

## Research Article

Darius Hennekeuser\*, Daryoush Daniel Vaziri, David Golchinfar, Dirk Schreiber and Gunnar Stevens

# Balancing automation and control: user perceptions of tool invocations in conversational agents

<https://doi.org/10.1515/icom-2025-0053>

Received October 22, 2025; accepted December 11, 2025;

published online February 6, 2026

**Abstract:** Advances in large language models (LLMs) have enabled conversational agents (CAs) to invoke external functions dynamically, yet little is known about how different tool-calling (TC) strategies affect user experience. This study compares two approaches – automatic execution based on conversational context versus user-confirmed execution – using a German-language CA in a business-travel scenario. In a between-subjects experiment, 451 employed adult participants (aged 18–65) were randomly assigned to one of the two TC conditions, completed a structured travel request interaction, and then rated their perceived trust (PTru), autonomy (PA), transparency (PTrn), ease of use (PEU), and usefulness (PU). Our findings indicate that TC strategy had a significant overall impact on users' combined perceptions, and that demographic factors – particularly age – played a critical role in shaping these responses, highlighting the need for adaptive, inclusive design practices that balance automation with user control.

CCS Concepts • Human-centered computing → Human computer interaction (HCI) → Interactive systems and tools  
• Human-centered computing → Interaction design → Systems and tools for interaction design.

**Keywords:** conversational agent; function calling; tool calling; large language models

## 1 Introduction

The expanding capabilities of large language models (LLMs), such as OpenAI's GPT series, Meta's LLaMA series, and Mistral AI's Mistral series, offer significant potential for advancing conversational agents (CAs).<sup>1–3</sup> A key development is the ability of LLMs to connect CAs with external services, addressing long-standing challenges in functionality and integration.<sup>4</sup> While earlier, rule-based CAs already supported interactions with external systems, these integrations were typically implemented through rigid, rule-based pipelines in which user utterances were matched to predefined intents and templates. Once an intent was identified, the system deterministically triggered backend actions – such as database queries or external API calls – through handcrafted mappings between intents, entities, and functions. As documented in prior work, for example, classical architectures executed “information retrieval from the Backend through external APIs calls or Database requests” only after successful intent recognition.<sup>5</sup> Such systems were effective in narrow domains but suffered from brittleness, limited language understanding, and high manual engineering effort. For example, prior research has shown that users desire richer integrations, such as controlling home devices or transferring health data.<sup>6,7</sup> Such demands highlight that CAs are no longer judged only by their language processing skills but also by how seamlessly they can execute useful actions across systems – a shift from earlier rule-based approaches, where function execution depended on brittle intent–action mappings, toward more flexible, context-aware integrations enabled by LLMs. Such demands highlight that CAs are no longer judged only by their language processing skills but also by how seamlessly they can execute useful actions across systems.

To support these integrations, emerging system architectures combine LLMs with external knowledge and services. Retrieval-augmented generation (RAG), for instance, allows CAs to access real-time information sources beyond their training data.<sup>8</sup> More recently, LLMs have been trained with tool-calling (TC) capabilities, enabling them not only to

\*Corresponding author: Darius Hennekeuser, University of Applied Sciences Bonn-Rhein-Sieg, Sankt Augustin, Germany, E-mail: [darius.hennekeuser@h-brs.de](mailto:darius.hennekeuser@h-brs.de), <https://orcid.org/0009-0005-8443-1872>

Daryoush Daniel Vaziri, David Golchinfar and Dirk Schreiber, University of Applied Sciences Bonn-Rhein-Sieg, Sankt Augustin, Germany. <https://orcid.org/0000-0003-2025-9637> (D. D. Vaziri). <https://orcid.org/0000-0002-7785-3891> (D. Golchinfar)

Gunnar Stevens, University of Siegen, Siegen, Germany. <https://orcid.org/0000-0002-4490-7187>

retrieve information but also to act – by invoking functions, APIs, or other services in machine-readable formats such as JSON.<sup>12,9</sup> This represents a step change: CAs can now dynamically trigger external processes, broadening their role from conversation partners to versatile digital assistants.

Yet the way TC is implemented introduces new design trade-offs. Functions can be executed automatically when sufficient conversational context is available, streamlining interactions but potentially reducing user control. Alternatively, execution can be delayed until the user explicitly confirms the tool-call, increasing transparency and autonomy but adding extra steps. These two approaches raise fundamental questions about user experience: Do people trust automated executions, or do they prefer confirmation? How do such strategies shape perceptions of autonomy, transparency, usefulness, and ease of use?

Despite the importance of these questions, research has not yet examined how different TC strategies affect user perceptions in CAs. Moreover, demographic differences such as age and gender may interact with these strategies, shaping preferences in ways that current design guidelines do not account for. Addressing this gap is critical for advancing inclusive and adaptive CA design.

**Research Question:** How do different TC approaches (automatic execution vs. user-confirmed execution) affect user perceptions of trust, autonomy, transparency, efficiency and usefulness in interactions with a CA?

In this comparative quantitative study with two groups, we aim to understand how different TC techniques affect user interaction metrics within a chatbot environment. Specifically, we explore two distinct approaches: function execution based solely on conversational context (CA 1) and function execution that requires a user confirmation prompt before proceeding (CA 2). By using the same LLM with two distinct techniques (automatic function execution vs. confirmation-based function execution), we create two experimental groups to evaluate the comparative effects of these TC strategies. The study scenario involves participants completing a business trip application form by interacting with a chatbot. Therefore, the target group was defined as current employees.

Using Multivariate Analysis of Variance (MANOVA) and Permutational Multivariate Analysis of Variance (PERMANOVA), we analyze whether these TC techniques influence a set of dependent variables, such as perceived trust (PTru), perceived autonomy (PA), perceived transparency (PTRn), perceived ease of use (PEU), and perceived usefulness (PU). Additionally, we investigate whether

age and gender, in combination with the CA-setting (automatic execution vs. user-confirmed execution), and as single independent variables, impact the dependent variables. MANOVA/PERMANOVA allows us to assess if the combined impact of the techniques results in significant differences across multiple dependent measures. Following the multivariate analyses, we conduct Analyses of Variance (ANOVAs) to further explore univariate effects, identifying which specific dependent variables contribute to the observed multivariate differences. This approach enables us to determine whether prompting users for confirmation before executing functions significantly enhances the quality of interaction compared to executing functions autonomously based on context alone. The findings from this study provide insights into optimizing TC strategies for enhanced user experience in conversational AI systems.

## 2 Literature review and hypotheses development

### 2.1 Augmenting LLMs with external integrations

The generative pre-training of large language models (LLMs) has demonstrated large gains on varying language generation tasks, such as question answering, machine translation, reading comprehension, summarization and coding.<sup>10–15</sup> However, pure text generative tasks usually do not involve the integration of external services or functions.

As LLMs increasingly serve as the core of autonomous, goal-directed systems, recent research describes this development under the concept of agentic AI. Agentic AI systems exhibit capabilities such as planning, tool use, interpretation of user goals, and adaptive decision-making. Unlike earlier symbolic agents, modern agentic systems rely on LLM-driven orchestration, where agency emerges from generative reasoning and dynamic tool use rather than explicit rule-based planning.<sup>16</sup> In tool-learning frameworks, this agency is operationalized through the model's ability to understand a user's intent, select appropriate tools, and iteratively execute actions based on feedback, enabling LLMs to function as controllers rather than mere text generators.<sup>17</sup>

Retrieval-Augmented Generation (RAG) enhances LLMs by integrating real-time external data, improving responses beyond static training.<sup>8</sup> In education, RAG-powered chatbots pull course materials to tailor student support.<sup>18–20</sup> In customer service, RAG with LLMs accesses a restaurant's knowledge graph to answer specific queries.<sup>21</sup> Overall, RAG + LLM integration boosts accuracy and relevance across

domains, enabling Conversational Agents (CAs) to handle broader queries with contextually rich interactions.<sup>22</sup>

While RAG-applications paired with LLMs provides an external integration of textual knowledge in the LLM, it does not enable the LLM to execute certain tasks based on the conversation. To address the inability to perform specific tasks, Tool-Calling (TC) has emerged as a valuable approach. TC enables LLMs to interact with external systems – APIs, databases, or computational tools – to execute particular functions within the context of a conversation, effectively expanding the model’s capabilities beyond text generation.<sup>23–25</sup> From the perspective of agentic AI research, TC represents a shift from passive text generation to the active use of tools, a capability identified as a core ingredient of modern agent architectures. Recent foundational work highlights that LLM-based agents can function as controllers that reason about user goals, select appropriate tools, and execute multi-step plans to achieve them. This aligns with broader developments in agentic AI, where autonomy increasingly emerges from the orchestration of tool use, feedback integration, and iterative decision-making rather than symbolic planning alone.<sup>16</sup>

Early systems such as Toolformer demonstrate how LLMs can autonomously decide when and how to use tools such as calculators or knowledge databases, which helps them handle tasks requiring precise computations or up-to-date information. Toolformer achieves this by training the LLM to recognize situations where external information or computation enhances response accuracy and then calling the appropriate API accordingly.<sup>26</sup>

Similarly, models like Granite-Function Calling integrate multi-task learning approaches to improve generalizability and handle complex workflows, which involve calling multiple functions in sequence or even creating new tools as needed.<sup>23,27</sup> Such advancements are crucial for enhancing LLMs’ reliability and versatility across domains with high precision demands, such as finance and healthcare.<sup>25</sup>

TC thus represents a foundational shift in how LLMs can handle complex, real-world tasks, moving from static models to dynamic systems capable of interacting with the environment. This advancement enables LLMs to serve as robust agents that not only respond to queries but also execute specialized functions, thus broadening the scope of what conversational AI can achieve.<sup>23,28</sup>

Current model series of OpenAI, like GPT-4o or models of MistralAI support TC to enable users to connect to user defined functions or APIs to build applications designed for individual use cases.<sup>9,29</sup> An example on how TC with such

models works, is shown in the mistral docs and will be described here.<sup>9</sup>

The LLM is provided with functions formatted in JSON schema:

```
[
  {
    "type": "function",
    "function": {
      "name": "retrieve_payment_status",
      "description": "Get payment status of a transaction",
      "parameters": {
        "type": "object",
        "properties": {
          "transaction_id": {
            "type": "string",
            "description": "The transaction id.",
          }
        },
        "required": ["transaction_id"],
      },
    },
  },
  // ... additional function definitions ...
]
```

Functions can then be executed based on the conversational context. E.g., the prompt “What’s the status of my transaction T1001?” would lead to a tool-call, also formatted in JSON-schema:

```
{
  "name": "retrieve_payment_status",
  "arguments": {"transaction_id": "T1001"}
}
```

The tool-call can then be handled as an external function (e.g., an API) and a response from the API, such as {"status": "Paid"} can be sent back to the LLM to continue the conversation with the result of the tool-call.<sup>9</sup>

Altogether, TC represents a foundational mechanism through which LLMs gain the ability to act – not just to speak. As the literature on tool learning shows, connecting LLMs with external tools is central to the emergence of agentic capabilities, enabling CAs to execute complex workflows, handle real-world tasks, and operate as autonomous digital assistants rather than purely conversational systems.<sup>17</sup>

## 2.2 Conceptual background of user perception dimensions

Before turning to the hypotheses, it is useful to briefly outline the key user-perception dimensions that guide our analysis. Prior research on human–AI interaction and technology acceptance highlights five constructs as central for evaluating CAs: perceived trust, perceived autonomy,

perceived transparency, perceived ease of use, and perceived usefulness. Perceived trust (PT<sub>ru</sub>) reflects whether users believe the system acts reliably and without causing harm, a foundational predictor of continued engagement.<sup>30</sup> Perceived autonomy (PA), grounded in Self-Determination Theory (SDT), captures users' sense of volition and control in the interaction, which is especially relevant when CAs proactively execute actions.<sup>31</sup> Transparency concerns users' understanding of how and why a system behaves in certain ways, which becomes increasingly important when AI models invoke external services.<sup>32</sup> Perceived ease of use (PEU) and perceived usefulness (PU) stem from the Technology Acceptance Model (TAM) and represent how effortless and beneficial users expect the system to be.<sup>33</sup> Together, these dimensions form the theoretical basis for assessing how different tool-calling strategies influence user experience. We discuss each construct in detail in the following sections.

While established usability instruments such as the System Usability Scale (SUS)<sup>34</sup> and the User Experience Questionnaire (UEQ)<sup>35</sup> are widely used to assess general usability, parts of their constructs – particularly learnability and ease-of-use – conceptually overlap with TAM's perceived ease of use (PEU). However, TAM focuses specifically on the cognitive evaluations that predict technology acceptance, making PEU and PU more appropriate for studying how users assess tool-calling strategies in conversational agents.<sup>33</sup>

## 2.3 Hypothesis development

To investigate how Tool-Calling (TC) techniques in CAs influence user perceptions, we propose that prompting users for confirmation before executing functions versus autonomous function execution significantly impacts user outcomes. Specifically, the combined dependent variables – perceived trust (PT<sub>ru</sub>), perceived autonomy (PA), perceived transparency (PT<sub>rn</sub>), perceived ease of use (PEU), and perceived usefulness (PU) – are expected to differ between these two techniques (H1). Furthermore, demographic factors, specifically age and gender, are examined as moderators, since existing literature underscores their influence on user perceptions in technology interactions. Age moderates PA, as younger individuals tend to exhibit heightened privacy concerns, affecting their sense of autonomy when interacting with AI-driven systems.<sup>36</sup> Moreover, Generation Z, as digital natives, highly values transparency in interactions, suggesting age differences in responses to transparent or autonomous AI behaviours.<sup>37</sup> Additionally, gender differences have been shown to moderate technology acceptance factors such as PEU and PU. Women, particularly older women, place more emphasis on PEU,<sup>38</sup> whereas

men tend to prioritize PU, focusing on productivity and performance expectancy.<sup>38,39</sup> Consequently, we hypothesize significant moderation effects by gender (H2) and age (H3) on the relationship between TC techniques and the combined dependent variables. Therefore, we derive the following main hypothesis:

**H1:** There will be a significant difference in the combined scores of the dependent variables between CA 1 (autonomous function execution) and CA 2 (user-confirmation prompted execution).

**H2:** Gender will significantly moderate the relationship between TC technique (CA 1 vs. CA 2) and the combined dependent variables.

**H3:** Age will significantly moderate the relationship between TC technique (CA 1 vs. CA 2) and the combined dependent variables.

The subsequent chapters will elaborate further on the proposed relationships between TC techniques and the dependent variables, including detailed exploration through sub-hypotheses.

### 2.3.1 Perceived trust in CAs

The perception of trust among users of conversational agents (CAs) is shaped by multiple factors, including interactivity, humanness, perceived usefulness (PU), and the social and technological contexts in which these agents are used. Studies have shown that higher interactivity – where users can engage in fluid, responsive conversations – boosts perceived trust (PT<sub>ru</sub>), as it makes interactions seem more genuine and aligned with human expectations.<sup>40,41</sup> Humanness, characterized by the chatbot's ability to mimic human-like qualities such as empathy and contextual understanding, further strengthens PT<sub>ru</sub>, as users tend to be more receptive to agents that feel personable and adaptive.<sup>41,42</sup>

Perceived ease of use (PEU) and perceived usefulness (PU) are other critical contributors to PT<sub>ru</sub>. When users find chatbots helpful in achieving their goals efficiently and without technical hurdles, they are more likely to PT<sub>ru</sub> and continue using these systems.<sup>43,44</sup> Initial trust is essential, especially in newer technologies, as it often determines long-term user engagement and loyalty. Furthermore, the perceived risk associated with the use of CAs – particularly concerning privacy and data security – can modulate PT<sub>ru</sub> levels, with higher perceived risks diminishing PT<sub>ru</sub> and, consequently, usage intention.<sup>40,44</sup>

PT<sub>ru</sub> in CAs is intricately linked to user concerns regarding privacy and personal data processing. Several studies have highlighted that users' PT<sub>ru</sub> in these technologies is significantly impacted by their perceptions of how

their personal data is handled, as well as the associated risks of data misuse.<sup>40,41,45,46</sup>

Large language models (LLMs) with tool-calling (TC) capabilities may heighten concerns about privacy and data processing, particularly when users are not informed about which parts of their conversational data trigger these tool-calls. TC allows LLMs to interact dynamically with external systems, enabling them to retrieve, process, or store data based on the user's input. Without explicit user awareness or consent, this capability may feel invasive, as users may be unaware of which specific queries or types of information prompt external processing. Therefore, we formulate H1a, H2a and H3a:

**H1a:** There will be a significant difference in PTru scores between CA 1, which executes functions based solely on conversational context, and CA 2, which executes functions only after receiving a user confirmation prompt.

**H2a:** Gender will significantly moderate the relationship between TC technique and PTru.

**H3a:** Age will significantly moderate the relationship between TC technique and PTru.

### 2.3.2 Perceived autonomy in CA interactions

Perceived autonomy (PA), a fundamental concept in Self-Determination Theory (SDT) developed by Deci and Ryan, is pivotal in understanding user interactions with conversational agents (CAs). SDT emphasizes three essential psychological needs – autonomy, competence, and relatedness – that drive human motivation and well-being. Among these, PA – the sense of volition and alignment of actions with personal values – plays a crucial role in fostering intrinsic motivation and satisfaction.<sup>31</sup>

Studies have found that findings of SDT are also applicable to use cases around CA applications. Specifically, PA in interactions with CAs is shaped by four primary factors. First, sufficient choice and transparency enhance user autonomy by providing multiple options and clear explanations for decisions or recommendations. This transparency is particularly important in real-time applications, like navigation systems, where the immediate consequences of AI decisions directly impact users.<sup>47</sup> Second, personalization of options based on user preferences increases autonomy by fostering alignment with individual values and interests. Third, decision-making assistance, especially in scenarios requiring support, can enhance PA when designed to empower rather than control users. Finally, privacy and data security concerns play a critical role. A lack of clarity or invasiveness in data use threatens users' sense of autonomy, diminishing their willingness to engage with CAs.<sup>47</sup>

In addition to technology adoption, PA is an antecedent to key technology acceptance factors, such as perceived usefulness (PU) and perceived ease of use (PEU). Systems that satisfy autonomy needs tend to foster higher user satisfaction and acceptance, while those that diminish autonomy can have the opposite effect. For instance, in chatbot interactions, users reported greater satisfaction when they felt free to engage with the system based on their preferences. Conversely, high levels of proactive guidance, which felt overly directive, reduced PA and satisfaction. This effect extended to mental well-being, with autonomy satisfaction mediating positive outcomes. However, excessive guidance limited the system's overall ability to improve well-being, suggesting the need for a balance between assistance and user-centered design.<sup>48,49</sup>

In our study setting, PA might differ for users for CA 1 and CA 2, since CA 1 triggers tool-calls (TCs) just based on the conversational context and does not require the user's confirmation of the tool-call execution. This might make users feel less autonomous about the conversation and the data processing. Therefore, we formulate the following H2a, H2b and H2c:

**H1b:** There will be a significant difference in PA scores between CA 1, which executes functions based solely on conversational context, and CA 2, which executes functions only after receiving a user confirmation prompt.

**H2b:** Gender will significantly moderate the relationship between TC technique and PA.

**H3b:** Age will significantly moderate the relationship between TC technique and PA.

### 2.3.3 Perceived transparency in CA interactions

Another factor which is essential in AI applications is perceived transparency (PTRn), as it can significantly enhance users' trust, particularly when paired with features like control and visualization. According to a study by Yu and Li, visualizing the impact of control increased PTRn and trust in an e-learning platform. This visualization allowed users to understand better the algorithm's behaviour, which indirectly explained why recommendations were suitable. It helped users refine their mental model of how the system operated, contributing to higher perceived trust (PTru) and comprehension.<sup>50</sup>

Furthermore, PTRn can also improve user perceptions in social AI applications like chatbots. Research conducted by Xu et al. has found that providing upfront explanations about how a chatbot works reduces feelings of unpredictability and discomfort, increases users' willingness to engage with the AI, and enhances their perception of the

chatbot's social intelligence. This effect is especially strong among users with less prior knowledge of AI, making PTrn a key factor in building trust and fostering positive interactions.<sup>51</sup>

Additionally, PTrn can enhance the effectiveness of AI-driven digital endorsers. Wang and Qiu have found that when AI-based endorsers disclose their artificial nature, consumers report higher trust, engagement, and purchase intentions. This effect is mediated by mind perception, which refers to the extent to which users attribute human-like qualities to AI. PTrn boosts both perceived agency (abilities like reasoning and planning) and perceived experience (empathy and emotional depth), making AI endorsers seem more relatable and credible, and ultimately increasing their persuasive power.<sup>52</sup>

In the context of our study, PTrn may vary depending on how the large language model (LLM) handles tool-calls and user involvement. CA 1, which executes tool-calls automatically without notifying or confirming with users, may appear less transparent, leaving users unsure about what actions were taken or how their data was used. In contrast, CA 2, which asks for explicit user confirmation before executing tool-calls, may enhance PTrn by making system actions visible and controllable.

**H1c:** There will be a significant difference in PTrn scores between CA 1, which executes functions based solely on conversational context, and CA 2, which executes functions only after receiving a user confirmation prompt.

**H2c:** Gender will significantly moderate the relationship between TC technique and PTrn.

**H3c:** Age will significantly moderate the relationship between TC technique and PTrn.

### 2.3.4 Perceived ease of use in CA interactions

The increasing complexity of digital systems and software solutions has made perceived ease of use (PEU) a critical determinant in the adoption and sustained utilization of technology. Despite technological advancements that continue to extend the functional capabilities of systems, user acceptance remains contingent not only on the usefulness of these systems but also on how easily they can be used.

A foundational definition of perceived PEU was established by Davis, who defined it as “the degree to which a person believes that using a particular system would be free of effort.” This definition, emerging from the context of the Technology Acceptance Model (TAM), highlights that effort is a finite resource users allocate among multiple tasks.<sup>53</sup> Thus, a system that minimizes the effort required

for operation is inherently more attractive and more likely to be adopted.

Indeed, as highlighted in recent research on conversational agents (CAs), systems that are easy to use – offering intuitive interaction, user-friendly interfaces, and low cognitive demand – significantly increase the likelihood of acceptance and sustained usage. The studies reviewed show that the lower the effort required to engage with the system, the higher the probability that users will adopt it. Conversely, if users perceive a system as too complicated or cumbersome, this can act as a major barrier to adoption, regardless of the system's usefulness.<sup>54,55</sup>

The significance of PEU extends beyond mere convenience. Systems that are difficult to navigate or operate can impede users' ability to achieve intended outcomes, thereby undermining both personal productivity and organizational efficiency. As Davis (1989) posited, even when users recognize a system's potential usefulness, they may reject it if the effort required to utilize it outweighs the anticipated benefits. Hence, PEU plays a pivotal role in determining whether users are willing to engage with a system, even when its functional benefits are evident.<sup>53</sup>

In the context of our study, PEU is likely to differ between CA 1 and CA 2, as they follow different approaches to function execution. CA 1, which triggers tool-calls automatically based on conversational context, may reduce the perceived effort required from users because it handles tasks without requiring explicit confirmation. However, this could also come at the cost of reduced (perceived autonomy) PA or perceived transparency (PTrn). On the other hand, CA 2, which requires explicit user confirmation before executing tool-calls, may increase user effort due to additional steps in the interaction but may enhance users' sense of control and awareness.

Thus, investigating how users perceive the PEU between these two LLM behaviours is essential, as it can influence both short-term interaction satisfaction and long-term technology adoption. Understanding this trade-off between automation (reduced effort) and control (potentially increased effort) will help design more effective and user-friendly AI-driven conversational systems.

**H1d:** There will be a significant difference in perceived PEU scores between CA 1, which executes functions automatically without user confirmation, and CA 2, which requires explicit user confirmation before executing functions.

**H2d:** Gender will significantly moderate the relationship between TC technique and perceived PEU.

**H3d:** Age will significantly moderate the relationship between TC technique and perceived PEU.

### 2.3.5 Perceived usefulness in CA interactions

In the realm of information systems research, perceived usefulness (PU) has long been recognized as a pivotal determinant of whether individuals and organizations choose to adopt new technologies. As originally framed by Davis, PU is defined as “the degree to which a person believes that using a particular system would enhance his or her job performance”.<sup>53</sup> This definition underscores a core principle: users are more inclined to embrace technologies when they expect tangible gains – such as increased efficiency, effectiveness, or overall productivity – in their tasks.

The influence of PU is evident across various contexts, from productivity tools in corporate environments to consumer-facing applications. A system might possess cutting-edge features, or robust technical capabilities yet remain underutilized if end users do not perceive a clear link between the system’s functionality and their personal or professional objectives. This practical, outcome-focused view places PU at the heart of the TAM,<sup>53</sup> where it consistently emerges as a stronger predictor of user intentions and behaviour than many other factors. This practical, outcome-oriented perspective positions PU at the core of the TAM,<sup>53</sup> where it consistently proves to be a significant predictor of user intentions and behavior.

Recent empirical work on CAs further underscores the centrality of PU in shaping user adoption. For instance, a study of older adults’ interactions with chatbots during the COVID-19 pandemic revealed that PU had a direct positive effect on usage intentions, outstripping demographic factors such as age and gender in predictive power.<sup>54</sup> In a similar vein, an integrative review of health care chatbots identified performance expectancy – closely mirroring the concept of PU – as the most influential factor determining whether both patients and professionals choose to adopt a given system.<sup>55</sup> These findings collectively reinforce the notion that when users see tangible, outcome-focused benefits – whether in terms of efficiency gains or improved health outcomes – they are far more inclined to integrate new technologies into their routines.

In the context of CA 1, which automatically executes tool-calls based on context, users may experience greater efficiency because the system handles tasks proactively, potentially increasing PU. By contrast, CA 2, which requires explicit user confirmation, may make task completion feel more deliberate but also more effortful – even if it offers greater control and clarity. How users weigh automation against the need for manual oversight will likely affect whether they perceive the system as “useful” for achieving meaningful outcomes.

Therefore, examining whether the automated approach of CA 1 or the user-confirmation approach of CA 2 leads to stronger PU, can offer vital insights for designing conversational experiences that balance efficiency, control, and ultimate task performance.

**H1e:** There will be a significant difference in PU scores between CA 1, which automates function execution, and CA 2, which requires explicit user confirmation before executing functions.

**H2e:** Gender will significantly moderate the relationship between TC technique and PU.

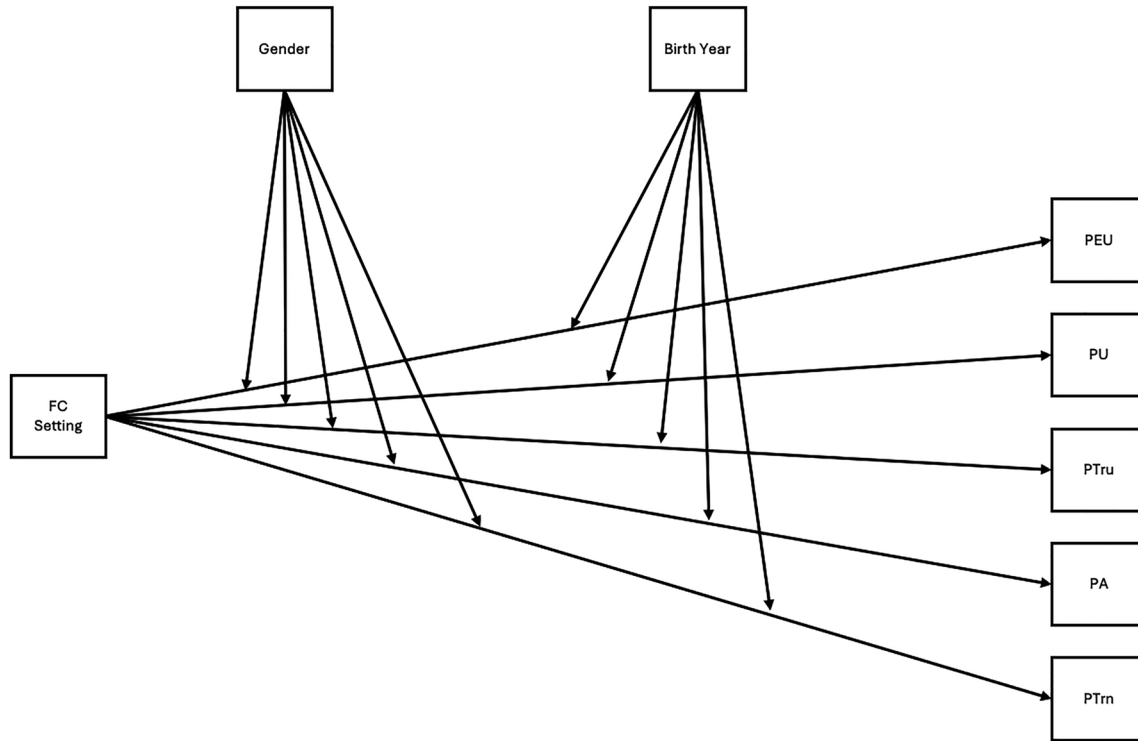
**H3e:** Age will significantly moderate the relationship between TC technique and PU.

### 2.3.6 Direct effects of demographics

While our conceptual model emphasizes Age and Gender as moderators of the effect of chatbot setting on user perceptions, the statistical analysis also includes direct effects of these demographic variables on the dependent variables. Including Age and Gender as predictors ensures proper model specification when testing interaction terms (e.g., Group  $\times$  Age, Group  $\times$  Gender). This means that beyond potential moderation, we also test whether Age and Gender independently explain variance in perceived autonomy, trust, transparency, ease of use, and usefulness. Although these direct effects are not part of the core theoretical hypotheses, they are reported in order to provide a complete and transparent account of the statistical results.

### 2.3.7 Conceptual model

To bring together our proposed relationships, we have developed a comprehensive conceptual model (see Figure 1). At its core, this model posits that the mode of tool-calling (TC) – automatic execution versus user-confirmed execution – directly impacts five key user perceptions: perceived trust (PT<sub>ru</sub>), perceived autonomy (PA), perceived transparency (PT<sub>rn</sub>), perceived ease of use (PEU), and perceived usefulness (PU) (H1). In addition, we theorize that two demographic factors, age and gender, will moderate these effects: age shaping the strength and direction of each functional link (H3), and gender influencing how users weigh control versus convenience (H2). By mapping both the direct paths from TC technique to each outcome and the interaction paths via age and gender, this model provides a unified framework for testing our hypotheses in a single, integrated analysis.



**Figure 1:** Hypothesized model: function-calling technique's effects on user perceptions, with age and gender as moderators.

## 3 Methodology

### 3.1 Research design

This chapter details the methodological approach used to compare two distinct tool-calling (TC) strategies within a conversational agent (CA). The study was conducted in two separate phases, each featuring one of the two approaches. The following sections describe the study design, participant recruitment, materials used, procedure, and data analysis.

#### 3.1.1 Study design overview

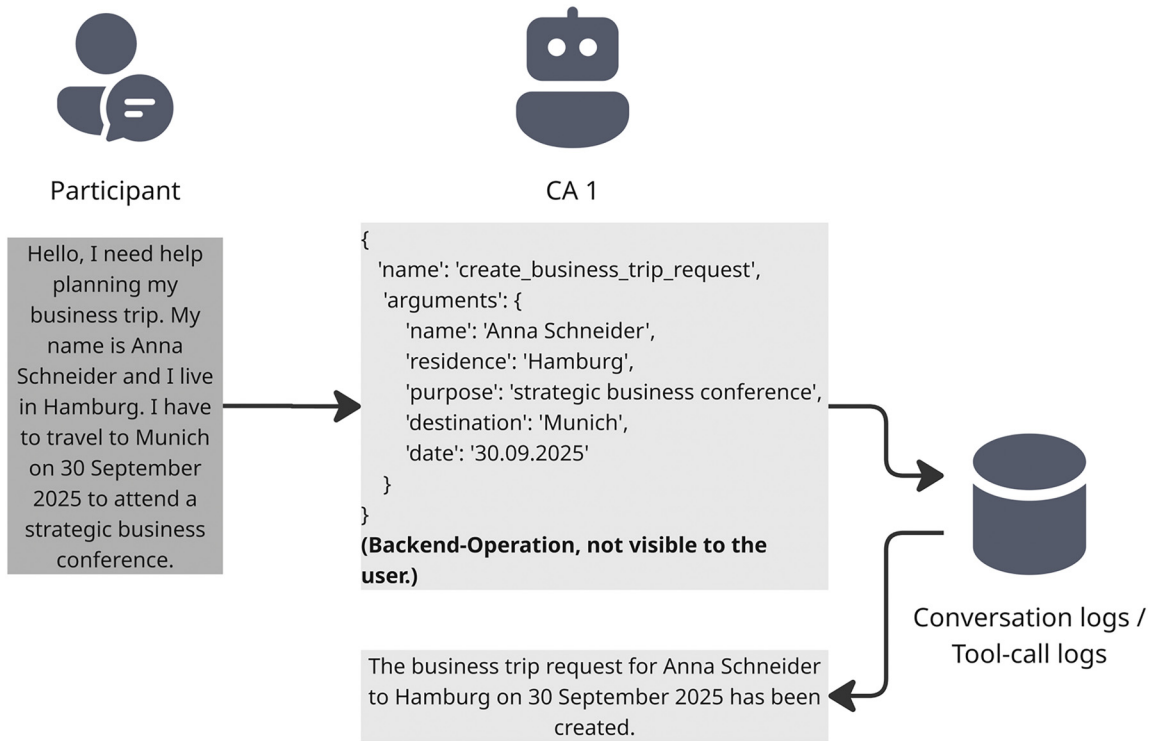
The study employed a between-subjects experimental design to investigate the effects of two distinct tool-calling (TC) methods implemented in a German-language conversational agent (CA). The experiment consisted of two phases, each utilizing a different TC strategy.

The study was situated within a business-trip planning scenario, in which participants interacted with the CA to complete a standardized travel request. We chose this scenario because it offers a clear, structured workflow that involves providing information, checking inferred details, and triggering external functions – steps that are essential for comparing different tool-calling strategies in a

controlled way. Importantly, the task might be familiar to most employed participants, which helps reduce domain-specific learning effects and allows observed differences to more confidently be attributed to the tool-calling behavior rather than to the task itself.

**Phase 1:** In the first phase, the CA (powered by CA 1) automatically executed external functions as soon as it inferred sufficient context from the participant's input. The process was entirely automated, and no confirmation from the participant was required before executing the function. Figure 2 illustrates this workflow, where the CA directly processes the participant's request for a business trip without seeking further approval.

**Phase 2:** In the second phase, the CA (powered by CA 2) followed a confirmation-based approach. After gathering sufficient context from the conversation, it presented the inferred details to the participant for confirmation before proceeding with the tool-call. The confirmation gateway parsed the LLM output for the tool-call and automatically prompted the large language model (LLM) to ask for confirmation. Only after receiving explicit approval via the tool-call *confirm\_business\_trip\_request* did the CA execute the function. Figure 3 depicts this process, highlighting the confirmation step as a critical difference compared to Phase 1.



**Figure 2:** Exemplary conversation flow with CA 1.

Both phases employed the same scenario and instructions to ensure that any observed differences in user experience or performance could be attributed solely to the TC strategy rather than variations in task content or context.

Participants were randomly invited to one of the two phases, meaning each participant interacted with only one version of the conversational agent (CA). After completing the assigned scenario, participants were directed to a questionnaire measuring the dependent variables: perceived trust (PTru), perceived autonomy (PA), perceived transparency (PTrn), PEU, and perceived usefulness (PU).

### 3.1.2 Participants and recruitment

Participants were recruited via Bilendi, an online panel provider established in 1999 that supplies research samples for both academic and commercial studies. Bilendi operates research panels across several European countries and maintains a database of more than two million registered members. Panelists are profiled along demographic and behavioural dimensions, allowing researchers to apply detailed inclusion criteria when drawing samples. Recruitment procedures follow common industry standards such as double opt-in registration and compliance with ESOMAR guidelines, helping ensure data quality and panel

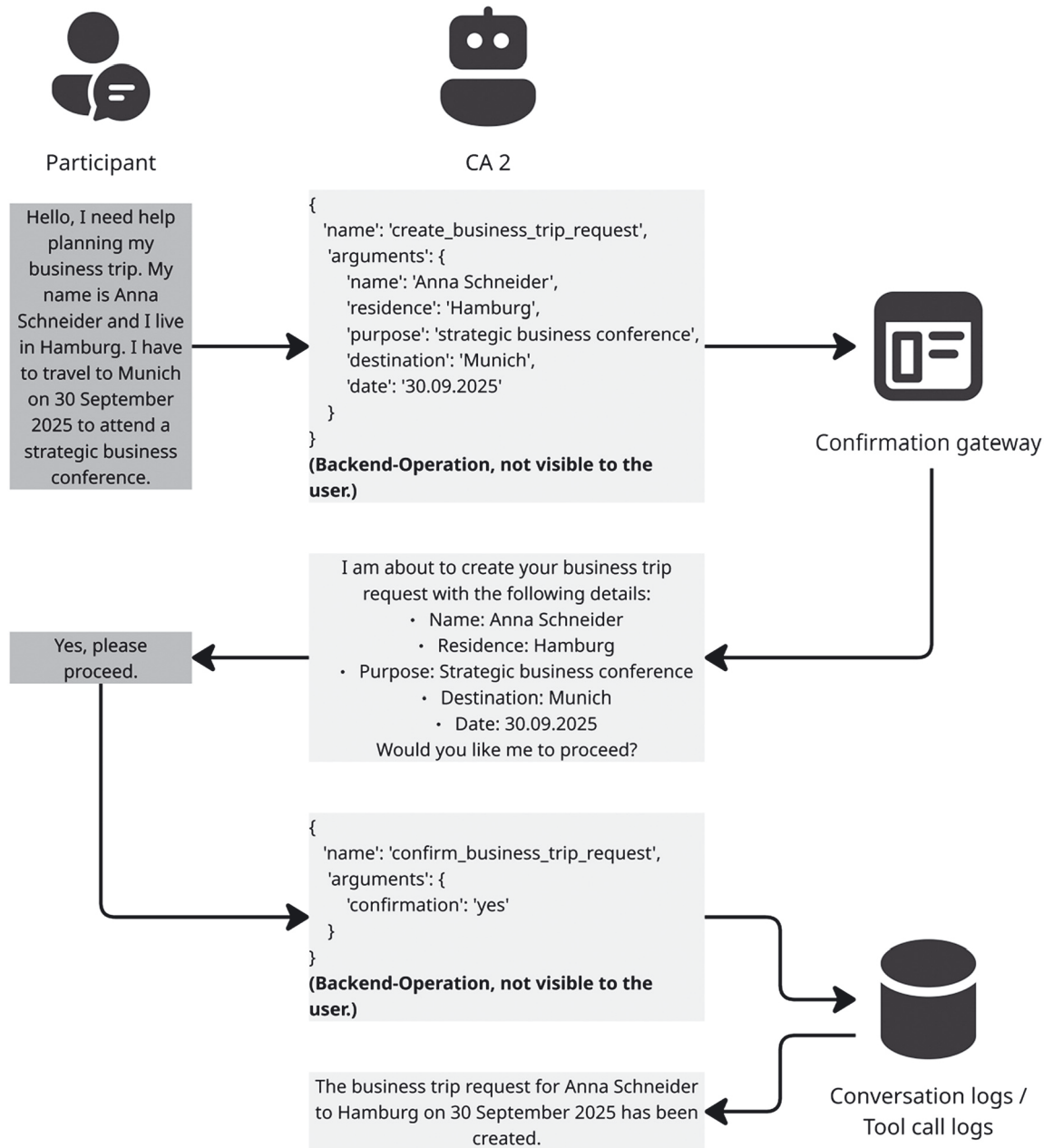
reliability.<sup>56</sup> Individuals invited to participate were required to meet three primary criteria:

1. They had to be adults (18 years or older).
2. They needed to reside in Germany and be fluent in German, as both the conversational agent (CA) interface and questionnaire were presented exclusively in German.
3. They had to confirm that they were currently employed, either full-time or part-time, to match the work-related scenario.

Upon completing the study, participants received an incentive of 2 Euros, deemed suitable compensation for the time and effort invested. A roughly equal number of participants was allocated to each phase of the experiment (CA 1 vs. CA 2), aiming to secure balanced sample sizes and enhance the comparability of results.

### 3.1.3 Instruments

We measured perceived trust (PTru), perceived autonomy (PA), perceived transparency (PTrn), perceived ease of use (PEU), and perceived usefulness (PU) using validated measurement scales from prior research. Specifically, PTru was measured using items adapted from Corritore et al. (2005), demonstrating strong reliability in their study (Cronbach's



**Figure 3:** Exemplary conversation flow with CA 2.

$\alpha = 0.84$ ).<sup>57</sup> PA items were adapted from Nguyen et al. (2022), based on SDT, also exhibiting high reliability (Cronbach's  $\alpha = 0.93$ ).<sup>31,58,59</sup> The items measuring PTRn were adapted from Wang and Benbasat (2016), who studied its impact on user trust in recommendation agents.<sup>32</sup> Lastly, PEU and PU were measured using the original items from Davis' TAM.<sup>53</sup> All measurement items and their validation results are presented in the results section. The measurement items were first translated using DeepL and then independently reviewed by two native German speakers, who made

any necessary adaptations. These revised translations were subsequently compared and discussed until a final consensus version was agreed upon.

### 3.1.4 Procedure

Data collection was carried out in two phases – one for each tool-calling (TC) condition – although the general procedure remained consistent across both phases.

Step 1: Recruitment.

Eligible participants were invited via the Bilendi panel.

Step 2: Screening.

Participants confirmed that they were currently employed and within the required age range. Those who did not meet these requirements were not allowed to continue in the study. They also consented to anonymous data processing.

Step 3: Interaction with the Chatbot.

Here, participants were directed to the chatbot interface. They were briefed on the scenario of creating a business travel request for a trip from Hamburg to Munich on September 30, 2025, in the role of Anna Schneider.

- In Phase 1 (CA 1), the CA automatically invoked external functions upon recognizing the necessary travel information.
- In Phase 2 (CA 2), the CA requested explicit user confirmation before calling each external function.

Throughout the conversation, participants could pose questions or provide details to the chatbot until they successfully received a confirmation message indicating that the travel request had been created.

Step 4: Return to Questionnaire and Completion.

Once participants were satisfied that the chatbot had processed their request, they navigated back to the online questionnaire. They were then asked to confirm they had finished the task successfully, provide demographic information (age, gender), and rate their experiences based on multiple-item scales covering perceived trust (PTru), perceived autonomy (PA), perceived transparency (PTrn), perceived ease of use (PEU), and perceived usefulness (PU).

Step 5: Debriefing and Incentive.

Upon submitting the final responses, participants were shown a thank you page, which reiterated the confidentiality of the data and indicated that their incentive of 2 Euros would be processed through the Bilendi platform. Figure 4 illustrates the process of a participant's involvement.

### 3.1.5 Data analysis

A rank-transformed MANOVA served as our primary statistical technique to investigate the impact of tool-calling (TC) strategy on the five dependent variables. In this approach, each outcome was averaged (PA\_avg, PTru\_avg, PTrn\_avg, EOU\_avg, PU\_avg) and replaced by its overall sample rank to mitigate violations of multivariate normality and homogeneity of covariance. We then fit the factorial model using Pillai's Trace to assess whether the combined ranks of PA\_avg, PTru\_avg, PTrn\_avg, EOU\_avg and PU\_avg differed between automatic and user-confirmed execution groups.

In addition, a PERMANOVA was conducted as a complementary nonparametric approach. PERMANOVA is based on distance matrices and permutation testing and does not require multivariate normality or homogeneity of covariance matrices. Including PERMANOVA allows for validation of results under more flexible assumptions, ensuring robustness of findings.

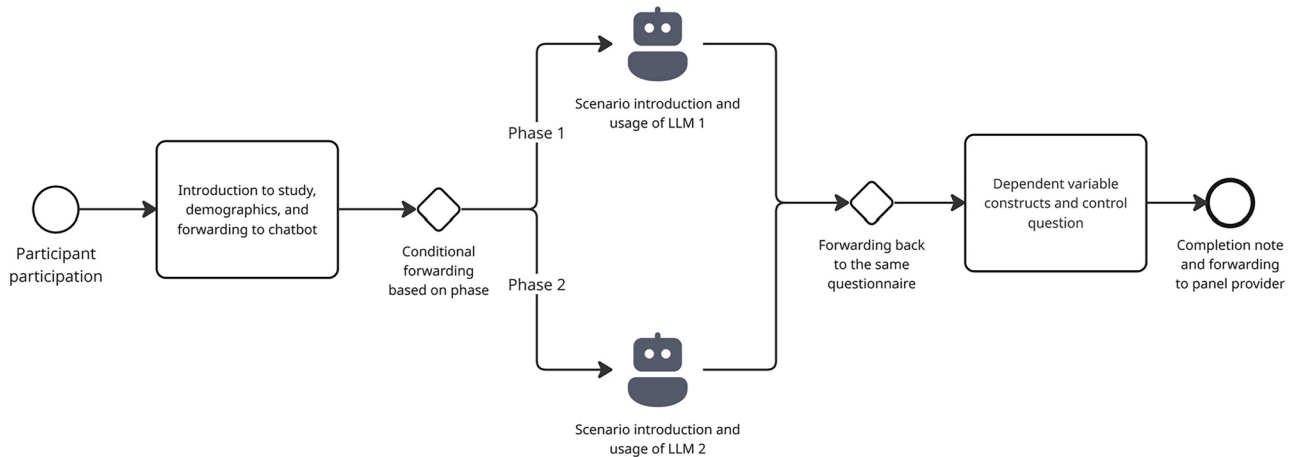
Prior to conducting the MANOVA and PERMANOVA, the dataset was screened for missing values, outliers, and potential violations of normality, linearity, and homoscedasticity. Since a statistically significant difference emerged from the MANOVA, follow-up univariate ANOVAs were conducted to determine which specific dependent variables were notably influenced by the TC strategy.

## 3.2 LLM selection, training procedure and prompt design

Deciding upon which LLM to use for our study, we used the Berkeley Function-Calling Leaderboard, which is a benchmark developed by researchers at UC Berkeley to evaluate the performance of large language models (LLMs) specifically on their ability to correctly invoke and execute functions based on natural language instructions.<sup>60</sup> Due to limited computing resources, we were bound to use a model of a size of maximum 8 billion parameters. Additionally, as the participants of the study were German speaking, we needed a model with German speaking capabilities. Therefore, we only considered models which were providing information about German speaking capabilities. Based on these constraints the Qwen2.5-7b-Instruct offered a promising overall accuracy of 56.7 on the leaderboard and multilingual capabilities.<sup>61</sup>

However, the Qwen2.5-7b-Instruct demonstrated weaknesses in specific benchmark categories critical to our research goals, notably in Irrelevance Detection and Relevance Detection, achieving accuracy scores of only 77.78 % and 69.08 %, respectively. Irrelevance Detection refers to scenarios where none of the provided functions are relevant to the user's query, and the model is expected not to invoke any function. Conversely, Relevance Detection involves scenarios where at least one provided function is relevant and should be invoked, though precise correctness of the invocation parameters is not assessed.<sup>60</sup>

To maintain a strong overall performance while minimizing hallucinations, we decided to further fine-tune the model specifically targeting improvements in these two categories. Our dataset consisted of various self-created German tool-calling (TC) examples and the apigen-synthtrl dataset by argilla-warehouse and the ToolACE dataset from Team-ACE.<sup>62,63</sup> Additionally, we used data specifically



**Figure 4:** Study workflow: Participant engagement, LLM interaction, and data collection flow.

designed to make the model learn when to ignore and when to invoke provided functions, by also adding samples where the conversation was not relevant to the provided functions. The full dataset consisted of 92,496 samples.

The training process consisted of one training epoch targeting only 50 % of the LLM’s layers with the highest signal-to-noise ratio, calculated using Random Matric Theory, in order to reduce GPU memory use, accelerate training, while maintaining the performance of full fine-tuning methods.<sup>64</sup> As a result, the model achieved an overall accuracy of 55.04 %, with notable improvements in specific categories – 83.33 % for relevance detection and 80.37 % for irrelevance detection.

### 3.3 Technical infrastructure

In this study, the conversational agent (CA) environment operates through three core services, each running as a separate Docker container:

1. Database Service (PostgreSQL<sup>65</sup>).

The first service manages all data storage, including user-session information, conversation transcripts, tool-call logs, and any pending trip requests. By preserving this data in a dedicated database container, the system can reliably track both user activities and model-generated outputs. Storing these records is essential for auditing of system actions (e.g., invoked tool calls). PostgreSQL was chosen as a reliable, industry-standard relational database that ensures stable storage of session and tool-call data.

2. User Interface (Streamlit<sup>66</sup>).

The second service provides the main user interface and business logic. Implemented in Streamlit, it hosts a web-based interface enabling participants to interact with the CA. Upon receiving user inputs, the application consults the database to retrieve session context and then dispatches requests to the large language model (LLM) service. Once the LLM responds, the application processes any returned outputs (such as tool-call requests) and updates the database accordingly. This layer thus orchestrates the end-to-end workflow: authenticating users, storing messages, and ensuring that each step of the conversation is recorded. A screenshot of the user interface can be found in the Appendix A.1. Streamlit was used because it enables fast prototyping of simple web interfaces, which is sufficient for a controlled research demo.

3. LLM Service (vLLM<sup>67</sup>).

The third service handles inference by running a pretrained LLM in a container configured to accept chat completion requests. When the application forwards user messages, the LLM produces appropriate responses and may request the execution of external functions. All computations related to language processing occur in this service, which can be scaled or replaced independently of the database or application layers. This separation ensures that updates to the model architecture or parameters do not disrupt the rest of the system. vLLM was selected because it provides efficient LLM inference, an OpenAI-compatible API, and native support for tool-calling required by our system.

## 4 Results

### 4.1 Data screening, preparation, and participant demographics

A total of 1,063 individuals accessed the online questionnaire. Of these, 505 participants ultimately completed the survey – 251 in Group A and 254 in Group B (Group A interacted with CA 1, Group B interacted with CA 2). Another 242 participants began the survey but did not finish (135 in Group A; 107 in Group B). Additionally, 171 respondents were removed due to exceeding predefined demographic quotas (79 in Group A; 92 in Group B). A further 145 participants were excluded after failing the control question (72 in Group A; 73 in Group B).

After excluding non-completes and ineligible respondents, 487 cases remained, of which 25 participants were removed due to outlier management (see 4.2). The remaining 462 participant's answers were used for further descriptive analysis.

Descriptive statistics for age indicated a mean of 45.94 (SD = 12.64) for the entire sample, with the youngest participant being 19 years old and the oldest 78 years old. Broken down by experimental group, Group A ( $n = 233$ ) had a mean age of 45.91 (SD = 12.45), whereas Group B ( $n = 229$ ) had a mean age of 45.97 (SD = 12.86).

Regarding gender, the final sample comprised 234 females, 228 males. In Group A, 119 participants were female and 114 were male. In Group B, 115 were female and 114 were male.

All variables were transformed using a custom scaling process based on known scale ranges. Specifically, variables from the following constructs were scaled according to their maximum values (min-max scaling): perceived autonomy (PA) (scale from 1 to 5), perceived trust (PTru) (scale from 1 to 7), perceived transparency (PTrn) (scale from 1 to 7), ease of use (EOU) (scale from 1 to 7), and perceived usefulness (PU) (scale from 1 to 7). Subsequently, the Likert items for each construct were averaged to generate the following mean scores: PA\_avg, PTru\_avg, PTrn\_avg, EOU\_avg, and PU\_avg.

### 4.2 Outlier management and testing statistical assumptions

Since the model investigated gender as a moderator, it was crucial to ensure a statistically significant and robust data base. Due to the low number of respondents identifying as diverse or choosing not to specify their gender (three participants were diverse, one participant did not specify), these cases were excluded from the analysis to maintain

statistical validity. In addition, one participant was excluded because they were under 18 years of age, in line with the study's eligibility criteria.

Outliers were initially identified and removed using two methods: the interquartile range (IQR) method and the Mahalanobis distance method, both applied separately within each experimental group. The IQR method identified univariate outliers based on the lower and upper bounds ( $Q1 - 1.5 \times IQR$ ,  $Q3 + 1.5 \times IQR$ ) calculated within each group, reducing the sample by 16 cases.

Subsequently, multivariate outliers were assessed using Mahalanobis distance. The threshold for identifying outliers was determined using the chi-squared distribution, with degrees of freedom equal to the number of averaged constructs ( $df = 5$ ) and a significance level of 0.001. This cutoff is recommended by Tabachnick and Fidell (1996) as a standard criterion for detecting extreme multivariate outliers in psychological research.<sup>68</sup> This resulted in a cutoff value of 20.52, meaning that any case with a Mahalanobis distance exceeding this value was considered an outlier. Based on this criterion, 4 cases were flagged, reducing the dataset from 466 to 462 participants.

To test the assumptions for the MANOVA, we examined absence of multicollinearity and singularity, linearity, multivariate normality and homogeneity of covariance matrices according to Warne.<sup>69</sup> All test results are described in the following sections.

Multicollinearity was assessed separately within each group using Pearson correlation matrices of the dependent variables. In Group A, pairwise correlations ranged from  $r = 0.46$  to  $0.67$ , while in Group B they ranged from  $r = 0.46$  to  $0.62$ . These moderate correlations indicate that the dependent variables are sufficiently related to justify MANOVA, without being so highly intercorrelated as to suggest redundancy. Tabachnick & Fidell (2012) recommend that inter-construct correlations should remain below  $r = 0.90$  to avoid redundancy, which our findings (Group A:  $r = 0.46$ – $0.67$ ; Group B:  $r = 0.46$ – $0.62$ ) comfortably meet.<sup>70</sup> In addition to supporting the absence of multicollinearity, these correlations also demonstrate the approximately linear relationships among dependent variables required for MANOVA. Furthermore, because Age was included as a covariate, we tested linearity between Age and each dependent variable by comparing linear and quadratic models. In all cases, the quadratic term was non-significant, confirming that the covariate relationships could be assumed to be linear.

Multivariate normality was assessed using Mardia's test across the five averaged constructs. The results indicated a significant departure from normality ( $HZ = 4.47$ ,

$p < 0.001$ ), suggesting that the assumption of multivariate normality was not met. Consistently, Shapiro–Wilk tests for univariate normality also showed significant deviations from normality for each dependent variable: PA ( $W = 0.951$ ,  $p < 0.001$ ), PTru ( $W = 0.968$ ,  $p < 0.001$ ), PTrn ( $W = 0.977$ ,  $p < 0.001$ ), EOU ( $W = 0.969$ ,  $p < 0.001$ ), and PU ( $W = 0.954$ ,  $p < 0.001$ ). Together, these results demonstrate violations of the assumption of normality, both univariate and multivariate.<sup>71</sup>

Lastly, the assumption of homogeneity of covariance matrices was tested using Box’s M test.<sup>72</sup> The result was statistically significant,  $\chi^2(15) = 31.54$ ,  $p = 0.007$ , indicating that the covariance matrices across groups were not equal at the conventional  $\alpha = 0.05$  level. However, it is important to note that Box’s M test is highly sensitive to sample size and deviations from multivariate normality and may become significant even with minor violations when large samples are involved. In light of this, some authors recommend using more conservative significance thresholds, such as  $\alpha = 0.001$ .<sup>73,74</sup> Applying such criteria would suggest a more cautious interpretation of this result. Despite the significant test statistic, the extent of the violation may not be practically meaningful.

We attempted various transformation techniques, including logarithmic, square root, and inverse, to address the lack of normality and homoskedasticity. However, none of these transformations succeeded in normalizing the data or achieving homoskedasticity. Therefore, to preserve data integrity and interpretability, we proceeded with the analysis without transforming the data.

According to Warne,<sup>69</sup> when the assumptions of multivariate normality and homoscedasticity (equality of covariance matrices) are not met, it is recommended to proceed with MANOVA using the Pillai’s Trace statistic. Pillai’s Trace is considered the most robust MANOVA test statistic under assumption violations, as it is less sensitive to deviations from normality and unequal covariance matrices compared to other statistics like Wilks’ Lambda, Hotelling’s Trace, and Roy’s Largest Root. Therefore, despite the lack of normality and homoscedasticity in our data, we conducted the MANOVA analysis using Pillai’s Trace as the primary test statistic.

Additionally, PERMANOVA was conducted. Unlike MANOVA, PERMANOVA does not assume multivariate normality and is robust to deviations from covariance homogeneity, particularly in balanced designs. It relies on permutation testing and geometric partitioning of a distance matrix, making it well-suited for analysing multivariate differences under relaxed distributional assumptions. Therefore, PERMANOVA was employed

as a complementary method to validate the robustness of the findings derived from the MANOVA, using a distribution-free framework.<sup>75</sup>

### 4.3 Measurement reliability

To assess the internal consistency of each multi-item scale, Cronbach’s alpha values were computed. All five scales demonstrated reliability above 0.70. In particular, perceived autonomy (PA), perceived transparency (PTrn), ease of use (EOU), and perceived usefulness (PU) all achieved alpha values above 0.90. Perceived trust (PTru) showed an alpha of 0.83. Table 1 summarizes the item groupings and Cronbach’s alpha values for each scale.

### 4.4 Results of MANOVA and PERMANOVA

A MANOVA was conducted to examine the effects of tool-calling (TC) strategy (statistically treated as a group variable in the analysis), gender, birth year, and their interactions on the combined dependent variables: PA\_avg, PTru\_avg, PTrn\_avg, EOU\_avg, and PU\_avg. As shown in Table 2, significant multivariate effects were observed for Group (Pillai’s Trace = 0.024,  $F(5, 452) = 2.26$ ,  $p = 0.047$ ), Gender (Pillai’s Trace = 0.035,  $F(5, 452) = 3.23$ ,  $p = 0.007$ ), and Birth year (Pillai’s Trace = 0.030,  $F(5, 452) = 2.84$ ,  $p = 0.016$ ). Neither the Group  $\times$  Gender interaction (Pillai’s Trace = 0.017,  $p = 0.158$ ) nor the Group  $\times$  Birth year interaction (Pillai’s Trace = 0.017,  $p = 0.159$ ) reached statistical significance.

To complement the MANOVA and account for potential violations of multivariate normality and covariance homogeneity, a PERMANOVA was conducted using Euclidean distance and 9,999 permutations. The model included group, gender, birth year, and their interactions.

The overall model was statistically significant,  $F(5, 456) = 3.33$ ,  $p = 0.001$ , indicating that the multivariate distribution of the dependent variables differed significantly across the predictor combinations. The model accounted for approximately 3.52 % of the total variance in the data ( $R^2 = 0.035$ ), with the remaining 96.48 % attributed to residual (unexplained) variance (see Table 3).

The results of both MANOVA and PERMANOVA consistently indicate that TC strategy, gender, and birth year are associated with significant differences in the combined outcome variables of PA\_avg, PTru\_avg, PTrn\_avg, PEU\_avg, and PU\_avg. The significant interaction between group and birth year suggests that the effect of TC strategy may vary depending on participants’ age. While MANOVA assumptions were partially violated, the converging evidence from the distribution-free PERMANOVA supports the robustness of these findings.

**Table 1:** Scale reliability and Cronbach’s alpha values.

Scale	Items	Cronbach’s alpha
PA	PA1: I felt like I could decide how to complete the tasks.	0.927
	PA2: I felt like I could pretty much be myself when completing the tasks.	
	PA3: When doing the tasks, I had an opportunity to decide for myself how to go about my work.	
	PA4: I felt like I had flexibility to decide how to complete the tasks.	
	PA5: I could control how the tasks were done.	
	PA6: I felt like I could manage my own work while completing the tasks.	
PTru	PTru1: I expect this chatbot will not take advantage of me.	0.805
	PTru2: I believe this chatbot is trustworthy.	
	PTru3: I believe this chatbot will not act in a way that harms me.	
	PTru4: I trust this chatbot.	
PTrn	PTrn1: The chatbot makes its reasoning process clear to me.	0.939
	PTrn2: It is apparent to me how the algorithm of the chatbot handles the data of incoming inquiries.	
	PTrn3: It is apparent to me how the algorithm of the chatbot generates the answers.	
	PTrn4: I do not understand how the chatbot performs its job.	
	PTrn5: I easily understand the chatbot’s reasoning process.	
	PTrn6: It is easy for me to understand the inner workings of the chatbot.	
EOU	EOU1: Learning to operate the chatbot would be easy for me.	0.923
	EOU2: I would find it easy to get the chatbot to do what I want it to do.	
	EOU3: My interaction with the chatbot would be clear and understandable.	
	EOU4: I would find the chatbot to be flexible to interact with.	
	EOU5: It would be easy for me to become skillful at using the chatbot.	
	EOU6: I would find the chatbot easy to use.	
PU	PU1: Using the chatbot in my job would enable me to accomplish tasks more quickly.	0.979
	PU2: Using the chatbot would improve my job performance.	
	PU3: Using the chatbot in my job would increase my productivity.	
	PU4: Using the chatbot would enhance my effectiveness on the job.	
	PU5: Using the chatbot would make it easier to do my job.	
	PU6: I would find the chatbot useful in my job.	

**Table 2:** MANOVA multivariate tests (Pillai’s Trace).

Effect	Pillai’s Trace	F	Df	p	Sig.
Intercept	0.231	27.16	5,452	<0.001	***
Group	0.024	2.26	5,452	0.047	*
Gender	0.035	3.23	5,452	0.007	**
Group × gender	0.017	1.60	5,452	0.158	n.s.
Birth year	0.030	2.84	5,452	0.016	*
Group × birth year	0.017	1.60	5,452	0.159	n.s.

Significance codes – \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ , n.s., not significant.

**Table 3:** PERMANOVA results.

Source	Df	Sum of sqs	R <sup>2</sup>	F	p	Sig.
Model	5	2.940	0.0352	3.33	0.001	***
Residual	456	80.504	0.9648			
Total	461	83.444	1.0000			

Significance codes – \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ , n.s., not significant.

### 4.5 Univariate effects: follow-up ANOVAs

Before conducting univariate ANOVAs on each of the five dependent variables (*PA\_avg*, *PTru\_avg*, *PTrn\_avg*, *EOU\_avg*, and *PU\_avg*), assumptions of normality and homogeneity of variances were tested in accordance with the recommendations by Ståhle and Wold (1989).<sup>76</sup> Residual normality was assessed using the Shapiro–Wilk test,<sup>77</sup> and Levene’s test<sup>78</sup> was used to examine the homogeneity of variances across groups.

The Shapiro–Wilk test indicated statistically significant deviations from normality in the residuals for all five dependent variables ( $p < 0.001$ ). However, Levene’s test showed no significant violations of homoscedasticity for any variable ( $p > 0.05$ ), suggesting that the assumption of equal variances was met.

Even if normality was violated, ANOVA is considered robust to such violations – particularly in studies with large sample sizes and approximately equal group sizes. This is supported by extensive empirical research,<sup>72,79–82</sup> which demonstrates that the Type I error rate and statistical power of ANOVA remain largely unaffected under such conditions.

**Table 4:** Summary of ANOVA results for each dependent variable.

Effect	PA_avg (p_adj)	PTru_avg (p_adj)	PTrn_avg (p_adj)	EOU_avg (p_adj)	PU_avg (p_adj)
Group	n.s. (1.000)	n.s. (1.000)	n.s. (0.213)	n.s. (1.000)	n.s. (1.000)
Gender	n.s. (1.000)	n.s. (1.000)	n.s. (1.000)	n.s. (1.000)	0.012*
Age	n.s. (0.236)	n.s. (0.440)	0.005**	n.s. (0.440)	<0.001***
Group × age	n.s. (0.409)	n.s. (1.000)	n.s. (1.000)	n.s. (0.337)	n.s. (1.000)
Group × gender	n.s. (1.000)	n.s. (1.000)	n.s. (1.000)	n.s. (0.262)	n.s. (1.000)

Significance codes – \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ , n.s., not significant.

Therefore, one-way and interaction ANOVAs were conducted to explore the individual effects of group, gender, birth year, and their interactions on each outcome variable. To control for inflation of the Type I error rate across the multiple ANOVAs,  $p$ -values were adjusted using the Holm method. The table below summarizes the F-values and adjusted  $p$ -values for each effect across the five dependent variables (Table 4):

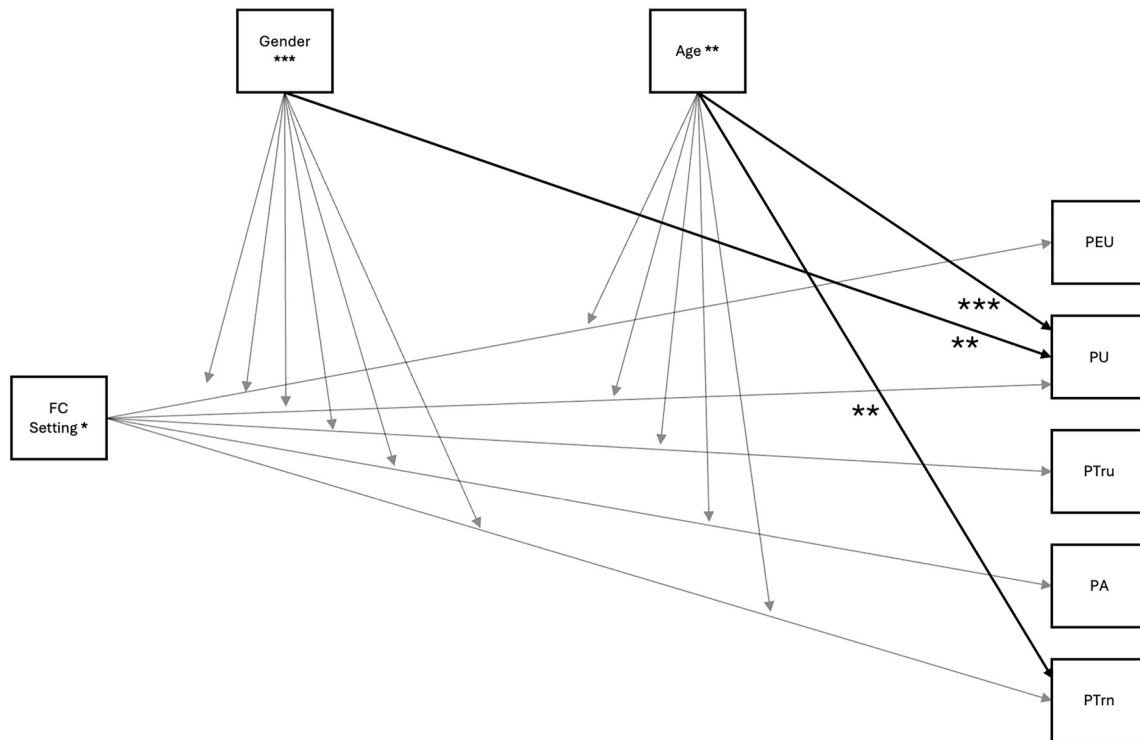
The ANOVA results reveal that the TC strategy did not have a significant effect on any individual dependent variable when considered separately. However, the MANOVA indicated that TC strategy significantly influenced the combined set of outcomes, suggesting a distributed impact across the overall construct profile rather than a strong effect on a single measure. In contrast, birth year showed strong main effects on both transparency (PTrn) and usefulness (PU), while gender significantly affected

usefulness (PU). These findings suggest that demographic variables – particularly age and gender – exert robust effects on specific constructs, whereas the TC strategy influences the broader multivariate perception profile.

### 4.6 Final model

In this section, we present our integrated conceptual framework in Figure 3. It synthesizes the hypothesized direct effects of tool-calling (TC) technique on five key user perceptions – trust, autonomy, transparency, ease of use, and usefulness – as well as the moderating roles of age and gender on these relationships. Path coefficients are annotated with significance levels: \* $p < 0.05$ , \*\* $p < 0.01$ , and \*\*\* $p < 0.001$ .

Figure 5 shows that, in line with our hypotheses, TC technique exerts a modest but significant multivariate effect



**Figure 5:** Function-calling effects on user perceptions, moderated by age and gender (\* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ ).

on the combined user perceptions (H1). Gender did not significantly moderate the relationship between TC strategy and user perceptions (H2 not supported), nor did age (H3 not supported). Instead, both gender and age exhibited significant direct effects on user perceptions, with gender influencing perceived usefulness (PU) and age influencing both transparency (PTrn) and usefulness (PU).

In the univariate follow-up tests, none of the direct effects on individual perceptions – PTru (H1a), PA (H1b), PTrn (H1c), PEU (H1d), or PU (H1e) – reached significance. Likewise, no TC  $\times$  gender interactions were significant for any outcome (H2a–H2e not supported). Age also did not significantly moderate the impact of technique on any of the outcomes (H3a–H3e not supported).

The following section discusses these results considering existing research and theoretical, as well as practical implications.

## 5 Discussion

This study investigated how different tool-calling (TC) strategies in CAs – automatic execution versus user-confirmed execution – affect user perceptions of trust, autonomy, transparency, ease of use, and usefulness. Drawing on a between-subjects experimental design with 462 participants, we evaluated the impact of these strategies across demographic factors such as gender and age.

While the overall multivariate analyses revealed statistically significant differences between TC strategies, the univariate results suggest that these effects are distributed across the outcome profile rather than concentrated in any single construct. By contrast, age and gender exerted direct effects on user perceptions, with age influencing transparency and usefulness, and gender influencing usefulness. These findings indicate that demographic factors play a more prominent role in shaping individual perceptions, whereas TC strategy primarily impacts the broader multivariate experience profile. In the following sections, we discuss the theoretical and practical implications of these results.

### 5.1 Theoretical implications

Building on the significant multivariate result for our full models (MANOVA/PERMANOVA), we now turn to its theoretical implications. The fact that tool-calling (TC) strategy, age, and gender jointly shape users' perceptions suggests that existing theories of technology acceptance and human–agent interaction must extend beyond a simple focus on automation versus control. In particular, our

findings connect directly to the Technology Acceptance Model (TAM), as both perceived usefulness (PU) and perceived ease of use (PEU) were significantly influenced by demographic factors. Rather than assuming uniform effects of execution strategy, our results indicate that user characteristics – especially age and gender – play a decisive role in shaping how CA interactions are evaluated. This underscores the importance of integrating demographic context into TAM-based accounts of adoption and sustained use of CAs.

Our finding that the age of the participants significantly predicts PA suggests that age influences how users experience control in CA interactions. Although prior CA research has rarely explored age-related differences in autonomy, Sheldon, Houser-Marko, and Kasser (2006) report that individuals tend to experience increased autonomy and well-being as they progress from early adulthood into middle age, potentially making older users particularly sensitive to autonomy-supportive mechanisms.<sup>83</sup> Our results indicate that demographic characteristics, particularly age and gender, had a significant effect on participants' perceptions of autonomy (PA). Rather than being primarily driven by the tool-calling strategy itself, differences in PA appear more strongly linked to these demographic factors, underscoring the importance of considering user characteristics when evaluating conversational agent interactions.

Regarding the age and gender effects on PU, extensive empirical research in HCI and TAM contexts has examined how user demographics moderate PU. Gender differences have been consistently observed. For example, in workplace studies of new software adoption, men's usage decisions rely more strongly on PU, whereas women rely more on PEU and social influence.<sup>39,84</sup> Venkatesh and Morris (2000) followed 342 employees over 5 months and found men's intentions were predominantly driven by PU, while women were influenced more by PEU and by others' opinions (subjective norm).<sup>39</sup>

Age differences in PU show a somewhat different pattern. Classic TAM-based studies indicate older adults do not necessarily doubt a technology's PU but often struggle more with usability. For instance, a study of office IT use found older staff perceived email and word processors as significantly less easy to use than younger users did, yet they rated the PU of these tools similarly to younger users.<sup>84</sup> Interestingly, our study revealed a contrasting result: age had a significant impact on PU, suggesting that older and younger adults differ in how useful they perceive CAs with external services. Further research is therefore required

to better understand these age-related differences, particularly in contexts where CAs interact with external functions and APIs.

For the last significant effect, research on PTrn across age groups is still limited. However, emerging evidence suggests age might play a role. For example, Gedrimiene et al. (2023) found that older users rated an AI system as significantly less transparent than younger users, despite reporting similar levels of trust. This aligns with the idea that younger, more digitally native users may require less explanation to feel a system is transparent, while older users benefit from more explicit clarity. Our findings reinforce this pattern, highlighting the need for age-aware transparency strategies in CA design, especially when external services are involved.<sup>85</sup>

The multivariate analyses confirmed our first hypothesis (H1): the tool-calling strategy alone had a statistically significant effect on the combined set of dependent variables. In addition, both age and gender showed significant main effects at the multivariate level, highlighting that demographic characteristics also play an important role in shaping user perceptions. These results suggest that while the choice of execution strategy matters, user demographics independently influence how the interaction is experienced. The moderation hypotheses (H2 and H3) therefore received partial support in that gender and age were significant predictors in the MANOVA, although their specific univariate moderation effects were less consistent.

To conclude the theoretical discussion, it is important to reflect on the hypothesized effects that were not supported in our study. Although the overall multivariate model was statistically significant – indicating that TC strategy, age, and gender jointly shape user perceptions – the majority of individual hypotheses (H1a–H1e, H2a–H2e, and H3a–H3e) did not reach significance on the univariate level. Specifically, we did not find significant differences in PTru, PTrn, PEU, or PU based solely on the TC strategy (H1a, H1c, H1d, H1e), nor did gender or age significantly moderate any of these relationships (H2a–H2e and H3a–H3e).

Nonetheless, the significant multivariate effect of tool-calling strategy underscores the need for further research to clarify its specific impact on individual user perceptions. Future studies should investigate under which conditions and in which contexts execution strategies influence trust, autonomy, transparency, ease of use, and usefulness more directly, as the absence of clear univariate effects in our study leaves important questions open.

The null results at the univariate level are nonetheless valuable, as they suggest that users' experiences with

tool-calling CAs may be shaped more strongly by demographic characteristics – such as age and gender – than by the execution strategy itself. At the same time, the significant multivariate effect of the tool-calling strategy indicates that it does play a role in shaping overall user perceptions, even if its influence on individual outcomes was not directly observable. This raises important questions: under what circumstances does strategy affect specific perceptions such as trust, transparency, or usefulness? Do these effects become more salient in particular domains, such as healthcare or finance, or with repeated use over time? Future research should continue to investigate these dynamics, ideally through longitudinal and domain-specific studies, to better understand how demographic factors and execution strategies jointly shape perceptions of autonomy, ease of use, and other key outcomes.

## 5.2 Practical implications

The findings of this study carry several important practical implications for the design and deployment of CAs, particularly those that involve external TC capabilities. Although TC strategy did not significantly affect individual outcomes such as trust, autonomy, transparency, ease of use, or usefulness when considered separately, the overall multivariate analyses (MANOVA and PERMANOVA) demonstrated that execution mode exerted a modest but significant influence on the combined set of user perceptions. This suggests that while no single construct was strongly shaped by strategy alone, the aggregate profile of user experience was affected. From a practical standpoint, this indicates that TC design choices should not be dismissed as irrelevant: execution strategy shapes how users evaluate CAs holistically, even if its influence on individual perceptions is diffuse.

To illustrate the practical relevance of this aggregate effect, consider domain-specific applications. For example, in the healthcare domain, digital consent mechanisms face strong demands for transparency, human oversight, and user control.<sup>86,87</sup> Users tend to accept conversational agents for routine health inquiries but expect fallback to human agents when stakes are higher.<sup>88</sup> Confirmation-based execution could enhance patient safety by ensuring that sensitive health data is only transmitted with explicit consent. In finance, users interacting with conversational agents highly prioritize assurances around privacy, security, and risk before consenting to automation.<sup>89</sup> Even among users exposed to similar financial assistants, those with lower initial trust amplify concerns of danger and uncertainty.<sup>90</sup> Moreover, perceived risk moderates the translation of usability and trust into behavioral intent in banking CA adoption.<sup>91</sup> Automatic execution may streamline routine

account inquiries but requires heightened transparency for high-stakes transactions.

Drawing on insights from the TAM, and current HCI research, we propose actionable guidelines tailored to different user demographics, especially in terms of age and gender.

First, the results highlight that ease of use perceptions are strongly shaped by age, independent of execution strategy. Younger participants consistently rated PEU higher than older participants, suggesting that they found the CA more intuitive and less effortful to use across conditions. This aligns with TAM research showing that age influences ease of use perceptions, with older adults often reporting greater difficulty in adapting to new technologies.<sup>92–94</sup> Importantly, our data indicate that confirmation prompts did not differentially disadvantage older users; rather, the overall age-related gap in PEU persisted across both tool-calling strategies. For CA designers, this underscores the importance of addressing age-related usability differences through accessible interface design and supportive guidance, rather than relying solely on automation to reduce interaction complexity.

Second, our results imply that gender did not significantly moderate the effects of TC strategy on user perceptions, but it did have a significant impact on PU. This suggests that gender plays an important role in shaping how users evaluate the utility of CAs, even if it does not alter the effectiveness of specific TC strategies. For example, Venkatesh and Morris (2000) found that men and women rely on different predictors of technology use: men prioritize PU, whereas women are more influenced by PEU and social norms.<sup>39</sup> Although gender effects were not significant in our sample, these distinctions should not be ignored in broader applications. Inclusive design practices – such as allowing users to customize the CA's voice and persona or avoiding gendered stereotypes in agent behaviour – are essential for equitable user experiences.

Third, although our model did not find a significant main or interaction effect of the FC setting on PTRn, it remains a theoretically and practically relevant design consideration. Importantly, our results did show a significant main effect of age on PTRn, with younger participants rating transparency higher than older participants. This suggests that age shapes how users perceive the clarity of CA operations, independent of execution strategy. Design practices such as clearly labelling external data sources, providing user confirmations before executing functions, and offering brief explanations of system actions are still aligned with human–AI interaction guidelines.<sup>95</sup> Such measures may be particularly valuable for older users, for whom explicit

transparency cues can help bridge perceived gaps in system comprehensibility. These strategies may also prove useful in specific use cases or user groups where transparency becomes more salient.

Finally, practical deployment of TC CAs should incorporate adaptive interaction strategies informed by the significant multivariate effect observed in our study. Although no single dependent variable (trust, autonomy, transparency, ease of use, usefulness) was individually shaped by the TC setting, the significant multivariate result demonstrates that execution mode influences the overall profile of user perceptions. This implies that TC design choices meaningfully affect how users experience a CA holistically, even if the precise dimension of influence varies across contexts or user groups.

From a design perspective, this means that execution strategies should not be treated as interchangeable defaults. Instead, CA developers could:

- Differentiate execution mode by task type: apply automatic execution for low-risk, routine actions (e.g., retrieving account balances) while requiring confirmation for high-stakes tasks (e.g., financial transfers, health data sharing).
- Enable user-selectable modes: offer onboarding options where users choose a preferred interaction style – automatic, confirm-first, or adaptive – acknowledging that preferences may differ by demographic group.
- Adapt dynamically: use lightweight user modeling (e.g., inferring age or interaction preferences from behavior) to adjust confirmation frequency or explanation detail in real time.

These strategies operationalize the observed multivariate effect: even modest, diffuse influences of TC setting can shape the aggregate user experience. By embedding adaptive or configurable TC modes, CAs can balance efficiency with control, ensuring that automation design choices actively account for their demonstrated impact on user perceptions. This approach aligns with inclusive design principles,<sup>96</sup> supporting sustained adoption across diverse user groups.

In sum, this study underscores the importance of demographic-sensitive design in CA interactions. Developers and UX practitioners should adopt evidence-based strategies that account for users' cognitive and emotional needs, balancing automation with control, and ensuring PTRn in external service integration. Doing so will not only enhance PEU, but also promote PTRu, accessibility, and long-term engagement with CAs. Furthermore, this study emphasizes that human-in-the-loop approaches might be valuable

during the application of AI models – providing real-time oversight, corrective feedback, and adaptive intervention to ensure AI behaviors remain aligned with users' needs and contexts, not just during model training<sup>97</sup> but throughout deployment and use.

## 6 Conclusions

This study investigated how two distinct TC strategies in German-language CAs – automatic execution based on conversational context (CA 1) versus user-confirmed execution (CA 2) – shape users' perceptions of trust, autonomy, transparency, ease of use, and usefulness. Employing a between-subjects experimental design with 462 participants and complementary MANOVA and PERMANOVA analyses, we found that the overall TC approach exerts a modest but statistically significant multivariate effect on combined user perceptions. However, univariate follow-up tests revealed that this effect did not translate into significant differences on individual constructs of PTru, PTRn, PEU, or PU. Instead, demographic factors – particularly age and gender – emerged as stronger predictors, with age significantly influencing PU and PTRn, and gender showing a significant effect on PU.

Our findings extend technology acceptance and human-agent interaction theories by demonstrating that the benefits and drawbacks of automated versus user-confirmed tool-calls are better understood when considered alongside demographic context. The absence of univariate strategy effects on PTru, PTRn, PEU, and PU suggests that execution mode alone may not be the decisive factor shaping the given variables. Instead, age and gender exerted stronger direct influences: younger participants rated PU and PTRn higher than older participants, and men rated PU higher than women.

For practitioners, this underscores the importance of designing adaptive conversational agents that address demographic differences directly – for example, by simplifying explanations and usability features for older users while emphasizing utility and task performance for younger ones. While TC strategy should not be ignored given its significant multivariate effect, tailoring interfaces to demographic user profiles appears to be the more effective route for improving trust, usability, and long-term adoption.

## 7 Limitations and future research

This study has several limitations that should be considered. First, our use of a single German-language

business-travel scenario may limit the generalizability of findings to other domains, tasks, languages, or cultural contexts. Future research should replicate this comparison across diverse use cases (e.g., healthcare, finance), languages, and interaction modalities (e.g., voice vs. text). Second, we recruited via a single online panel (Bilendi), which may introduce self-selection bias and yield a relatively homogeneous sample. We also did not measure participants' technical literacy, prior experience with chatbots or AI, or task-specific expertise – all of which likely moderate PEU, PTru, and related perceptions. Subsequent studies should purposively sample across different experience levels and include objective assessments of AI familiarity. Third, the cross-sectional design captures only immediate, post-interaction perceptions and cannot speak to how these perceptions evolve over time or with repeated use. Fourth, although monetary incentives (2 € per participant) were necessary to ensure timely recruitment, they may have affected participants' motivation, effort, or satisfaction in ways that could bias self-reported measures. Lastly, it should be acknowledged as a limitation that several of our Cronbach's alpha coefficients exceeded 0.90 – values at this level may indicate item redundancy and should be taken into account when interpreting the scales' reliability.<sup>98</sup> Future studies should employ longitudinal designs, explore a broader range of application scenarios, incorporate additional user characteristics, and consider non-monetary or varied incentive structures to better understand how TC approaches influence sustained engagement and real-world adoption.

**Acknowledgments:** The authors wish to acknowledge that large language models were employed for translation, paraphrasing, and summarization of certain passages. All outputs were subsequently reviewed and edited by the authors to ensure accuracy and consistency with the original meaning.

**Research ethics:** Not applicable.

**Informed consent:** Informed consent was obtained from all individuals included in this study, or their legal guardians or wards.

**Author contributions:** All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

**Use of Large Language Models, AI and Machine Learning Tools:** See Acknowledgments, tools used: ChatGPT.

**Conflict of interest:** All other authors state no conflict of interest.

**Research funding:** None declared.

**Data availability:** The raw data can be obtained on request from the corresponding author.

## Appendix

See Figure 6.



Figure 6: Screenshot of the German Chat-UI.

## References

1. Meta-Llama (Meta Llama), 2025. <https://huggingface.co/meta-llama> (accessed 2025-05-08).
2. Models - OpenAI API, 2025. <https://platform.openai.com> (accessed 2025-05-08).
3. Models Overview | Mistral AI Large Language Models, 2025. [https://docs.mistral.ai/getting-started/models/models\\_overview/](https://docs.mistral.ai/getting-started/models/models_overview/) (accessed 2025-05-08).
4. Hennekeuser, D.; Vaziri, D.; Golchinfar, D.; Stevens, G. What I Don't like about You? A Systematic Review of Impeding Aspects for the Usage of Conversational Agents. *Interact. Comput.* **2024**, *36* (5), 293–312.
5. Adamopoulou, E.; Moussiades, L. Chatbots: History, Technology, and Applications. *Mach. Learn. Appl.* **2020**, *2*, 100006.
6. Kowalski, J. Older Adults and Voice Interaction: a Pilot Study with Google Home. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (New York, NY, USA, 2019)*, 2019; pp. 1–6.
7. Nallam, P.; Bhandari, S.; Sanders, J.; Martin-Hammond, A. A Question of Access: Exploring the Perceived Benefits and Barriers of Intelligent Voice Assistants for Improving Access to Consumer Health Resources Among Low-Income Older Adults. *Gerontol. Geriatr. Med.* **2020**, *6*, 2333721420985975.
8. Lewis, P. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv **2021**.
9. Function Calling | Mistral AI Large Language Models, 2024. [https://docs.mistral.ai/capabilities/function\\_calling/](https://docs.mistral.ai/capabilities/function_calling/) (accessed 2024-11-13).
10. Brown, T. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, 1877–1901.
11. OpenAI 2023. GPT-4 Technical Report. arXiv.
12. Radford, A.; Narasimhan, K. *Improving Language Understanding by Generative Pre-Training*; Technical Report, 2018.
13. Radford, A. Language Models are Unsupervised Multitask Learners. *OpenAI blog* **2019**, *1* (8), 9.
14. Touvron, H. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv **2023**.
15. Touvron, H. LLaMA: Open and Efficient Foundation Language Models. arXiv **2023**.
16. Abou Ali, M.; Dornaika, F.; Charafeddine, J. Agentic AI: A Comprehensive survey of Architectures, Applications, and Future Directions. *Artif. Intell. Rev.* **2025**, *59* (1), 11.
17. Qin, Y.; Hu, S.; Lin, Y.; Chen, W.; Ding, N.; Cui, G. Tool Learning with Foundation Models. *ACM Comput. Surveys* **2025**, *57* (4), 1–40.
18. Cabezas, D. Integrating a LLaMa-based Chatbot with Augmented Retrieval Generation as a Complementary Educational Tool for High School and College Students. In *Proceedings of the 19th International Conference on Software Technologies - ICSoft (2024)*, 2024; pp. 395–402.
19. Hennekeuser, D.; Vaziri, D. D.; Golchinfar, D.; Schreiber, D.; Stevens, G. Enlarged Education — Exploring the Use of Generative AI to Support Lecturing in Higher Education. *Int. J. Artif. Intell. Educ.* **2024**, *35*, 1096–1128.
20. Šarčević, A. Enhancing Programming Education with Open-Source Generative AI Chatbots. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, 2024; pp. 2051–2056.
21. Bhat, V. Retrieval Augmented Generation (RAG) Based Restaurant Chatbot with AI Testability. In *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService) (Los Alamitos, CA, USA, July 2024)*, 2024; pp. 1–10.
22. Vidivelli, S.; Ramachandran, M.; Dharunbalaji, A. Efficiency-Driven Custom Chatbot Development: Unleashing LangChain, RAG, and Performance-Optimized LLM Fusion. *Comput. Mater. Continua.* **2024**, *80* (2), 2423–2442.
23. Huang, S. Planning, Creation, Usage: Benchmarking LLMs for Comprehensive Tool Utilization in Real-World Complex Scenarios. In *Findings of the Association for Computational Linguistics: ACL 2024 (Bangkok, Thailand, Aug. 2024)*, 2024; pp. 4363–4400.
24. Qin, Y. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-World APIs. In *The Twelfth International Conference on Learning Representations*; OpenReview, 2024.
25. Theuma, A.; Shareghi, E. Equipping Language Models with Tool Use Capability for Tabular Data Analysis in Finance. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers) (St. Julian's, Malta, Mar. 2024)*, 2024; pp. 90–103.

26. Schick, T. Toolformer: Language Models Can Teach Themselves to Use Tools. *Adv. Neural Inf. Process. Syst.* **2023**, 68539–68551.
27. Abdelaziz, I. Granite-Function Calling Model: Introducing Function Calling Abilities via Multi-task Learning of Granular Tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track (Miami, Florida, US, Nov. 2024)*, 2024; pp. 1131–1139.
28. Zhuang, Y. ToolQA: A Dataset for LLM Question Answering with External Tools. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (Red Hook, NY, USA, 2024)*, 2024.
29. OpenAI Platform, 2024. <https://platform.openai.com> (accessed: 2024-11-13).
30. Schnall, R.; Higgins, T.; Brown, W.; Carballo-Dieguez, A.; Bakken, S. Trust, Perceived Risk, Perceived Ease of Use and Perceived Usefulness as Factors Related to mHealth Technology Use. *Stud. Health Technol. Inf.* **2015**, *216*, 467–471.
31. Ryan, R. M.; Deci, E. L. Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and well-being. *Am. Psychol.* **2000**, *55* (1), 68–78.
32. Wang, W.; Benbasat, I. Empirical Assessment of Alternative Designs for Enhancing Different Types of Trusting Beliefs in Online Recommendation Agents. *J. Manag. Inf. Syst.* **2016**, *33* (3), 744–775.
33. Venkatesh, V.; Bala, H. Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decis. Sci.* **2008**, *39* (2), 273–315.
34. Brooke, J. SUS: A Quick and Dirty Usability scale. *Usability Eval. Ind.* **1995**, 189.
35. Laugwitz, B. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work*; Holzinger, A., Ed.; Springer: Berlin Heidelberg, 2008; pp. 63–76.
36. Kim, W. Chapter 11 – Shopping with AI: Consumers’ Perceived Autonomy in the age of AI. In *Human-Centered Artificial Intelligence*; Nam, C. S., Ed.; Academic Press: Raleigh, 2022; pp 157–171.
37. Dobrowski, Z.; Drozdowski, G.; Panait, M. Understanding the Impact of Generation Z on Risk Management—A Preliminary Views on Values, Competencies, and Ethics of the Generation Z in Public Administration. *Int. J. Environ. Res. Publ. Health* **2022**, *19* (7), 3868.
38. Venkatesh; Morris, M. G.; Davis, G. B.; Davis, F. D. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* **2003**, *27* (3), 425–478.
39. Venkatesh, V.; Morris, M. G. Why Don’T Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior. *Soc. Sci. Res. Netw.* **2020**.
40. Alagarsamy, S.; Mehroliya, S. Exploring Chatbot Trust: Antecedents and Behavioural Outcomes. *Heliyon* **2023**, *9* (5), e16074.
41. Ding, Y.; Najaf, M. Interactivity, Humanness, and Trust: A Psychological Approach to AI Chatbot Adoption in e-commerce. *BMC Psychol.* **2024**, *12* (1), 595.
42. Seymour, W.; Van Kleek, M. Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants. *Proc. ACM Hum.-Comput. Interact.* **2021**, *5*, CSCW2–16.
43. M, J.; P, V. S.; Kryvinska, N. Exploring the Chatbot Usage intention-A Mediating Role of Chatbot Initial Trust. *Heliyon* **2024**, *10* (12), e33028.
44. Ramrath, M. Trust in AI Chatbots: The Perceived Expertise of Chatgpt in Subjective and Objective Tasks. In *Frontiers in Artificial Intelligence and Applications*; Lorig, F., Ed.; IOS Press: Amsterdam, 2024.
45. Brill, T. M.; Munoz, L.; Miller, R. J. Siri, Alexa, and Other Digital Assistants: A Study of Customer Satisfaction with Artificial Intelligence Applications. *J. Market. Manag.* **2019**, *35* (15–16), 1401–1436.
46. Nasirian, F. AI-Based Voice Assistant Systems: Evaluating from the Interaction and Trust Perspectives. *Americas Conference on Information Systems 2017* **2017**.
47. Sankaran, S.; Zhang, C.; Aarts, H.; Markopoulos, P. Exploring Peoples’ Perception of Autonomy and Reactance in Everyday AI Interactions. *Front. Psychol.* **2021**, *12*, 713074.
48. Cai, W. “Listen to Music, Listen to Yourself”: Design of a Conversational Agent to Support Self-Awareness While Listening to Music. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (New York, NY, USA, 2023)*, 2023.
49. Xiaohan Hu, X. X.; Chen, C. (Crystal) Investigating the Effects of Perceived Autonomy in Chatbot Advertising. *J. Interact. Advert.* **2023**, *23* (4), 323–338. <https://doi.org/10.1080/15252019.2023.2262456>.
50. Yu, L.; Li, Y. Artificial Intelligence Decision-Making Transparency and Employees’ Trust: the Parallel Multiple Mediating Effect of Effectiveness and Discomfort. *Behav. Sci.* **2022**, *12* (5), 127.
51. Xu, Y.; Bradford, N.; Garg, R. Transparency Enhances Positive Perceptions of Social Artificial Intelligence. *Hum. Behav. Emerg. Technol.* **2023**, *2023*, 1–15.
52. Wang, X.; Qiu, X. The Positive Effect of Artificial Intelligence Technology Transparency on Digital Endorsers: Based on the Theory of Mind Perception. *J. Retailing Consum. Serv.* **2024**, *78*, 103777.
53. Davis, F. D. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* **1989**, *13* (3), 319–340.
54. Iancu, I.; Iancu, B. Interacting with Chatbots Later in Life: A Technology Acceptance Perspective in COVID-19 Pandemic Situation. *Front. Psychol.* **2023**, *13*, 1111003.
55. Wutz, M.; Hermes, M.; Winter, V.; Köberlein-Neu, J. Factors Influencing the Acceptability, Acceptance, and Adoption of Conversational Agents in Health Care: Integrative Review. *J. Med. Internet Res.* **2023**, *25*, e46548.
56. Academic Research 2025. <https://www.bilendi.us> (accessed 2025-10-16).
57. Corritore, C. Measuring Online Trust of Websites: Credibility, Perceived Ease of Use, and Risk. In *AMCIS 2005 Proceedings*; Paper 370, 2005.
58. Gagné, M. The Role of Autonomy Support and Autonomy Orientation in Prosocial Behavior Engagement. *Motiv. Emot.* **2003**, *27* (3), 199–223.
59. Nguyen, Q. N.; Sidorova, A.; Torres, R. User Interactions with Chatbot Interfaces vs. Menu-based Interfaces: An Empirical Study. *Comput. Hum. Behav.* **2022**, *128*, 107093.
60. Yan, F. Berkeley Function Calling Leaderboard. In *42nd International Conference on Machine Learning*; Vancouver: Canada, 2025.
61. Yang, A. Qwen2.5 Technical Report. arXiv 2025.
62. Argilla-Warehouse/Apigen-Synth-Trl Datasets at Hugging Face 2025. <https://huggingface.co/datasets/argilla-warehouse/apigen-synth-trl> (accessed 2025-03-19).
63. Liu, W.; Huang, X.; Zeng, X.; Hao, X.; Wang, S.; Gan, W.; Liu, Z.; Yu, Y.; Wang, Z.; Wang, Y.; Ning, W.; Hou, Y.; Wang, B.; Wu, C.; Wang,

- X.; Liu, Y.; Wang, Y.; Tang, D.; Tu, D.; Shang, L.; Jiang, X.; Tang, R.; Lian, D.; Liu, Q.; Chen, E. Toolace: Winning the Points of LLM Function Calling. *arXiv* **2024**, <https://doi.org/10.48550/arXiv.2409.00920>.
64. Hartford, E. Spectrum: Targeted Training on Signal to Noise Ratio. *arXiv* **2024**.
  65. PostgreSQL 2025. <https://www.postgresql.org/> (accessed 2025-05-12).
  66. Streamlit - A Faster Way to Build and Share Data Apps 2021. <https://streamlit.io/> (accessed 2025-05-12).
  67. Quickstart - vLLM 2025. [https://docs.vllm.ai/en/latest/getting\\_started/quickstart.html](https://docs.vllm.ai/en/latest/getting_started/quickstart.html) (accessed 2025-05-12).
  68. Tabachnick, B. G.; Fidell, L. S. *Using Multivariate Statistics*; HarperCollins College Publ: New York, 1996.
  69. Warne, R. A Primer on Multivariate Analysis of Variance (MANOVA) for Behavioral Scientists. <https://doi.org/10.7275/SM63-7H70>.
  70. Tabachnick, B. G.; Fidell, L. S. *Using Multivariate Statistics*; Pearson: London, 2013.
  71. Mardia, K. V. Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika* **1970**, *57* (3), 519–530.
  72. Blanca, M.; Alarcón, R.; Arnau, J.; Bono, R.; Bendayan, R. Non-Normal Data: Is ANOVA Still a Valid Option? *Psicothema* **2017**, *4* (29), 552–557.
  73. Verma, J. P. *Repeated Measures Design for Empirical Researchers*; John Wiley & Sons, Inc: New York, 2015.
  74. Warner, R. M. *Applied Statistics: From Bivariate Through Multivariate Techniques*, 2nd ed.; Sage Publications, Inc.: Thousand Oaks, 2013.
  75. Anderson, M. J. Permutational Multivariate Analysis of Variance (PERMANOVA). In *Wiley StatsRef: Statistics Reference Online*; Kenett, R. S., Ed.; Wiley: Hoboken, 2017; pp 1–15.
  76. Strohle, L.; Wold, S. Analysis of Variance (ANOVA). *Chemom. Intell. Lab. Syst.* **1989**, *6* (4), 259–272.
  77. Shapiro, S. S.; Wilk, M. B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **1965**, *52* (3–4), 591–611.
  78. Brown, M. B.; Forsythe, A. B. Robust Tests for the Equality of Variances. *J. Am. Stat. Assoc.* **1974**, *69* (346), 364–367.
  79. Glass, G. V.; Peckham, P. D.; Sanders, J. R. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Rev. Educ. Res.* **1972**, *42* (3), 237–288.
  80. Harwell, M. R.; Rubinstein, E. N.; Hayes, W. S.; Olds, C. C. Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases. *J. Educ. Stat.* **1992**, *17* (4), 315.
  81. Lix, L. M.; Keselman, J. C.; Keselman, H. J. Consequences of Assumption Violations Revisited: a Quantitative Review of Alternatives to the One-Way Analysis of Variance “F” Test. *Rev. Educ. Res.* **1996**, *66* (4), 579.
  82. Schmider, E.; Ziegler, M.; Danay, E.; Beyers, L.; Bühner, M. Is It Really Robust? Reinvestigating the Robustness of ANOVA Against Violations of the Normal Distribution Assumption. *Methodology* **2010**, *6* (4), 147–151.
  83. Sheldon, K. M.; Houser-Marko, L.; Kasser, T. Does Autonomy Increase with Age? Comparing the Goal Motivations of College Students and their Parents. *J. Res. Pers.* **2006**, *40* (2), 168–178.
  84. Symasek, L.; Yeazitsis, T.; Weger, K.; Mesmer, B. Recent Developments in Individual Difference Research to Inform the Adoption of AI Technology. *Systems* **2025**, *13* (3), 156.
  85. Gedrimiene, E.; Celik, I.; Mäkitalo, K.; Muukkonen, H. Transparency and Trustworthiness in User Intentions to Follow Career Recommendations from a Learning Analytics Tool. *J. Learn. Anal.* **2023**, *10* (1), 54–70.
  86. Allen, J. W.; Earp, B. D.; Koplin, J.; Wilkinson, D. Consent-GPT: Is it Ethical to Delegate Procedural Consent to Conversational AI? *J. Med. Ethics* **2024**, *50* (2), 77–83.
  87. Goldschmitt, M.; Gleim, P.; Mandelartz, S.; Kellmeyer, P.; Rigotti, T. Digitalizing Informed Consent in Healthcare: A Scoping Review. *BMC Health Serv. Res.* **2025**, *25* (1), 893.
  88. Seitz, L.; Bekmeier-Feuerhahn, S.; Gohil, K. Can we Trust a Chatbot like a Physician? A Qualitative Study on Understanding the Emergence of Trust Toward Diagnostic Chatbots. *Int. J. Hum. Comput. Stud.* **2022**, *165*, 102848.
  89. Ng, M. In Private, Secure, Conversational FinBots We Trust. *arXiv* 2022.
  90. Schreibelmayr, S.; Moradbakhti, L.; Mara, M. First Impressions of a Financial AI Assistant: Differences Between High Trust and Low Trust Users. *Front. Artif. Intell.* **2023**, *6*, 1241290.
  91. Hasan, S.; Godhuli, E. R.; Rahman, M. S.; Mamun, M. A. A. The Adoption of Conversational Assistants in the Banking Industry: Is the Perceived Risk a Moderator? *Heliyon* **2023**, *9* (9), e20220.
  92. Cao, X.; Zhang, H.; Zhou, B.; Wang, D.; Cui, C.; Bai, X. Factors Influencing Older Adults’ Acceptance of Voice Assistants. *Front. Psychol.* **2024**, *15*, 1376207.
  93. Yu, S.; Chen, T. Understanding Older Adults’ Acceptance of Chatbots in Healthcare Delivery: An Extended UTAUT Model. *Front. Public Health* **2024**, *12*, 1435329.
  94. Zhong, R.; Ma, M.; Zhou, Y.; Lin, Q.; Li, L.; Zhang, N. User Acceptance of Smart Home Voice Assistant: a Comparison Among Younger, middle-aged, and Older Adults. *Univers. Access Inf. Soc.* **2024**, *23* (1), 275–292.
  95. Amershi, S. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (New York, NY, USA, 2019)*, 2019; pp 1–13.
  96. Wang, Y.-L.; Lo, C.-W. The Effects of Response Time on Older and Young Adults’ Interaction Experience with Chatbot. *BMC Psychol.* **2025**, *13* (1), 150.
  97. Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; Fernández-Leal, Á. Human-In-The-Loop Machine Learning: a State of the Art. *Artif. Intell. Rev.* **2023**, *56* (4), 3005–3054.
  98. Tavakol, M.; Dennick, R. Making Sense of Cronbach’s Alpha. *Int. J. Med. Educ.* **2011**, *2*, 53–55.