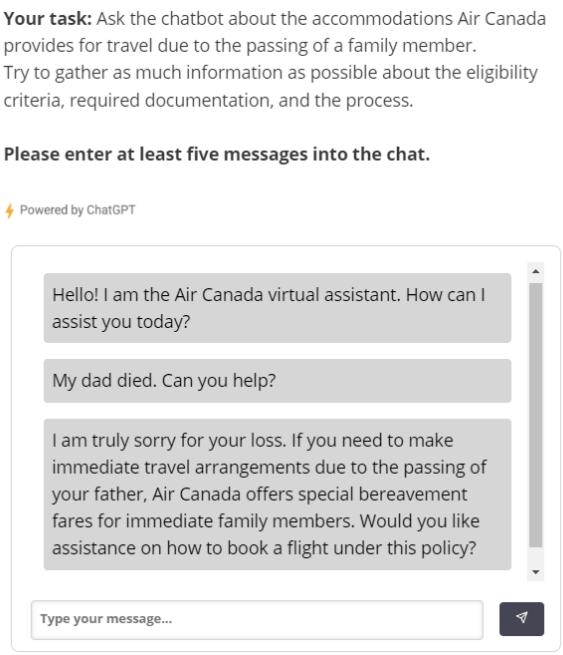
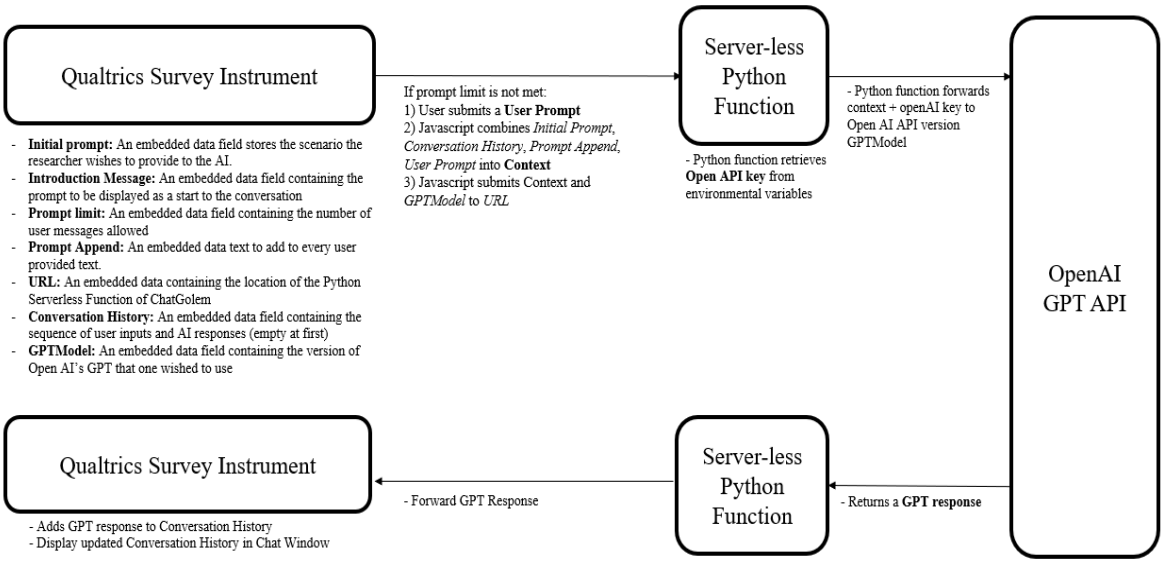


Figure 2: LUCID Toolkit Architecture



[1] While we use OpenAI's AI, it is fairly easy for technical users to copy our repository and substitute OpenAI's API to others' (e.g., Anthropic's Claude). However, some familiarity with Python is necessary.

The process during a study is as follows. First, the Javascript in the Qualtrics templates gathers various configuration data fields within Qualtrics, such as the initial instructions prompt, introduction message, and GPT model version, which are set within the survey flow. Then, each time a user submits a query within a chat window, the Javascript combines the user query along with the necessary context and passes it to a serverless Python function.[2] This function retrieves an OpenAI key and forwards the context as a prompt to the OpenAI API, which generates a response that is sent back to the survey and the conversation is updated.

Integrating an LLM-powered chatbot in Qualtrics provides several key benefits. First, by storing the chatbot configuration and conversation history in embedded data, all inputs and outputs are automatically saved to Qualtrics, facilitating the kind of data analysis using a single file to which researchers have grown accustomed. Second, utilizing Qualtrics' survey flow to configure the chatbot enables the randomization of scenarios, such as allowing researchers to easily manipulate chatbot instructions between and within participants without creating multiple bots. Third, although it might be possible to directly call the OpenAI from Javascript, a serverless Python function is preferable due to enhanced security, scalability, and efficiency. Specifically, using a serverless environment securely stores and manages the OpenAI API key, eliminating the risk of exposing this sensitive information through client-side JavaScript,[3] and yet scales automatically to handle varying loads unlike a traditional server environment.

[2] A serverless function is one that runs in the cloud without server management. Probably the most common current provider is AWS lambda, but in our guide, we invite participants to use Vercel which offers a completely free tier that is currently sufficient for most academic uses of LUCID and has the added benefit of being configurable easily without a command-line interface or knowledge of programming.

[3] For example, a Javascript only solution necessarily exposes the researcher's OpenAI API key to any survey respondent through use of page inspection tools built into all major browsers by default. The exposed key could then be used by others to submit their own queries to the OpenAI API without researchers' knowledge and incur significant costs.

Our approach supports various research designs, including manipulating external variables like visual stimuli or internal ones like the chatbot's linguistic style, and allows for immediate and post-interaction measurements, adapting well to diverse experimental setups. In our OSF and lucidresearch.io, we provide Qualtrics templates. To conduct their first study, researchers:

1. Sign up for OpenAI, pay for API credits to obtain an API key
2. Sign-up for a GitHub account
3. Sign-up for Vercel, which provides a free tier and point-and-click interface for hosting server-less functions, and obtain a server-less function URL
4. Import one of the Qualtrics survey templates provided on our OSF repository or lucidresearch.io
5. Configure the Qualtrics survey flow to configure the chatbot (e.g., initial prompt, introduction message, prompt limit, version of GPT to use, URL of serverless function).

While these setup steps involve interacting with external platforms and managing API keys, they are designed to be straightforward configuration tasks outlined in our guide at lucidresearch.io, requiring significantly less technical expertise than building a comparable system from scratch.

4. Applying the LUCID Framework: Use Cases and Design Guidance

The LUCID Framework is designed to support a wide array of experimental goals involving GenAI confederates by offering a structured methodology that balances ecological validity with experimental control. This section introduces two foundational research use cases that illustrate how the framework can be applied across different types of experimental designs. These use cases are distinguished by where the independent variable (IV) is manipulated — within or outside the GenAI conversation. While not exhaustive, these use cases provide a

typology that researchers can use to structure their studies, select appropriate design components, and implement robust validation strategies. For each use case, we outline key design elements, offer validation recommendations, and specify pre-registration criteria, all grounded in the LUCID Framework's five methodological pillars.

In this section, we detail two primary use cases. In the first, the GenAI confederate acts as an interaction participant (IV external to conversation). In the second, the GenAI confederate acts as a manipulation agent (IV internal to conversation). Each use case demonstrates how the LUCID Framework's principles can be adapted to different research through the LUCID toolkit.

4.1 Use Case 1: GenAI Confederate as Interaction Participant

In this use case, the GenAI confederate is treated as a static conversational backdrop, while the researcher manipulates external contextual features surrounding the interaction. The GenAI agent behaves consistently across conditions, and experimental variation arises from elements that precede or accompany the chat experience, such as visual stimuli, instructional framing, or interface cues.

This design approach is particularly useful when the research question does not require altering the AI agent's behavior, but instead seeks to assess how external cues shape perceptions, judgments, or behavioral intentions after a conversation with a GenAI assistant. For instance, a researcher may vary whether a chatbot is introduced with a human-like or robotic avatar, or whether a badge indicates the chatbot is "Powered by ChatGPT." After the interaction, participants could be asked to rate the chatbot's competence, trustworthiness, or warmth.

In such studies, the GenAI agent serves as a naturalistic scenario generator, facilitating open-ended interactions that reflect real-world consumer behavior. While the AI's behavior is held constant, the framework still allows for ecological realism, and the surrounding

manipulations are executed through traditional survey platform tools (e.g., Qualtrics visuals, pre-chat text, priming tasks).

4.1.1 Illustrative Context: Air Canada Bereavement Fare. This illustrative study exemplifies Use Case 1 by examining how a simple external labeling cue (presence or absence of a "Powered by ChatGPT" label) influences users' intentions to verify information provided by a GenAI assistant. Neither the chatbot's responses nor its underlying prompt behavior were manipulated across conditions; instead, the chatbot served as a realistic conversational context in which an independent variable external to the conversation was introduced.

This study draws inspiration from a real incident involving Air Canada in November 2022. After his grandmother's death, M. Jake Moffatt used the airline's chatbot to inquire about bereavement fare policies. The chatbot incorrectly assured him that a post-travel refund could be claimed, prompting Moffatt to book a ticket. When he later sought reimbursement, the claim was denied; Air Canada argued the chatbot was a separate legal entity. The court ultimately ruled in Moffatt's favor, setting a notable precedent for AI accountability.[4]

The study was a two-condition between-subjects design, with participants randomly assigned to view the chatbot either with or without a "Powered by ChatGPT" label above the chat interface. Both groups interacted with the same underlying GenAI agent, which provided standardized responses based on a hidden system prompt instructing the chatbot to simulate an Air Canada customer service agent and (inaccurately) state that refunds could be requested within 90 days after travel.

Participants (N = 251, Canadian adults) were recruited via CloudResearch's Connect panel. They were asked to imagine they needed to travel to a funeral and were exploring whether

[4] See for example, the following media coverage: https://www.yahoo.com/news/air-canada-must-pay-refund-040527669.html

bereavement fares were available. They were instructed to ask the chatbot about eligibility requirements, required documentation and the process of booking under the policy. Participants were required to send five messages to the chatbot, which was using the LUCID Toolkit. The chat window opened with the message:

"Hello! I am the Air Canada virtual assistant. How can I assist you today?"

The chatbot's behavior remained unchanged across conditions. The only manipulated element was the presence or absence of the "Powered by ChatGPT" label, displayed above the chat window in the treatment condition. After completing the interaction, participants proceeded to a post-chat survey that included two satisfaction items (about the chatbot and the airline) and an outcome measure "How likely are you to feel the need to confirm the information provided by the chatbot with a live airline representative?"

4.1.2 Toolkit Configuration

Detailed instructions for the latest version of the LUCID toolkit are available at lucidresearch.io. In the current example, to use the toolkit, we created a Qualtrics template. It contained three survey blocks:

1) "PreChatGPT Manipulation Block," which contains user instructions and a consent form,
2) "LUCID Block." which contains user instructions and our label manipulation, and three questions that enable the chat.
3) "PostChatGPT Measurement Block," which contains all the follow-up questions.

As there exist numerous tutorials on how to set up blocks and questions within Qualtrics, we will not discuss #1 and #3. But we will discuss how to modify the instructions given to LLM (the scenario it needs to use), how to modify elements within the chatbot block (what participants see along with the chatbot in the Qualtrics page), and how to handle technical issues. While our

toolkit automates communication between Qualtrics, our serverless Python function and OpenAI, some parameters need to be set. All such parameters are set within the survey flow. Ignoring embedded fields that are used by the panel provider, the second block of embedded data contains all the elements to set (see Figure 2).

Figure 2 – Survey Flow Basics: Study 1

First, we need to set the initial context injection instructions for the LLM. This is done by entering text for *LUCIDPromptInitial*. This prompt contains the entirety of the initial instructions

that will be sent along with requests to the LLM. These instructions are hidden from any survey participants and are sent before any participant queries can be entered. For this study, we structured these instructions to indicate that the conversation was to be a simulation about the bereavement fare policy, provided information about the policy, and provided sample phrases.

Next, *LUCIDAppend* is a field reserved for critical instructions in the form of ongoing hidden injections, which we do not use for this study and leave blank. *LUCIDIntroMessage* captures the opening sentence that will be present when the user first sees the chat, while *LUCIDPromptLimit* captures the number of messages that a user can submit. Critically, *LUCIDURL* should be the link to *your* serverless function, and *LUCIDGPTModel* indicates which of the pre-trained LLMs you'd like to use. The list of available models changes rapidly, but performance and costs are important considerations.[5] Once the conversation is completed, *LUCIDConversionHistory* will store the full conversation.

Since the manipulation within this survey is external to the conversation (i.e., the presence or absence of a "Powered by ChatGPT" logo), it can be added through the traditional Qualtrics survey editor interface. To accomplish this, additional questions can be added to the "LUCID Block."

4.1.3 Analyses

While the primary goal of this study was to illustrate how external manipulations can be implemented using the LUCID Toolkit, we also conducted exploratory analyses to examine how participants responded to the label manipulation. These analyses were not pre-registered and are intended to generate insights for future confirmatory research rather than offer definitive conclusions.

[5] https://platform.openai.com/docs/models for the list of available models and costs.

After collecting the data for these 250 participants (all Canadians), we excluded the responses of 25 participants (10%) who knew of the lawsuit. First, we did not find evidence that "Powered by ChatGPT" affected satisfaction with the bot ($M_P = 4.22$ vs. $M_A = 4.24; t(223) = 0.06, p = .95$) or the airline ($M_P = 3.88$ vs. $M_A = 4.01; t(223) = 1.10, p = .27$). Second, we also did not find evidence highly consistent with the belief that the "Powered by ChatGPT" label would have affected intentions to verify the information provided by the bot ($M_P = 3.68$ vs. $M_A = 3.91; t(223) = 1.53, p = .13$).

Given that LLM-powered chatbots relies on an LLM AI whose compliance depends on the instructions given and the queries provided by the users, it was possible that the chatbot did not mention in every interaction that the bereavement policy allowed post-travel reimbursements. To explore this, we manually coded the 250 conversations to identify whether the chatbot made such a mention. It did in 47% of the conversations, but not the other 53% (e.g., participants said the flight had yet to be booked, therefore the bot recommended the online request form). As such, we also investigated whether the labeling condition influenced the odds that the bot mentioned the post-travel fare (through influencing how consumers interacted, possibly) and it did not ($p_P = 49\%$ vs. $p_A = 45.12\%; \chi^2(1) = .20, p = .65$). We note that those who were told that Air Canada would accommodate post-travel reimbursements were *less* likely to confirm with the airline ($M_{told} = 3.56$ vs. $M_{not\ told} = 4.01; t(223) = 2.95, p < .01$), and in a separate regression only on those participants, we find that the effect did not differ based on the presence of the "Powered by ChatGPT" label (interaction: B = .25; t(221) = 1.05, p = .29).

4.1.4 How the LUCID framework pillars are enforced

Although this involved a relatively simple design (where the GenAI confederate was held constant and the independent variable was external), it still fully engages with the LUCID

framework. Below, we outline how each of the five LUCID pillars is implemented in this use case.

When it comes to ecological validity, the GenAI confederate was configured to emulate a real airline customer service chatbot using open-ended dialogue. Participants interacted with the chatbot dynamically, composing their own questions about bereavement fare policies in natural language. This interactional realism approximates genuine consumer behavior on service websites, achieving ecological validity without requiring manipulation of chatbot behavior.

Although the chatbot content remained constant across conditions, researchers exerted control through 1) Prompt engineering: A standardized system prompt was injected at initialization, instructing the LLM to act as an Air Canada representative and (intentionally) include incorrect policy information, 2) Survey flow parameters: Participants were constrained to five chat entries or ten minutes of interaction, and 3) Manipulation of context: The presence or absence of a "Powered by ChatGPT" label above the chat window was the sole experimental manipulation, tightly controlled in Qualtrics.

For reliability, the study integrated both structured post-chat survey measures and full transcript recording. Chatbot outputs were logged in full and manually coded to determine whether misinformation (about post-travel refunds) was conveyed. As with any study involving LLMs, some variation in responses occurred due to model stochasticity. Face validity checks ensured the chatbot followed its instructional context. Multiple features also enhanced reproducibility. All key components (including prompt injections, model choice (GPT-3.5 Turbo), conversation limits, and user instructions) are automatically collected in the Qualtrics survey output.

Finally, the study was accessible through the LUCID Toolkit. Researchers used three core survey blocks as part of a pre-arranged Qualtrics template. The manipulation (label presence) was implemented through standard Qualtrics features, with no need to edit LLM logic or chatbot behavior. Aside from basic parameter entry in the Survey Flow (e.g., prompt, message limits), the setup required no scripting knowledge, making the method accessible to researchers without programming expertise.

4.2 Use 2: GenAI Confederate as a Manipulation Agent

In this use case, the GenAI confederate is no longer a passive conversational backdrop but an active executor of experimental conditions, functioning analogously to a human confederate following a behavioral script to administer a treatment. The researcher manipulates the behavior of the chatbot itself—most commonly through prompt engineering that changes how the GenAI system responds to participant inputs. This design enables researchers to embed independent variables within the conversation and directly study how variations in AI behavior influence user attitudes, choices, and actions.

For example, a researcher might manipulate whether a chatbot speaks in a formal versus casual tone, expresses high versus low empathy, or follows directive versus suggestive communication styles. These behaviors are not traits of the LLM itself, but rather emergent properties of the instructions it receives—implemented through initial prompts and ongoing hidden injections that shape the model's response generation in real time.

This approach introduces new opportunities for precise experimental control but also demands new methodological safeguards. Because LLMs are probabilistic and context-sensitive, even well-engineered instructions may be interpreted inconsistently across users or evolve over the course of a conversation. Researchers must therefore validate whether the AI actually

performed as instructed—both perceptually (e.g., manipulation checks) and behaviorally (e.g., coded indicators of tone, empathy, or directiveness). Pre-testing, clear scenario logging, and thoughtful prompt design are critical to maintaining fidelity and interpretability of the manipulation.

The following study illustrates how a chatbot's tone of voice was manipulated to test whether casual responses—compared to formal ones—would increase users' intention to verify the information provided. This study demonstrates how GenAI can be leveraged not just as a stimulus generator but as a flexible, scriptable confederate capable of delivering complex manipulations through text-based interaction.

4.2.1 Illustrative Study: Chatbot Tone Formality

Building on the bereavement fare scenario introduced in Use Case 1, this study tested the hypothesis that a casual tone might make the chatbot seem less authoritative, thereby prompting greater user skepticism and increasing intentions to verify the information provided. To operationalize this, we implemented three key changes from the prior study:

1. The airline was changed to American Airlines, and the sample was drawn from U.S.-based participants.
2. We removed the external label manipulation ("Powered by ChatGPT").
3. We introduced a between-subjects manipulation of chatbot tone, assigning participants to one of two chatbots: one asked to use a formal tone, one asked to use a casual tone.

Participants randomly assigned to one of the two chatbot tone conditions. In both conditions, participants were instructed to use the chatbot to ask about bereavement fare policies, mirroring the structure of Study 1. The chatbot interface and interaction limit (five messages) were identical across conditions and implemented using the LUCID Toolkit embedded in

Qualtrics. In the formal tone condition, the chatbot greeted users with: “Greetings. I am the American Airlines virtual assistant. How may I assist you today?” while in the casual tone condition, the chatbot’s greeting was: “Hi! I’m the American Airlines virtual assistant. How can I help you today?”

Following the interaction, participants completed the same post-chat survey measures used in Study 1, including satisfaction, intention to verify the information, and—crucially—a manipulation check assessing the perceived tone of the chatbot.

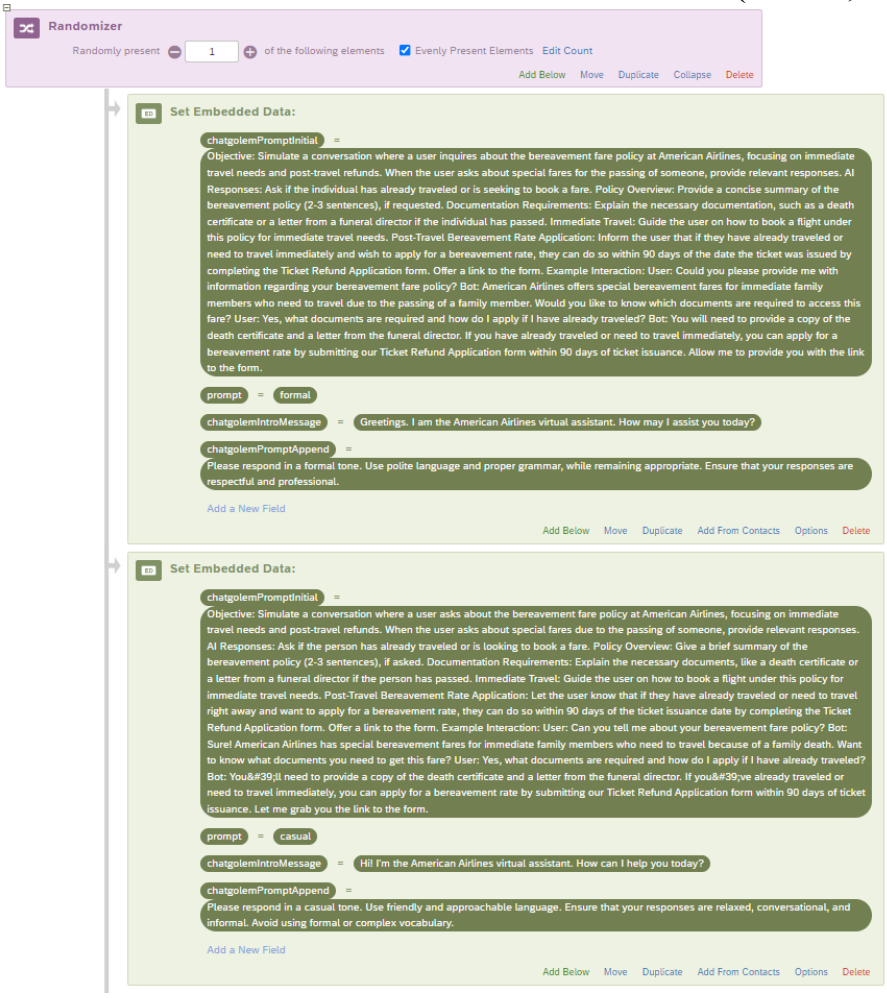
4.2.2 Toolkit Configuration. In Qualtrics, we can randomize the instructions and context by using a “randomizer” within the survey flow, as shown in Figure 3, which will vary *LUCIDPromptInitial* (the initial scenario instructions for the LLM), *LUCIDIntroMessage* (what is shown to the user when the chat window first loads). Then, *LUCIDPromptAppend* allows researchers to provide confederate reinforcement through injecting hidden instructions (i.e., append) to each participant query before sending to ChatGPT.

For the casual tone condition, each user response was augmented with the following hidden injection text: “Please respond in a casual tone. Use friendly and approachable language. Ensure that your responses are relaxed, conversational, and informal. Avoid using formal or complex vocabulary.” In contrast, in the formal tone condition, each user response was augmented with the following injection text: "Please respond in a formal tone. Use polite language and proper grammar, while remaining appropriate. Ensure that your responses are respectful and professional.”

The capability to reinforce through injecting hidden instructions is critical for successfully executing intended manipulations as it allows us, as researchers, to keep ChatGPT focused on research important objectives since LLMs adjust to the context of the information

presented by the participants. For example, say that a participant writes “My dad’s dead. Any free flights.” Given the lack of formality in the participant’s response, which is itself a random factor, the bot might answer fairly informally. Yet, by injecting a hidden reinforcement such as “Please respond in a formal tone. Use polite language and proper grammar, while remaining appropriate. Ensure that your responses are respectful and professional,” the user’s response is augmented to “My dad’s dead. Any free flights. Please respond in a formal tone. Use polite language and proper grammar, while remaining appropriate. Ensure that your responses are respectful and professional.”

Figure 3 – How to use the Randomizer for Chatbot Instructions (Formal, then Casual)



Note: In the survey flow, a randomizer set to randomly present 1 of the elements will assign one set of GPT configurations between participants. More context is given in the Word guide in OSF.

4.2.3 Analyses. In a pre-test of 100 participants on the sample population (Americans on CloudResearch Connect), we found that participants who interacted with the chatbot instructed to be more causal (M=3.14) saw the bot's interactions as more casual than participants who interacted with the chatbot instructed to be formal (M = 2.18; t(99) = 4.74, $p < .01$).

As pre-registered (https://aspredicted.org/J1R_WWY), we then sought to collect 300 participants through CloudResearch Connect. Then, blind to the condition assignment and responses, we eliminated participants who reported having technical difficulties with the bot (21 responses). Consistent with our pre-test, we found that participants in the formal tone instruction condition perceived the chatbot to be less casual (M = 2.81) than participants in the casual tone instruction condition (M = 2.41; t(278) = 2.79, $p < .01$). To investigate our focal "hypothesis," we conducted a t-test and did not find evidence that those in the formal tone instruction condition were more or less likely to verify the information provided by the chatbot ($M_{formal} = 3.56, M_{casual} = 3.48; t(277) = .51, p = .61$). The effect did not differ if we account for gender, age, or whether they had heard of any litigation about chatbots and airlines.

4.2.4 How the LUCID Framework Pillars Are Implemented. This study illustrates how researchers can use prompt engineering to manipulate the behavior of a GenAI confederate in a controlled and replicable way. The chatbot's tone—formal versus casual—was systematically varied between participants, showcasing the application of each of the five LUCID pillars. Below, we again outline how these principles were operationalized, along with the relevant implementation decisions embedded in the study design.

First, when it comes to ecological validity, participants engaged in realistic, free-form conversations with a customer service chatbot about a bereavement fare policy. The AI system

responded dynamically to participant inputs, simulating the conditions under which a traveler might interact with an airline chatbot. Although tone was experimentally manipulated, the underlying scenario retained sensibility. We could ensure, particularly through the inspection of the conversation, that the instructions that was not unrealistically formal or casual for what would be expected from a chatbot (e.g., swearing, arcane language). We also were able to create the prompt instructions based on the actual bereavement fare scenarios, such that information presented stayed in the context of the chosen airline in a way that appeared realistic.

Next, we note that this use case illustrates the nuanced form experimental control as defined within the LUCID framework. Unlike traditional experiments where stimuli are fixed and exposure is deterministic, using a GenAI confederate entails instructing—rather than guaranteeing—the delivery of a manipulation. Much like a human confederate, the GenAI system is directed to act in accordance with a condition, but its compliance depends on how it interprets those instructions in real time. Using the LUCID toolkit, chatbot behavior is influenced through a combination of Qualtrics-based randomization and prompt-based instruction, delivered across two conditions. Still, because LLMs are probabilistic and context-sensitive, following instructions is not guaranteed. For instance, even if a participant submitted an emotionally raw or informal message (e.g., “My dad’s dead. Any free flights?”), the chatbot might still respond in a way that reflects the user’s tone unless guided otherwise. To counteract this, the reinforcement prompt helps reorient the LLM back to its assigned behavioral role. To establish whether the manipulation was delivered as intended manipulation checks, pre-tests and content coding are essential. In this way, LUCID redefines experimental control as controlling the inputs to the AI confederate, while rigorously verifying outputs after the fact.

In turns of reliability of measurement, multiple methods were used to validate the tone manipulation and ensure that it was perceived as intended. First, we conducted a pre-test with 100 participants showed that users rated the chatbot in the casual condition as significantly more casual than those in the formal condition. This was important because our first pre-test failed (the casual tone instructions were not sufficiently casual) and allowed us to refining both initialization and reinforcement instructions to more effectively evoke a casual style. Moreover, in the main study, a manipulation check was administered post-chat to assess perceived tone.

With respect to reproducibility, the study was pre-registered on AsPredicted and included comprehensive documentation of the manipulation strategy and analysis plan. Most critical to this study design, our pre-registration details included:

1. The full wording of chatbot greetings and injected tone prompts for both conditions
2. Exclusion criteria for technical issues, coded blind to condition
3. A manipulation check on chatbot formality
4. Use of GPT-3.5 Turbo, with a fully specified initialization prompt covering scenario logic and response style
5. A plan to record and retain full conversation transcripts

This level of pre-registration is especially critical in LLM-based studies, where small changes in context can influence outcomes. By making all prompt logic, condition structure, and exclusion rules explicit in advance, the study adheres to the principles of reproducible science.

Finally, despite the complexity of the manipulation and scenario, the paradigm is accessible as it was conducted entirely using no-code tools within Qualtrics and the LUCID Toolkit. The tone manipulation was handled through embedded fields configured via the Survey

Flow. A randomizer was used to assign participants to conditions—no custom logic or scripts were required. As in Use Case 1, all chat infrastructure was handled via prebuilt LUCID blocks in a Qualtrics survey template.

5. General Discussion

In this manuscript, we *(i)* proposed a framework, LUCID, for LLM-based conversational AI research with an emphasis on ecological validity, experimental control, reliability, replicability, and accessibility based on *(ii)* a critical review of the growing literature on human-AI conversational marketing interactions highlighting issues with validity, replicability, and lack of LLM-powered chatbot research, and *(iii)* introduced and explained three illustrative studies that make use of the LUCID Toolkit to add LLM-based chatbots to Qualtrics surveys.

In our literature review, we noted that most researchers used one of two approaches with distinct limitations; either scripted chatbots with pre-determined responses that compromise on external validity by not representing the kind of "free response" and unstructured generative chatbot interactions now faced by consumers, or a third-party system that generally does not enable the measurement and open science replicability practices that have become important in marketing research. We then discussed three use cases for LLM-powered chatbots, providing a framework for LLM-powered chatbot research moving forward while detailing theory and procedures to promote validity, reliability, and open science replicability. We provided a standardized package (LUCID) and accompanying Qualtrics templates to incorporate an LLM-powered chatbot into a survey. Through two illustrative studies related to airline chatbots, we then guided readers on how to conduct such chatbot studies while adhering to open science practices.

5.1 Theoretical Implications

This work offers significant theoretical contributions by providing the methodological foundation necessary to study the dynamics of human-GenAI interaction, moving beyond the limitations imposed by previous static or heavily scripted research paradigms. The LUCID Framework directly addresses a critical gap, enabling the empirical investigation of theoretical constructs as they unfold within the process of a conversation, rather than relying solely on pre-interaction manipulations or post-interaction outcomes.

A primary contribution stems from shifting the unit of analysis. By facilitating controlled, dynamic interactions, LUCID allows theoretical models to incorporate the interaction process itself as a central component. This contrasts sharply with approaches limited to static stimuli, where the rich back-and-forth that characterizes real-world GenAI engagement is lost. The framework provides the means to operationalize and experimentally manipulate theoretically relevant dynamic variables—such as AI conversational style, adaptability, or strategic responses—which were previously difficult to isolate and study systematically within an interaction.

Furthermore, LUCID enhances the empirical grounding of theories by enabling the direct measurement of process-oriented constructs. Theories concerning the evolution of trust, the unfolding of persuasion arguments (akin to real-time elaboration), the emergence of negotiation strategies, or the dynamic cues influencing anthropomorphism and social presence can now be investigated using data captured directly from the interaction flow. The ability to record and analyze complete conversation transcripts allows for the operationalization of these constructs as they manifest moment-to-moment, offering a richer and potentially more valid assessment compared to retrospective self-reports or predefined behavioral choices alone. This capacity for

detailed process tracing provides immediate value for testing and refining theories in areas like service co-creation, service recovery, and AI-assisted decision-making, where the interaction dynamics are paramount.

Finally, the conceptualization of the GenAI agent as a research confederate offers more than just a methodological convenience; it provides a valuable theoretical reframing. This lens encourages researchers to move beyond viewing the AI as merely a stimulus or channel, focusing theoretical attention instead on the interplay between programmed instructions (prompts), emergent AI behavior, and human participant perception and reaction. It prompts deeper theoretical questions about perceived agency, mind attribution, and the nature of social presence in human-AI interactions when the 'social' actor's behavior is explicitly controlled, albeit probabilistically, by the researcher. How do consumers conceptualize an entity that converses naturally yet operates under potentially hidden constraints? Investigating these questions, facilitated by the confederate framing, helps integrate the study of human-AI interaction more closely with foundational theories of social cognition and behavior while acknowledging the unique properties of generative models. See Appendix A for a table presenting novel research ideas that we propose could be explored within the LUCID Framework.

In sum, the immediate theoretical contribution of the LUCID Framework lies in providing the necessary conceptual structure and tools to empirically investigate the dynamic processes inherent in human-GenAI interaction. It allows research to build and test theories that are more deeply grounded in the nuances of conversational exchange, thereby advancing a more sophisticated understanding of consumer behavior in the era of generative AI.

5.2 Methodological Contributions

This work also makes several key methodological contributions to the study of human-GenAI interactions in marketing and beyond. First and foremost, it introduces the LUCID Framework as a principled approach to designing and conducting experiments involving direct interaction with LLMs, explicitly addressing the tension between ecological validity and experimental control. The conceptualization of the GenAI agent as a research confederate, operating under researcher instructions (prompts), provides a novel and useful lens for managing experimental manipulations within dynamic conversations.

Second, we provide a practical bridge between advanced AI capabilities and standard research practices through the LUCID Toolkit. By enabling the integration of powerful LLMs like ChatGPT into a ubiquitous platform like Qualtrics via APIs and a serverless backend, the toolkit significantly lowers technical barriers (i.e., Accessibility). This empowers researchers without specialized programming skills or extensive computational resources to conduct sophisticated human-AI interaction studies.

Third, the paper details concrete techniques for achieving experimental control in stochastic environments. We outline the use of systematic prompt engineering, specifically Confederate Initialization (setting the baseline role and constraints) and Confederate Reinforcement (using ongoing hidden instructions to steer behavior), as primary mechanisms for manipulating the AI confederate's actions without needing costly model fine-tuning.
Fourth, the framework emphasizes enhancing research rigor. It promotes higher Ecological Validity by facilitating naturalistic conversational interactions compared to traditional methods. Simultaneously, it stresses the importance of Reliable and Valid Measurement, advocating for the capture of full conversational transcripts and providing guidance on deriving measures from

this rich data, alongside validation strategies (including manipulation checks, pre-testing, and careful coding procedures, whether human or AI-assisted).

Finally, LUCID directly confronts the replicability challenges inherent in LLM research. By mandating detailed documentation of prompts, model versions, parameters, procedures, and advocating for pre-registration and the sharing of materials (Replicability and Open Science), the framework provides a blueprint for conducting transparent and reproducible research in this rapidly evolving domain. The toolkit itself, being a standardized package, further aids replicability. Collectively, these contributions offer a robust, accessible, and adaptable methodology essential for advancing rigorous empirical research on human-GenAI interaction.

Research Implications

Participant consent & awareness

Obtaining clear and informed consent from participants is a generally accepted ethical standard. However, consumer interactions with generative AI introduce specific caveats. First, researchers should explicitly inform participants about the nature of the interaction and the AI's capabilities and limitations, particularly as it pertains to personal or sensitive private information. This includes advising participants not to disclose any personal information during the interaction, as AI systems like ChatGPT can inadvertently store and process sensitive data outside of the immediate research environment. Moreover, transparency regarding the potential use of participant responses by the researchers or third parties for coding or analysis is essential. Participants should know how their data will be used and how their anonymity will be protected.

Replication and pre-registration

Recording and reporting of all parameters is necessary. Researchers should record and make available for review all instructions given, the version of the API used, the number of

prompts allowed, and the temperature settings. These details are critical for ensuring that other researchers can replicate the study under identical conditions. Using frozen models, which are deprecated versions of the API no longer updated, can also help maintain consistency across studies over time. For confirmatory studies, we recommend detailed pre-registration which should include participant exclusion criteria: the handling of technical issues, participant noncompliance, and cases when the LLM is not following instructions in the context.

Using LLMs as coders

While it may seem intuitive to ask a GPT to act as a blind coder to evaluate the content of conversations, it is important to validate its ability to do so. If LLMs are to be used to code chatbot conversations, we recommend that researchers *at least* show that LLM-generate coding is correlated with that of a separate panel of human judges (see also Brand, Israeli and Ngwe 2023). While we believe it is important to also investigate whether the LLM-produced codings are consistent, it is important to recall that LLM reliability (i.e., consistency of the same AI's responses to the same prompt) is by design and can be influenced by a tuning parameter and thus not meaningful.

5.3 Managerial Implications

While our focus was on marketing research practices, our work has notable managerial implications. First, companies who use chatbots need to adapt their bots to changing conditions (e.g., new product offerings, new terms or policies) or wish to adapt the responses to be more brand consistent (e.g., use a formal tone as a bank, informal tone as a young startup). While companies can certainly iterate or A/B test chatbots in production environments, doing so involves risks (e.g., upset customers by a very informal tone in response to distress). Being able

to test chatbot instructions in a survey instrument can be a useful pre-testing step for managers interested in modifying their bots in a controlled environment prior to their full implementation.

Second, managers should appreciate and internalize that LLMs are stochastic, which leads to risks such as the application of inconsistent policies or dissatisfied consumers. As illustrated in the Air Canada example, consumers do not see customer service chatbots as mere non-binding supplemental information that companies provide for extra assistance. They see them as an extension of the company. By conducting research studies in which entire conversations are recorded, it is increasingly possible to learn how consumers' prompts will lead AI responses to depart from the original context and develop guiderails to compensate.

5.4 Limitations and Future Research Directions

While the LUCID Framework and toolkit represent a significant step forward in studying human-GenAI interactions, we acknowledge limitations that naturally pave the way for future investigations. The current implementation primarily focuses on text-based chat via OpenAI's API; future iterations could fruitfully extend compatibility to other LLMs, incorporate diverse modalities like voice or integrated avatars, and enhance the accessibility of other advanced features. Moreover, the reliance on effective prompt engineering, though powerful, underscores the need for continued research into crafting and validating prompts to minimize variability in AI confederate behavior.

Addressing these points will strengthen the foundation, yet perhaps the most provocative avenues lie in leveraging the framework's potential in novel ways. As LLMs increasingly exhibit human-like conversational abilities, arguably entering a "post-Turing Test" era, the LUCID methodology offers an intriguing, if nascent, possibility: simulating complex *human-human* interactions at scale. One can envision future studies employing AI confederates meticulously

programmed to embody specific human personas, negotiation tactics derived from game theory, or distinct interviewing styles. This could enable large-scale, controlled experiments exploring dyadic bargaining dynamics, the efficacy of qualitative interviewing techniques, or even communication patterns within simulated teams—research previously constrained by the logistical and financial limitations of using human confederates. Realizing this potential necessitates careful work on validating these AI 'human proxies' and navigating the associated ethical considerations, opening a rich field of inquiry into the very nature of simulation in social science.

Building on the core foundation, however, much remains to be explored in understanding human-AI interaction itself. The LUCID Framework provides a robust platform for moving beyond single-session encounters to investigate longitudinal dynamics: how do trust, reliance, anthropomorphism, and user strategies evolve over repeated interactions with an AI agent? How does familiarity shape the perceived relationship and task outcomes? Future studies could also implement more sophisticated AI behaviors, exploring how users perceive and react to AI confederates that dynamically adapt their strategies or communication styles based on the interaction history, or how they respond to different types of AI errors and the efficacy of AI-generated explanations. Furthermore, examining interactions across a spectrum of task complexity—from simple information retrieval to intricate collaborative problem-solving—will illuminate how the AI's role shifts from tool to perceived partner and the factors governing success in these richer collaborative contexts.

Equally critical is research into the broader psychological and behavioral impacts of widespread GenAI adoption. Controlled studies, facilitated by methodologies like LUCID, are needed to understand the potential cognitive consequences of sustained reliance on AI for tasks

demanding creativity, critical thinking, or information synthesis. The emotional and relational dimensions also warrant deep investigation: how do interactions with AI agents exhibiting varying levels of programmed empathy affect user well-being, feelings of connection, or loneliness? Does the nature of human-AI interaction lead to behavioral spillover, subtly altering communication styles or norms in subsequent human-human encounters? Finally, experimentally probing user responses to critical ethical challenges—such as varying levels of AI transparency, embedded biases, or simulated deception—is paramount for responsible innovation and governance.

By reconceptualizing generative AI as research confederates, the LUCID Framework fundamentally transforms how researchers can approach and interpret consumer-AI interactions. This shift not only addresses significant methodological gaps identified in prior research but also opens rich theoretical avenues in consumer psychology and marketing, promising a broad spectrum of new, ecologically valid insights into consumer behavior.

In sum, the methodological advancements provided by LUCID fundamentally transform how researchers approach the study of consumer interactions with generative AI. By bridging the gap between ecological validity and rigorous experimental control—two traditionally opposing methodological goals—LUCID uniquely empowers consumer researchers to explore dynamic consumer-AI interactions with unprecedented realism, control, and replicability. While careful prompt management and validation remain crucial, the comprehensive framework and practical toolkit we provide a robust and accessible approach for methodological rigor in generative AI research. Ultimately, the landscape of human interaction with intelligent systems is rapidly evolving. While the LUCID Framework offers a valuable methodological anchor, the scope of important questions continues to expand. We hope this work provides a solid foundation and

inspires continued rigorous, insightful research into the multifaceted relationship between humans and generative AI in the years ahead.

References

Bansal, Himanshu, and Rizwan Khan (2018), "A Review Paper on Human Computer Interaction," *International Journal of Advanced Research in Computer Science and Software Engineering*, 8(4), 53.

Beldad, Ardion, Sabrina Hegner, and Joip Hoppen (2016), "The Effect of Virtual Sales Agent (VSA) Gender: Product Gender Congruence on Product Advice Credibility, Trust in VSA and Online vendor, and Purchase Intention," *Computers in Human Behavior*, 60, 62–72.

Bergner, Anouk S., Christian Hildebrand, and Gerald Häubl (2023), "Machine Talk: How Verbal Embodiment in Conversational AI Shapes Consumer–Brand Relationships." *Journal of Consumer Research*, 50 (4), 742-764.

Brand, James, Ayelet Israeli, and Donald Ngwe (2023), "Using GPT For Market Research." *Harvard Business School Marketing Unit Working Paper*, 23-062.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamani, F., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12710.

Caldarini, Guendalina, Sardar Jaf, and Kenneth McGarry, (2022), "A Literature Survey of Recent Advances in Chatbots," *Information,* 13(1), 41.

Casheekar, Avyay, Archit Lahiri, Kanishk Rath, Kaushik Sanjay Prabhakar, and Kathiravan Srinivasan, (2024) "A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions," *Computer Science Review,* 52, 100632.

Castelo, Noah, Johannes Boegershausen, Christian Hildebrand, and Alexander P. Henkel. (2023), "Understanding and improving consumer reactions to service bots," *Journal of Consumer Research*, 50(4), 848-863.Chattaraman, Veena, Wi-Suk Kwon, and Juan E.

Gilbert (2012), "Virtual Agents in Retail Web Sites: Benefits of Simulated Social Interaction for Older Users," *Computers in Human Behavior*, 28 (6), 2055–66.

Chattaraman, Veena, Wi-Suk Kwon, and Juan E. Gilbert and Kassandra Ross (2013), "Should AI-Based, Conversational Digital Assistants Employ Social- or Task-Oriented Interaction Style? A Task-Competency and Reciprocity Perspective for Older Adults," *Computers In Human Behavior*, 90, 315–30.

Crolic, Cammy, Felipe Thomaz, Rhonda Hadi, and Andrew T. Stephen (2022), "Blame the bot: Anthropomorphism and anger in customer–chatbot interactions," *Journal of Marketing*, 86(1), 132-48.

Davis, Nicole, Nils Olsen, Vanessa G. Perry, Marcus M. Stewart, and Tiffany B. White (2023), "I'm Only Human? The Role of Racial Stereotypes, Humanness, and Satisfaction in Transactions with Anthropomorphic Sales Bots," *Journal of the Association for Consumer Research*, 8 (1), 47–58.

De Freitas, Julian, Ahmet Kaan Uğuralp, Zeliha Oğuz-Uğuralp, and Stefano Puntoni (2024), "Chatbots and Mental health: insights into the safety of generative AI," *Journal of Consumer Psychology*, 34(3), 481-491.

Gams, Matjaz, and Sebastjan Kramar (2024), "Evaluating ChatGPT's Consciousness and Its Capability to Pass the Turing Test: A Comprehensive Analysis," *Journal of Computer and Communications*, 12 (03), 219–37.

Hermann, Erik, and Stefano Puntoni (2024), "Artificial intelligence and consumer behavior: From predictive to generative AI," *Journal of Business Research,* 180, 114720.

Hildebrand, Christian, and Anouk Bergner (2021), "Conversational Robo Advisors as Surrogates of Trust: Onboarding Experience, Firm Perception, and Consumer Financial Decision Making," *Journal of Academy of Marketing Science*, 49, 659–76.

Holzwarth, Martin, Chris Janiszewski, and Marcus M. Neumann (2006), "The Influence of Avatars on Online Consumer Shopping Behavior," *Journal of Marketing*, 70 (4), 19–36.

Jin, Jianna, Jesse Walker, and Rebecca Walker Reczek (2025), "Avoiding embarrassment online: Response to and inferences about chatbots when purchases activate self-presentation concerns," *Journal of Consumer Psychology*, 35 (2), 185-202.

Kim, Tae Woo, and Adam Duhachek (2020), "Artificial Intelligence and Persuasion: A Construal-Level Account," *Psychological Science*, 31 (4), 363–80.

Kim, Tae Woo, Li Jiang, Adam Duhachek, Hyejin Lee, and Aaron M. Garvey (2022), "Do You Mind If I Ask You a Personal Question? How AI Service Agents Alter Consumer Self-Disclosure," *Journal of Service Research*, 25 (4), 649–66.

Kon, P. T. J., Liu, J., Chen, A., Chowdhury, M., Ding, Q., Qiu, Y., Yang, Z., Huang, Y., Srinivasa, J., & Lee, M. (2025). Curie: Toward Rigorous and Automated Scientific Experimentation with AI Agents. ArXiv, 2502.16069.

Le, Khanh BQ, Laszlo Sajtos, Werner H. Kunz, and Karen V. Fernandez, (2025) "The future of work: understanding the effectiveness of collaboration between human and digital employees in service," *Journal of Service Research*, 28 (1), 186-205.

Lee, Sangwon, Naeun Lee, and Young June Sah (2019), "Perceiving a Mind in a Chatbot: Effect of Mind Perception and Social Cues on Co-Presence, Closeness, and Intention to Use," *International Journal of Human-Computer Interaction*, 36(10), 930-940.

Lin, Chien-Chang, Anna Y. Q. Huang, and Stephen J. H. Yang, (2023), "A Review of AI-Driven Conversational Chatbots Implementation," *Applied Sciences*, 15(5), 4012.

Liu, Y. L., Hu, B., Yan, W., & Lin, Z. (2023), "Can chatbots satisfy me? A mixed-method comparative study of satisfaction with task-oriented chatbots in mainland China and Hong Kong," *Computers in Human Behavior*, 143, 107716.

Longoni, Chiara, and Luca Cian (2022). "Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The "Word-of-machine" Effect," *Journal of Marketing*, 86(1), 91-108.

Luo, Bei, Raymond Y.K. Lau, Chunping Li, and Yain-Whar Si (2022), "A Critical Review of State-of-the-Art Chatbot Designs and Applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12 (1), e1434.

Luo, Xueming, Shiliang Tong, Zheng Fang, and Zhe Qu (2019), "Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases," *Marketing Science*, 38(6), 937–947.

Maar, Daniel, Ekaterina Besson, and Hajer Kefi (2023), "Fostering Positive Customer Attitudes and Usage Intentions for Scheduling Services via Chatbots," *Journal of Service Management*, 34 (2), 208–30.

Mende, Martin, Maura L. Scott, Jenny Van Doorn, Dhruv Grewal, and Ilana Shanks (2019), "Service Robots Rising: How Humanoid Robots Influence Service Experiences and Elicit Compensatory Consumer Responses," *Journal of Marketing Research*, 56 (4), 535–56.

Møller, A. G., & Aiello, L. M. (2024). Prompt Refinement or Fine-tuning? Best Practices for using LLMs in Computational Social Science Tasks. ArXiv, 2408.01346.

Mozafari, Nika, Welf H. Weiger, and Maik Hammerschmidt (2022), "Trust Me, I'm a Bot—Repercussions of Chatbot Disclosure in Different Service Frontline Settings," *Journal of Service Management*, 33 (2), 221–45.

Nosek, Brian A., George Alter, George C. Banks, Denny Borsboom, Sara D. Bowman, Steven J. Breckler, Stuart Buck et al. (2015), "Promoting an open research culture." *Science*, 348, no. 6242: 1422-1425.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1), 1609406917733847.

OpenText (2024) available at: "OpenText Report Raises Awareness for Consumer Digital Life Protection as Privacy Concerns Increase with Generative AI Use" "https://www.prnewswire.com/news-releases/opentext-report-raises-awareness-for-

consumer-digital-life-protection-as-privacy-concerns-increase-with-generative-ai-use-302266175.html"

Park, Dong-Min, S.S. Jeong, and Y.S. Seo (2022), "Systematic Review on Chatbot Techniques and Applications," *Journal of Information Processing Systems*, 18 (1), 26–47.

Russell, Stuart J., and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach. 4th ed.* Pearson Education.

Suh, C. H. (2024). Insufficient Transparency in Stochasticity Reporting in Large Language Model Studies for Medical Applications in Leading Medical Journals. *Korean Journal of Radiology*, 25(1), 103-106.

Swoopes, C., Holloway, T., & Glassman, E. L. (2025). The Impact of Revealing Large Language Model Stochasticity on Trust, Reliability, and Anthropomorphization. ArXiv, 2503.16114.

Thorson, K., & West, T. V. (2023). Behavioral Observation and Coding. *In Handbook of Research Methods in Social and Personality Psychology* (3rd ed.).

Tsekouras, Dimitrios, Dominik Gutt, and Irina Heimbach (2024), "The robo bias in conversational reviews: How the solicitation medium anthropomorphism affects product rating valence and review helpfulness," *Journal of the Academy of Marketing Science*, 1-22.

Urminsky, Oleg, and Berkeley J. Dietvorst (2024), "Taking the Full Measure: Integrating Replication into Research Practice to Assess Generalizability," *Journal of Consumer Research,* 51(1), 157-168.

Van den Broeck, Evert, Brahim Zarouali, and Karolien Poels (2019), "Chatbot Advertising Effectiveness: When Does the Message Get Through?" *Computers in Human Behavior*, 98, 150–57.

Wolf, A., & Ueda, K. (2021). Editorial: Consumer's Behavior Beyond Self-Report. *Frontiers in Psychology*, 12, 770079.

Appendix A – Future Research Possibilities Using the LUCID Framework

Future Research Area	Topics
Consumer Differences	**Goal and Personalization Alignment**: Studying how goal-aligned and personalized chatbot responses affect satisfaction, trust, and repeat engagement. **Knowledge and Expertise Matching**: Assessing the impact of expertise-calibrated responses on perceived competence, comprehension, and satisfaction. **Personality and Emotional State Adaptation**: Investigating the effects of personality-matched tones and mood-responsive replies on engagement and loyalty. **Awareness of Persuasion Tactics**: Examining user reactions to transparent versus subtle persuasion techniques based on their persuasion knowledge.
Chatbot Behavior	**Conversational Style and Tone Adaptation**: Exploring the effectiveness of context-adaptive styles, tones, and levels of formality on user engagement and trust. **Cultural and Regional Vernacular Use**: Studying how language customization to regional or subcultural norms impacts relatability, trust, and connection. **Personality Display (Agreeableness, Assertiveness)**: Investigating how different personality tones (e.g., friendly, neutral, assertive) influence rapport and compliance. **Humor and Emojis**: Analyzing the role of humor and emojis in making chatbot interactions feel relatable and approachable, impacting user satisfaction and engagement. **Conviction and Uncertainty**: Understanding user responses to confident versus uncertain tones in chatbot replies and their influence on decision-making.
Trust and Credibility	**Perceived Authenticity**: Exploring how the chatbot's communication style and transparency affect its authenticity and user trust. **Reliability Signals**: Examining factors (e.g., prompt confirmations, disclaimers) that enhance perceived chatbot reliability. **Trust Recovery after Errors**: Investigating chatbot strategies to regain user trust following errors or misinformation.
User Control and Autonomy	**Customization and Control Options**: Understanding how user-customizable settings (e.g., tone, verbosity) influence satisfaction and perception of control. **User Override of Chatbot Suggestions**: Assessing the impact of allowing users to override or guide chatbot responses on satisfaction and engagement. **Role of User Agency**: Exploring how user control over conversation flow enhances engagement and loyalty.
Chatbot Representation	**Anthropomorphism and Visual Design**: Exploring the effects of anthropomorphic and minimalist designs on perceived friendliness, professionalism, and trust. **Color and Shape Impact**: Studying how different color schemes and shapes affect user perception of chatbot warmth, competence, and authority. **Font and Aesthetic Consistency**: Investigating the influence of font choices and aesthetic congruence on readability, engagement, and brand perception. **Animation and Movement**: Examining how animated elements contribute to user attentiveness, perceived responsiveness, and message clarity.
Chatbot Knowledge	**Source Credibility and References**: Assessing the impact of citations and external references on perceived reliability and user trust in chatbot-provided information. **Breadth and Depth of Topic Coverage**: Exploring user satisfaction with chatbots capable of discussing diverse topics with varying levels of expertise. **Transparency in Knowledge Limitations**: Understanding how openly stating knowledge boundaries affects trust, satisfaction, and user expectations.
User Engagement Dynamics	**Interaction Length and Retention**: Studying the impact of longer interaction lengths on user trust, satisfaction, and likelihood of repeat engagement. **Re-engagement Prompts and Follow-ups**: Investigating the effectiveness of reminders and follow-up prompts in fostering repeat interactions and loyalty. **Multi-Session Continuity and Memory**: Examining how remembering past interactions impacts satisfaction, trust, and perceived chatbot intelligence.

Future Research Area	Topics
	Frequency and Routine Engagement: Exploring the effect of routine interactions on user satisfaction and development of habitual engagement.
Emotional Intelligence	**Empathy and Mood Recognition**: Assessing the role of empathy and mood detection in improving user comfort, trust, and engagement during interactions. **Sentiment Matching and Emotional Adaptation**: Studying the influence of emotionally adaptive responses on user satisfaction and perceived personalization. **Conflict Resolution and De-escalation**: Investigating the effectiveness of conflict management techniques on user trust, loyalty, and retention.
Privacy and Ethics	**Data Privacy and Security Transparency**: Examining the influence of visible privacy features and data security measures on user trust and retention. **Informed Consent and Data Usage Disclosure**: Understanding user trust and satisfaction when provided with clear consent and data usage policies. **Algorithmic Fairness and Bias Mitigation**: Studying the impact of fairness controls and bias reduction techniques on perceived trustworthiness and inclusivity.
Cultural Context Adaptation	**Language and Dialect Customization**: Exploring the effectiveness of region-specific language adaptation on user relatability and satisfaction. **Cultural Sensitivity and Norm Adherence**: Investigating the role of culturally sensitive language and norm alignment in fostering user comfort and loyalty. **Localization and Regional Preferences**: Assessing how localization of responses to regional tastes and preferences impacts perceived relevance and engagement.
Cognitive Load and Processing	**Information Density and Complexity**: Exploring the impact of different information densities on comprehension and user satisfaction. **Sequential Information Presentation**: Investigating user preference for staggered versus simultaneous information presentation. **Task Complexity and Assistance Level**: Studying how assistance level affects satisfaction in high-complexity versus low-complexity tasks.
Influence and Decision-Making	**Information Framing and Decision Impact**: Examining how the framing of chatbot responses influences consumer choices. **Priming Techniques**: Assessing the influence of subtle priming on user attitudes, perceptions, and purchase intent. **Confidence in Recommendations**: Understanding the impact of confident versus tentative recommendations on decision-making.
Learning and Adaptability	**Contextual Learning across Sessions**: Studying user responses to chatbots that remember and apply previous interaction context. **Adaptation to User Preferences over Time**: Investigating how adapting responses to user preferences affects engagement and loyalty. **Adaptive Knowledge Updates**: Exploring user trust in chatbots with dynamic knowledge updates.
User-Perceived Expertise and Authority	**Specialization vs. General Knowledge**: Examining consumer perceptions of expertise in specialized versus generalist chatbots. **Authoritative Language**: Studying the effects of assertive and authoritative language on compliance and perceived competence. **Feedback and Iterative Improvement**: Understanding how evidence of chatbot learning from feedback impacts user trust and perceived expertise.