Understanding the Effects of Conversational Agent Personality on the Credibility of LLM-Based Conversational Search

Hyeonjeong Byeon
University of Washington
Seattle, USA
hjbyeon@uw.edu

Uran Oh
Ewha Womans University
Seoul, Republic of Korea
uran.oh@ewha.ac.kr

Gary Hsieh
University of Washington
Seattle, USA
garyhs@uw.edu

Abstract

The rise of Large Language Models (LLMs) has ushered in a wave of conversational search engines that allow people to engage in dialogues with LLM-infused chatbots to seek information. As people tend to infer personalities from digital social interactions, and given that personality cues have been shown to affect credibility, these perceptions of chatbot design may shape how users assess the credibility of information in conversational search. In this study, we conducted a controlled online study with 190 participants who assessed conversational search results with chatbots designed to exhibit different levels of personality traits. We found that in conversational search, personality can affect perceptions of credibility. Specifically, perceived conscientiousness and agreeableness of a chatbot can increase credibility, while perceived extraversion and neuroticism can decrease the credibility of the information. This research contributes to our understanding of how conversational interfaces and their personality and persona designs can impact credibility. We also provide design implications for conversational search interfaces based on our findings.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

conversational agent, conversational search interface, credibility, personality

ACM Reference Format:
Hyeonjeong Byeon, Uran Oh, and Gary Hsieh. 2026. Understanding the Effects of Conversational Agent Personality on the Credibility of LLM-Based Conversational Search. In *2026 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '26), March 22–26, 2026, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3786304.3788843

1 Introduction

Building on decades of research in dialog-based information access [49], the advent of Large Language Model (LLM)-infused conversational agents has brought renewed momentum to conversational search, reshaping how users interact with information systems. Instead of typing keywords into a text box to find relevant information as we have done with traditional search engines (*e.g.,* Google's Search, Microsoft's Bing), people can now engage in dialogues and use complete sentences and receive responses that are also structured in complete sentences from conversational chatbots (*e.g.,* OpenAI's ChatGPT, Microsoft's Bing Chat, Google's Gemini). While the idea of 'conversational search' has appeared in multiple years of research [4, 22, 36], the implementation of conversational search interfaces, especially in an open-domain context faced several technical challenges [55]. Addressing these technical difficulties, LLM-powered search systems instantly gained popularity and settled in as a prevalent system in our everyday lives. ChatGPT for example, has attracted 1 million users within a week after its initial launch and has about 180 million monthly active users [11]. According to a survey with adults in the United States [14], the most prevalent uses of ChatGPT was for information gathering (36.1%), followed by entertainment (33.4%), and problem-solving (22.2%). This rise of AI-powered conversational searches raises questions about how to effectively design these interfaces.

A key design consideration for conversational agents in prior research has been their personality [57]. Prior work has found that imbuing personalities in conversational agent design has been shown to influence user trust [9, 66], affection [7–9, 48, 61], engagement in conversation [56, 66], and self-disclosure [24, 66]. For example, in a sales context, by tailoring chatbot design to align with the brand personality, consumers were almost twice as willing to trade up to more costly options and add-on services. Consumers mentioned that they enjoyed the ability to connect to the chatbots' personality [29]. Also, in a mental healthcare context, where the patients' engagement and self-disclosure about their symptoms are important for an effective treatment, chatbots that are designed with a conscientious personality were able to elicit higher engagement and longer responses from the patient [47].

However, while adding personalities in chatbot may help increase user engagement, its use may have an inadvertent affect on credibility for conversational search.

Information credibility, which encompasses the trustworthiness, expertise, reliability, and accuracy of information [53], is thus crucial to the success of search interfaces [65]. In interpersonal communication, a speaker's personality has been shown to affect credibility. For example, people's perception of the speaker can shape their willingness to find the presented information credible. In contexts such as mock courtrooms, organizations, or work units, a speaker's conversational style, influenced by their personality, can impact judgments about the quality and believability of the information shared [15, 63]. As chatbots are designed to mimic human conversations, personality of the chatbot may also affect credibility. As the rise of these "intelligent" chatbots ushers in a paradigm shift in information search, understanding how designs of the chabot

CHIIR '26, Seattle, WA, USA

ACM ISBN 979-8-4007-2414-5/26/03
https://doi.org/10.1145/3786304.3788843

may affect the credibility of conversational searches is a timely and important research topic.

To advance our knowledge and to provide guidance for interface design for conversational search, we conducted an online experiment with 190 participants to explore whether chatbot personalities can affect information credibility. Our findings suggest that the information delivered by chatbots with personalities high in conscientiousness, high in agreeableness, low in extraversion, and low in neuroticism were perceived to have greater credibility.

In this work, we offer these following contributions: (1) we provide empirical evidence showing that the personality of the conversational search chatbot can affect information credibility, (2) we advance our understanding on how different personality traits can influence credibility and show that some personalities improve, while other personalities reduce credibility, and (3) we offer guidelines and implications for conversational search design.

2 Related Work

2.1 Credibility and Trust in Chatbots

Credibility and trust, while related, are distinct constructs [59]. Credibility refers to the believability of information or advice provided by a source, whether that source is a person, system, or process [30]. It is primarily concerned with how accurate, truthful, and reliable the information itself is perceived to be [31]. Trust, in contrast, is a longitudinal evaluation directed toward the source of that information. It develops gradually through repeated demonstrations of reliability and dependability, shaping the user's willingness to rely on the source in future interactions [52].

Trust has been a construct of much prior study for chatbots, with research examining how design features such as anthropomorphic elements, small talk, and embodiment in human-like forms can increase trust and foster long-term user engagement during interaction [37, 40]. However, as conversational agents are increasingly used for information search and retrieval, credibility now emerges as a key factor in how users evaluate conversational agents. As people now turn to conversational interfaces to find and receive information, the question arises of how the design and use of conversational interactions can affect credibility. In the information-seeking context, credibility has been shown to play an important role in influencing user actions. For example, information perceived as credible leads to higher intentions to share the content [51], higher chances of the information receiver being persuaded by the message [64], and a higher likelihood of the audience taking action based on the information [39]. Much prior work has also explored how the design of search engines and search result pages can influence credibility [10, 20, 23, 34, 54].

In interpersonal communication, perceived qualities of the communicator, such as competence, character, extroversion, sociability, composure, and attractiveness, have been shown to significantly impact their credibility [5, 43, 44]. Furthermore, individual differences, including personality traits, have been found to shape these qualities [3, 41]. Given that chatbots are designed to emulate human interaction, it is plausible that similar dynamics apply when assessing the credibility of information conveyed by chatbots. Specifically, the perceived personality of a chatbot may play a critical role in shaping how users evaluate the credibility of its responses.

To address this gap, this study aims to investigate how perceptions of a chatbot's personality influence the perceived credibility of the information it provides in the context of conversational search. This leads to the following research question:

- **RQ1**. Can perceptions of chatbot's personality affect credibility of information in conversational search?

2.2 Chatbot Personality and Credibility

In our work, we utilize the well-established Five Factor Model (FFM), also referred to as the Big Five Trait Taxonomy [25, 33], to measure and study the effects of chatbot personality on credibility. The model consists of five characteristics: conscientiousness, agreeableness, extraversion, openness, and neuroticism. In this section, we hypothesize how each personality trait may contribute to chatbot's credibility.

Conscientiousness includes traits relating to competence, being organized, dutifulness, achievement striving, self-discipline, and deliberation [33]. These personality traits are strongly associated with job performance, and individuals who demonstrate these traits tend to be seen as reliable and dependable by their coworkers. This reliability fosters trust, which, in turn, enhances their credibility when sharing information [21, 50]. Also, prior studies have shown that individuals with more conscientious personalities tend to exhibit higher levels of competence [6, 58]. Similar effects were also found in the chatbot context. In a study where a chatbot was designed to act as an academic advisor, participants pointed out that chatbots with high conscientiousness seemed more competent and rated higher trust scores than chatbots designed with less conscientiousness [38]. When a chatbot was used in a mental healthcare context to provide medical information, chatbots designed with high conscientiousness were perceived to be professional and give the most useful advice [47]. Base on these prior works, we hypothesize as follows:

- **H1**. Chatbots high in conscientiousness will be perceived as higher in credibility.

Next, previous studies also suggest a positive effect of agreeable personalities on credibility. Agreeableness includes traits related to trust, altruism, compliance, modesty, and tender-mindedness [33]. Agreeableness plays an important role in interpersonal interactions by fostering a sense of trust and rapport between individuals [27]. For example, in a work unit environment, individuals with agreeable personality traits easily bond with coworkers and build stronger employer-employee trust relations [63]. Also, Fulton *et al.,*[21] suggests that agreeableness is a key factor in explaining the charismatic appeal of speakers, as it contributes to the emotional and relational aspects of credibility. The relationship between chatbots' agreeableness and credibility is relatively less explored. However, there is evidence that agreeable personalities imbued in chatbots can make users perceive these 'sociable' qualities of credibility, significantly affecting user experiences and perceptions. For example, when users were presented with three chatbots with different levels of agreeableness, chatbots designed with low agreeableness were preferred the least to interact with among the three, as the disagreeable chatbot was designed with antisocial language markers [60]. Another study suggests that users perceived agreeable chatbots with high levels of trust because of their empathetic behaviors [38]. Thus, we hypothesize the following:

- **H2**. Chatbots high in agreeableness will be perceived as higher in credibility.

Extraversion can be explained as a personality trait characterized by qualities such as sociability, assertiveness, talkativeness, enthusiasm, and a high level of activity or energy. Individuals who are extroverted tend to seek out social interactions, enjoy being in the company of others, and are often perceived as outgoing and energetic. They are generally more comfortable in group settings, are likely to initiate conversations, and often express themselves with confidence and enthusiasm [33]. Directly, extraversion has been considered factor of credibility [43]. Indirectly, extraversion can also affect the attractiveness of the speaker, which in itself is also another factor of credibility in the source-attractiveness model [44].

In the chatbot context, while the link between extraversion and crediblity has not been empirically shown, extraverted chatbots are perceived to be 'likeable' and preferred by users. One study compared extraverted and introverted chatbots interacting with the user about everyday conversations, such as weekend plans, music, books and travel. The results indicate that highly extraverted conversational agents are generally better received in terms of social presence and communication satisfaction [2]. Volkel *et al.,*[62] studied different levels of extraverted personalities of chatbots in a healthcare context, and found out that users preferred to interact with an extraverted chatbot after repeated use. Participants appreciated the friendly interaction experience with the extraverted chatbot, and the human-likeliness of the conversation. These traits contributed to the attractive style of conversation, resulting a higher preference among the users. Here, we hypothesize that:

- **H3**. Chatbots high in extroversion will be perceived as higher in credibility.

Openness describes the breadth, depth, originality, and complexity of an individual's mental and experiential life [33]. Individuals with high in openness lead the conversation with other people by introducing diverse topics [42]. Openness to experience is considered important in relationship building in interpersonal interactions [35], as individuals high in openness are often seen as more knowledgeable and insightful [13]. For instance, prior work has explored personality traits of workers in work units. Researchers found out that high levels of interaction supports proactive information exchange in work units, leading to higher level of trust among the group [50]. In addition, intellectual openness contributes to the 'competence' quality of credibility. Along with conscientiousness, openness is also considered as a key factor that influences one's competence and job performance [16]. Competent individuals are likely to build better trust relations with colleagues, directly influencing credibility. While there has not been a study exploring the relationship between openness and credibility in the chatbot context, based on prior research in interpersonal contexts, we hypothesize that:

- **H4**. Chatbots high in openness will be perceived as higher in credibility.

Neuroticism contrasts emotional stability and even-temperedness with negative emotionality, such as feeling anxious, nervous, sad, and tense [33]. Individuals with low neuroticism tend to exhibit composure, a trait closely linked to credibility. While neuroticism or emotional stability is a key dimension of the Big Five personality model, the relationship between neuroticism and credibility has received less attention compared to other personality traits. In the design of chatbots for mental health, neuroticism has been specifically excluded from studies as it is deemed undesirable [47]. Given the lack of prior work, we simply pose the following research question:

- **RQ2**. Can perceptions of chatbot's neuroticism affect credibility of information in conversational search?

3 Method

3.1 Procedure

Our study sought to understand how different chatbot personalities can affect the perceived credibility of information during conversational search. We conducted a mixed experimental design with personality trait as a between-subjects factor and personality level as a within-subjects factor.

Participants were randomly assigned to one of five personality conditions (Conscientiousness, Agreeableness, Extraversion, Openness, and Neuroticism). Each participant viewed three conversational search exchanges–covering questions about travel recommendations, cooking recipes, and movie plots (order randomized)–all reflecting their assigned personality trait but at different levels (high, baseline, or low). Participants were told that these exchanges were with three different chatbots designed to answer users' questions, but were not informed about the chatbots' intended personality designs or that all three exchanges reflected the same underlying trait.

After reading each information seeking exchange, participants evaluated the credibility of the information. Following this, participants were asked to assess the personality of the chatbot. Upon completing all three exchanges and their respective evaluations, participants were given an attention check question: *"Which of the following is not a question the user asked during the exchange?"*. Participants who failed to pass this attention check question were later excluded from the data analysis.

Finally, participants provided demographic information, including gender and age. They were also asked about their personal experience with chatbots. This included questions about the types of questions they typically ask chatbots and how frequently they ask different types of questions.

Before conducting our experiment, to provide transparency to our study and ensure reproducibility, we pre-registered our hypotheses on AsPredicted[1]. The overall procedure of our study was reviewed and approved by the university's human subjects division.

[1]https://aspredicted.org/f3m9-ctj9.pdf

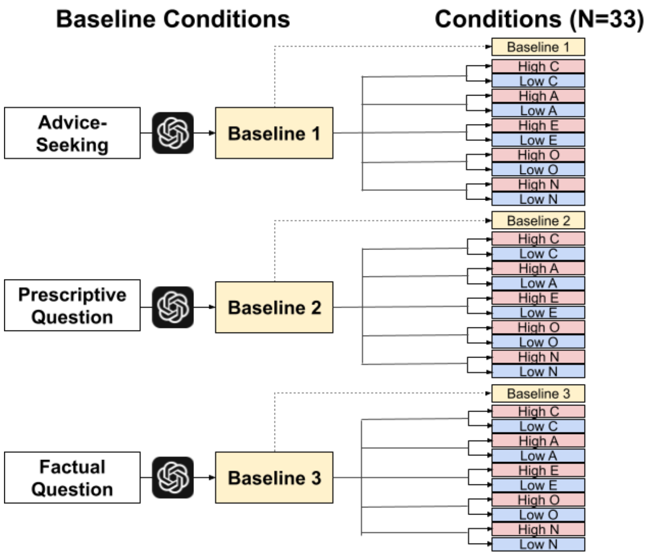

Figure 1: An overview of the process for generating stimulus materials

3.2 Stimulus Materials

We developed our stimulus materials in two stages. First, we created three baseline conversational search exchanges. We then systematically manipulated them to reflect different personality traits and levels. Figure 1 provides an overview of this process.

To create the baseline exchanges, we first selected appropriate question types for conversational search contexts. Prior research on social Q&A sites identified six question types [28], three of which aim to elicit preferences or personal opinions (Identification, (Dis)Approval, and Quality). As our study targets information-seeking scenarios whose primary goal is knowledge acquisition and actionable guidance, rather than subjective taste or personal value judgments, we focus on three question types that are more representative of search tasks: **Advise** ("Directed at generating a new (or specifically tailored) solution, approach, or plan rather than locating or implementing an already existing solution. Grounded in the questioner's desire to inform future action."), **Prescriptive** ("Directed at pursuing an already developed solution to a problem or challenge. Grounded in the questioner's desire to learn steps or strategies that are known (through experience) to address or resolve the issue at hand."), and **Factual** ("Directed at seeking an answer that is objectively or empirically true, such as existing information, data, or settled knowledge.").

For each question type, we designed a question of general interest that would be personally relevant to participants, while not being heavily influenced by their prior knowledge or education level. We developed the following questions, matching each topic with its relevant question type:

- Question 1. (Advice-seeking question) I'm planning a trip to Iceland. Can you advise me on five must-visit landmarks?
- Question 2. (Prescriptive question) How do I make a garlic lemon spaghetti?
- Question 3. (Factual question) What is the plot of Song of the Sea?

We then used GPT-4 to generate responses to these three questions. In the prompt, we specified that responses should not exceed half a page in length to prevent exchanges from being overly long while also ensuring consistency in length across conditions, thereby controlling for the potential influence of text length heuristics on perceived credibility [12]. These three question-response pairs served as our baseline condition.

Next, using the three baseline exchanges as starting points, we systematically created personality-manipulated conditions. Inspired by previous work demonstrating that GPT models can successfully reflect assigned personality traits [32], we prompted GPT-4 to rewrite the response to reflect either high or low levels of one of the Big Five personality traits, while preserving the factual content. Although the Big Five are conceptually independent from each other, most studies suggest that there are small-to-moderate associations between them [19]. We chose to manipulate one trait at a time while allowing the other traits to be implicitly generated by the model, as this approach could more naturally simulate conversational behaviors that reflect the interplay of traits typical of individuals high or low in the manipulated trait.

We used the following prompts to generate the personality variations:

```
[Q1/Q2/Q3 chatbot response] Return this information like a person who has a personality [high/low] in [Conscientiousness/Agreeableness/Extraversion/Openness/Neuroticism].
```

This manipulation created 30 additional exchanges (3 question x 5 personality x 2 levels). Combined with the three baseline exchanges, we used 33 conversational search exchanges in total for our study. The exchanges used in the experiment are provided in the Supplementary Materials.

3.3 Measures

The primary outcome of interest is information credibility. Using scales developed in prior work [18], participants were asked to rate perceptions of the information on 7-point scales ranging from 1 = "not at all"to 7 = "extremely" for believability, accuracy, trustworthiness, bias, and completeness. This credibility measure demonstrated high internal reliability (Cronbach's $\alpha = 0.85$).

For the personality assessment, we use the Ten-Item Personality Inventory (TIPI) [26], which is a brief measurement of one's personality based on the Big Five. Participants were asked to rate the chatbot's personality and their own personality on a 7-point scale (*e.g., "I see the chatbot as extroverted and enthusiastic,"*) where participants rated their agreement with this statement from 1-disagree strongly to 7-agree strongly. Each personality trait is assessed by two questions in the TIPI.

3.4 Recruitment & Participants

We conducted a power analysis using an effect size of 0.30, a significance level of α of 0.05, and a desired power of 0.70. The calculated minimum sample size was $N = 181$. Based on the power analysis [17], we initially recruited a rounded-up number of the minimum required sample size to ensure sufficient data. We recruited 200 participants through an online survey platform, Prolific[2]. Individuals 18 years or older, and fluent in English, were eligible to participate in the survey. For the data analysis, we excluded the responses of participants who failed to pass the attention check question. The final sample consisted of 190 participants (90 female, 95 male, 4 non-binary, 1 preferred not to answer), aged between 18 and 65, with most (86.3%) aged between 18 and 34.

3.5 Data Analysis

3.5.1 Manipulation Check. To ensure that our personality manipulation was effective, we conducted a manipulation check prior to the main analyses. Specifically, we tested whether participants perceived the high-level condition as higher than the baseline, and the baseline higher than the low condition. As the Shapiro-Wilk test indicated that our data did not follow a normal distribution, we conducted post-hoc pairwise comparisons using the Wilcoxon signed-rank test with Bonferroni correction for multiple comparisons.

Our results show that across all dimensions, the high condition indeed resulted in higher rated personality trait compared to the low condition (*Conscientiousness*: $W = 8.50, p < .001$; *Agreeableness*: $W = 12.00, p < .001$; *Extraversion*: $W = 13.50, p < .001$; *Openness*: $W = 17.50, p < .001$; *Neuroticism*: $W = 35.00, p < .001$).

The baseline conditions, however, though was generally rated between the high and low conditions, was sometimes not statistically different from the high and low. In the case of neuroticism, it even resulted in a lower perceived neuroticism compared to the low condition.

3.5.2 Analyses Approaches. Due to our manipulation check showing that the baseline condition was not consistently perceived to be in between the high and low conditions, we chose to first test our results focusing on comparing the high condition from the low condition – the two sets of conditions where we did observe significant differences in perceived personalities. This allows us to explore whether the higher personality trait conditions resulted in a different credibility perception compared to the lower personality trait conditions. For each of the personality traits, we compared the high and low conditions using the Wilcoxon-signed rank test.

In addition to comparing the manipulated conditions directly, we also explored using participants' perceived personality ratings as a predictor of credibility. We employed a linear mixed-effects model using participants' credibility ratings as the dependent variable, and the fixed effects included the five perceived personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), question type, and word count. Participant ID was included as a random intercept to account for individual differences and repeated measures within participants. This model offers several benefits compared to the former approach of simply comparing high and low conditions. First, it uses participants' perceived credibility ratings rather than relying on the manipulations—allowing us to examine the direct effect of perceived credibility. Second, it allows us to test the perceived credibility ratings in a single model. It is possible that when manipulating one personality trait, another personality trait may also have been influenced. This model allows us to statistically test and control for the effects of all personality traits. Additionally, given the potential correlations among the Big Five traits, we assessed multicollinearity among the personality predictors using pairwise correlations and variance inflation factors (VIFs).

However, as we will show, while we employed two different approaches in our analyses, the resulting findings between the two approaches were fairly consistent and shows similar effects.

[2]https://prolific.co

4 Results

4.1 Credibility Differences Across Personality Levels

For *Conscientiousness*, credibility ratings were significantly higher for the High level ($M = 5.57$) than for the Low level ($M = 3.85$) ($W = 0.00, p < .001$). This result confirms H1. Similarly, for *Agreeableness*, the High level ($M = 4.83$) was rated as significantly more credible than the Low level ($M = 3.81$), ($W = 112.00, p < .001$), as predicted in H2.

In contrast, the opposite pattern was found for *Extraversion* and *Neuroticism*. For *Extraversion*, credibility was significantly higher in the Low level ($M = 5.34$) than in the High level ($M = 4.71$), ($W = 117.50, p = .031$), contrary to H3. Similarly, for *Neuroticism*, the Low level ($M = 5.20$) was rated as significantly more credible than the High level ($M = 4.58$), ($W = 166.00, p = .015$). For *Openness*, there was no significant difference between the High ($M = 4.87$) and Low levels ($M = 5.06$), ($W = 260.50, p = .372$), providing no support for H3.

4.2 Predictors of Credibility

Similar to our analyses between high and low conditions using Wilcoxon-signed rank test, our mixed-effect linear regression model testing the perceived personalities as predictors revealed similar effects of personality on credibility. We again found that conscientiousness to positively predict credibility ($\beta = 0.443$, SE = 0.037, $t = 12.116$, $p < 0.001$), confirming H1. Similarly, agreeableness positively influenced credibility ($\beta = 0.120$, SE = 0.035, $t = 3.412$, $p < 0.001$), confirming H2. On the other hand, contrary to H3, extraversion negatively predicted credibility, although the effect was weaker compared to other traits ($\beta = -0.059$, SE = 0.031, $t = -1.885$, $p = 0.06$). Though not hypothesized, neuroticism also had a significant negative effect on credibility ($\beta = -0.185$, SE = 0.043, $t = -4.289$, $p < 0.001$). Finally, we did not find openness to significantly predict credibility (H4 not supported, $p = 0.306$). Multicollinearity diagnostics indicated low correlations among the personality predictors (all VIFs < 2.6), suggesting that the estimated effects were stable.

When examining our control variables, we also saw that the question type (factual) had a significant effect on credibility. Specifically, participants found chatbot responses for factual questions less credible compared to other question types ($\beta = -0.266$, SE = 0.080, $t = -3.317$, $p = 0.001$). The word count or the length of the text did not have an effect ($\beta = 0.001$, SE = 0.001, $t = 1.054$, $p = 0.292$).

Overall, please refer to Table 1 for a comparison of results between the two analyses.

Table 1: Summary of Credibility Ratings and Linear Regression Results by Personality Trait

Traits	Personality Levels Testing (Effect size)	Linear Model Coefficient	Hypotheses
Conscientiousness (H1)	$M_{High} = 5.57 > M_{Low} = 3.85^{***}$ $(r = .87)$	0.443^{***}	Confirmed
Agreeableness (H2)	$M_{High} = 4.83 > M_{Low} = 3.81^{***}$ $(r = .63)$	0.120^{***}	Confirmed
Extraversion (H3)	$M_{Low} = 5.34 > M_{High} = 4.71^{*}$ $(r = .40)$	-0.059	Disconfirmed
Openness (H4)	$M_{High} = 4.87 = M_{Low} = 5.06$ $(r = .15)$	-0.035	No support
Neuroticism	$M_{Low} = 5.20 > M_{High} = 4.58^{*}$ $(r = .41)$	-0.185^{***}	–

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

5 Discussion

Our findings offer empirical evidence showing that personality of the chatbot can significantly influence the perceived credibility of the information it provides. Furthermore, we found that the various chatbot personality traits can have different effects on credibility. Perceived conscientiousness and agreeableness of a chatbot can increase credibility, while perceived neuroticism and extraversion can decrease the credibility of the search results.

Chatbot personalities perceived as conscientious and agreeable positively affect credibility, likely due to underlying psychological mechanisms that align with theories established in individuals' personality and credibility. Conscientiousness, which is characterized by traits such as competence, professionalism, and expertise is often linked to perceptions of responsibility and dependability, as seen in interpersonal interactions [33]. This fosters trust, as users are more inclined to believe that a conscientious chatbot is providing well-considered and accurate information. Our findings align with and extend prior work suggesting that conscientiousness enhances perceived expertise and trustworthiness in both human [6, 58] and chatbot contexts [38, 47]. Similarly, the positive effect of agreeableness on credibility may stem from the fact that agreeableness is associated with sociability, empathy, and warmth, which likely promotes a supportive and pleasant atmosphere during interactions between users and chatbots. This interaction dynamic could drive users to trust and value the information more. Our results support previous studies indicating these mechanisms are at play in chatbot interactions [27], and suggest that carefully designing for these personalities can have tangible effects on the overall efficacy of the conversational search interface.

However, not all our hypotheses were supported by our empirical data. We had hypothesized that extraverted chatbots would be perceived as more credible due to their engaging and relatable communication style, which mirrors how people are often attracted to and enjoy interacting with extraverted individuals in interpersonal settings [45]. On the contrary, our findings revealed the opposite effect – extraversion actually decreased the perception of credibility in chatbots. One possible explanation could be the expectations users hold when interacting with a chatbot versus a human. In human-to-human interactions, extraversion might be linked to warmth, likability, and trustworthiness, enhancing the perception of credibility. However, in human-chatbot interactions, extraversion may not translate identically. The highly social and energetic behaviors typical of an extravert might lead users to question the chatbot's reliability, potentially undermining the chatbot's perceived credibility. However, further investigation is required, given that extraversion has long been included as an important component of credibility explained in interpersonal communication [43].

While we did not hypothesize any relationships between neuroticism and credibility due to the limited prior work, as it is often less considered in chatbot design due to it being perceived as a negative trait [47], as agents are typically designed to be supportive of the user's inquiries. Our empirical data does support this overall negative perception of neuroticism [46] and showed that neuroticism can decrease credibility perceptions. Lastly, in our experimental setting, openness did not significantly predict credibility. Our hypothesized relationship was neither supported nor disconfirmed. More research is needed to confirm the hypothesized link between openness and credibility.

Although not hypothesized, from the main model, we found a significant negative effect of factual questions on credibility. Compared to other question types, factual questions were considered less credible. One potential explanation may be that for the factual questions, people prefer more direct and succinct responses, as suggested in prior work [1]. Unlike advice-seeking or prescriptive questions, which might allow for more subjective or personalized responses in participants' expectations, factual questions may likely expose gaps between expectations and the response quality. Any perceived inaccuracies, lack of detail, or overly generic responses may have disproportionately reduced credibility for this question type. These factors could collectively explain why credibility was rated lower for factual questions. However, more research will be needed to explore this further.

5.1 Implications for Credible Conversational Search Interfaces

Given the importance of credibility in the information search context, our findings and the GPT model manipulations provide several design considerations for developing chatbots in conversational search, and chatbot design in general.

Our comparison between high- and low-conscientious chatbots revealed that the high-conscientious chatbot provided longer, more detailed responses, with an average word count of 208 compared to 141 for the low-conscientious chatbot. Upon closer examination, this difference was not due to verbosity but because the conscientious agent provided more information in their response. For instance, when providing a recipe, it offered alternative ingredient options, such as *"8 oz spaghetti (whole wheat for a healthier option),"* while this additional information was not observed in any other responses. Additionally, the high-conscientious chatbot presented information in a clear and organized manner. In the recipe example, it gave precise instructions with exact measurements and explicit timings for each step. This contrasts with the low-conscientious response, which provided more casual, approximate instructions, such as, *"Garlic, chop up a few cloves"* and *"Boil it in salty water 'till it's not crunchy"*, without much attention to detail. Furthermore, conscientious chatbots offered proactive guidance by anticipating user needs, such as reminding users to reserve tours in advance. This guidance was absent in the low-conscientious responses. In addition, the high-conscientious chatbot focused solely on the information, avoiding unnecessary remarks or emotional responses. Unlike other exchanges that started with an emotional response or a welcoming comment, the conscientious response would start with phrases like *"Certainly"* or *"Absolutely"* directly addressing the user's query without adding extraneous comments. These distinctions highlight how designers may enhanced the perceived conscientiousness of the chatbot's, by providing richer, more structured, and informative responses.

To design for credible conversational interface, we may also consider imbuing agents with agreeable personalities. Based on the generative models' manipulations, this can be achieved by using positive affirmations, empathetic language, and encouraging statements in the responses. For instance, when responding to a query about a movie plot, an agreeable chatbot provides a supportive and enthusiastic view of the movie. The chatbot uses affirming phrases such as, *"It's wonderful how the film weaves in all these elements from Irish mythology".* These expressions not only deliver the necessary information but validate the user's emotions, creating a sense of connection and positive reinforcement. In contrast, a disagreeable chatbot tends to take a more detached and skeptical tone in the response. Instead of affirming the user's perspective, the response includes statements such as, *"The whole thing's a lesson in why it's not a waste to care about old myths."* These responses may downplay the emotional depth of the film, leading to search results that discourages the user. Thus, implementing prosocial and positive responses can help users perceive the chatbot response as more credible.

As for neuroticism, we observed that the highly neurotic chatbots frequently added questions such as *"okay?"*, *"you know?"*, or *"right?"*, which may have reduced users' perceived confidence in the chatbot's responses. Additionally, we noticed that the neurotic chatbots would express concerns when, for example, providing a recipe, adding unnecessary worries at each ingredient and step, further generating doubtful responses. These behaviors stand in contrast to emotionally stable chatbots, which used more assertive language when providing responses. To maintain credibility, we recommend avoiding language markers that suggest uncertainty or concern, ensuring that chatbots communicate with confidence, emotional stability, and clarity throughout their interactions.

While focused on the conversational search context, our findings should also generalize to design of chatbots broadly. But this also raises the importance to acknowledge the potential misused of our findings–chatbot personalities may be manipulated to increase people's credibility perceptions when the information is of low integrity. It may be useful to integrate our findings into digital literacy efforts. By becoming more informed about these design strategies, users may more critically assess the trustworthiness of information provided by such systems and navigate their interactions more effectively.

5.2 Limitations and Future Work

While our study provides empirical insights into the effects of interface design on perceived information credibility, limitations and unanswered questions remain for future research. One key limitation is that to maximize experimental control, we chose to present screenshots of question and answer interactions, as opposed to allow participants to interact with the conversational searches directly. This may have undermined the ecological validity of our work and future work should extend our study to more interactive settings.

Our study design could also be extended to different question domains. While we focused on question types related to everyday life, future research could explore high-stakes scenarios or topics requiring specialized knowledge, such as medical inquiries or financial advice. Furthermore, our findings suggest that question type significantly influences perceived credibility. Would these results differ if factual questions pertained to different topics or domains? Investigating this possibility could expand our findings in specific search contexts.

6 Conclusion

In this paper, we explored the impact of interface design on the efficacy of conversational search engines, focusing on chatbot personalities and credibility. To investigate this, we conducted an online experiment with 190 participants, manipulating chatbot dialogues based on Big Five personality traits. Our findings revealed that chatbots designed with personalities high in conscientiousness and agreeableness, and low in extraversion and neuroticism, were perceived as more credible. These results offer empirical evidence contributing to a more nuanced understanding of how personality-driven design can affect conversational search results. Also, our work demonstrates the feasibility of using generative AI to create tailored interface and offers practical guidelines for designing more credible conversational search systems.

Acknowledgments

This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT)(IITP-2026-RS-2020-II201460).

References

[1] Lada A Adamic, Jun Zhang, Eytan Bakshy, and Mark S Ackerman. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*. 665–674.

[2] Rangina Ahmad, Dominik Siemon, and Susanne Robra-Bissantz. 2021. Communicating with machines: Conversational agents with personality and the role of extraversion. (2021).

[3] Jan Allbeck and Norman Badler. 2002. Toward representing agent behaviors modified by personality and emotion. *Embodied conversational agents at AAMAS* 2, 6 (2002), 15–19.

[4] Sandeep Avula and Jaime Arguello. 2020. Wizard of oz interface to study system initiative for conversational search. In *Proceedings of the 2020 conference on human information interaction and retrieval*. 447–451.

[5] Michael J Baker and Gilbert A Churchill Jr. 1977. The impact of physically attractive models on advertising evaluations. *Journal of Marketing research* 14, 4 (1977), 538–555.

[6] Murray R Barrick and Michael K Mount. 2012. Select on conscientiousness and emotional stability. *Handbook of principles of organizational behavior: Indispensable knowledge for evidence-based management* (2012), 19–39.

[7] Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 396–403.

[8] Timothy Bickmore and Justine Cassell. 2005. Social dialongue with embodied conversational agents. *Advances in natural multimodal dialogue systems* (2005), 23–54.

[9] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–11.

[10] Judee K Burgoon, Joseph A Bonito, Bjorn Bengtsson, Carl Cederberg, Magnus Lundeberg, and Lisa Allspach. 2000. Interactivity in human–computer interaction: A study of credibility, understanding, and influence. *Computers in human behavior* 16, 6 (2000), 553–574.

[11] Pew Research Center. 2023. A majority of Americans have heard of ChatGPT, but few have tried it themselves. https://www.pewresearch.org/short-reads/2023/05/24/a-majority-of-americans-have-heard-of-chatgpt-but-few-have-tried-it-themselves/. Accessed: 2024-07-15.

[12] Shelly Chaiken. 2014. The heuristic model of persuasion. In *Social influence*. Psychology Press, 3–39.

[13] Tomas Chamorro-Premuzic and Adrian Furnham. 2009. Mainly Openness: The relationship between the Big Five personality traits and learning approaches. *Learning and individual Differences* 19, 4 (2009), 524–529.

[14] Avishek Choudhury and Hamid Shamszare. 2023. Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *Journal of Medical Internet Research* 25 (2023), e47184.

[15] Robert J Cramer, Stanley L Brodsky, and Jamie DeCoster. 2009. Expert witness confidence and juror personality: Their impact on credibility and persuasion in the courtroom. *Journal of the American Academy of Psychiatry and the Law Online* 37, 1 (2009), 63–74.

[16] Stephen J Dollinger and Lisa A Orf. 1991. Personality and performance in "personality": Conscientiousness and openness. *Journal of Research in Personality* 25, 3 (1991), 276–284.

[17] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[18] Andrew J Flanagin and Miriam J Metzger. 2000. Perceptions of Internet information credibility. *Journalism & mass communication quarterly* 77, 3 (2000), 515–540.

[19] William Fleeson and Joshua Wilt. 2010. The relevance of Big Five trait content in behavior to subjective authenticity: Do high levels of within-person behavioral variability undermine or enable authenticity achievement? *Journal of Personality* 78, 4 (2010), 1353–1382.

[20] Brian J Fogg, Jonathan Marshall, Tami Kameda, Joshua Solomon, Akshay Rangnekar, John Boyd, and Bonny Brown. 2001. Web credibility research: A method for online experiments and early study results. In *CHI'01 extended abstracts on Human factors in computing systems*. 295–296.

[21] R Barry Fulton. 1970. The measurement of speaker credibility. *Journal of Communication* 20, 3 (1970), 270–279.

[22] Souvick Ghosh. 2019. Investigating result presentation in conversational ir. In *Proceedings of the 2019 conference on human information interaction and retrieval*. 421–424.

[23] Alexandru L Ginsca, Adrian Popescu, Mihai Lupu, et al. 2015. Credibility in information retrieval. *foundations and trends® in information retrieval* 9, 5 (2015), 355–475.

[24] Ulrich Gnewuch, Meng Yu, and Alexander Maedche. 2020. The effect of perceived similarity in dominance on customer self-disclosure to chatbots in conversational commerce. (2020).

[25] Lewis R Goldberg. 2013. An alternative "description of personality": The Big-Five factor structure. In *Personality and Personality Disorders*. Routledge, 34–47.

[26] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[27] William G Graziano, Meara M Habashi, Brad E Sheese, and Renée M Tobin. 2007. Agreeableness, empathy, and helping: a person× situation perspective. *Journal of personality and social psychology* 93, 4 (2007), 583.

[28] F Maxwell Harper, Joseph Weinberg, John Logie, and Joseph A Konstan. 2010. Question types in social Q&A sites. *First Monday* (2010).

[29] Christian Hildebrand and Anouk Bergner. 2019. AI-driven sales automation: Using chatbots to boost sales. *NIM Marketing Intelligence Review* 11, 2 (2019), 36–41.

[30] Carl Iver Hovland, Irving Lester Janis, and Harold H Kelley. 1953. Communication and persuasion. (1953).

[31] Carl I Hovland and Walter Weiss. 1951. The influence of source credibility on communication effectiveness. *Public opinion quarterly* 15, 4 (1951), 635–650.

[32] Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547* (2023).

[33] Oliver P John, Sanjay Srivastava, et al. 1999. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. (1999).

[34] Yvonne Kammerer and Peter Gerjets. 2012. Chapter 10 how search engine users evaluate and select web search results: The impact of the search engine interface on credibility assessments. In *Web search engine research*. Emerald Group Publishing Limited, 251–279.

[35] Simona C Kaplan, Cheri A Levinson, Thomas L Rodebaugh, Andrew Menatti, and Justin W Weeks. 2015. Social anxiety and the big five personality traits: The interactive relationship of trust and openness. *Cognitive behaviour therapy* 44, 3 (2015), 212–222.

[36] Abhishek Kaushik, Vishal Bhat Ramachandra, and Gareth JF Jones. 2020. An interface for agent supported conversational search. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 452–456.

[37] Johannes Maria Kraus, Florian Nothdurft, Philipp Hock, David Scholz, Wolfgang Minker, and Martin Baumann. 2016. Human after all: Effects of mere presence and social interaction of a humanoid robot as a co-driver in automated driving. In *Adjunct proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications*. 129–134.

[38] Mohammad Amin Kuhail, Justin Thomas, Salwa Alramlawi, Syed Jawad Hussain Shah, and Erik Thornquist. 2022. Interacting with a chatbot-based advising system: Understanding the effect of chatbot personality and user gender on behavior. In *Informatics*, Vol. 9. MDPI, 81.

[39] Shiri Lev-Ari and Boaz Keysar. 2010. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of experimental social psychology* 46, 6 (2010), 1093–1096.

[40] Monika Lohani, Charlene Stokes, Marissa McCoy, Christopher A Bailey, and Susan E Rivers. 2016. Social interaction moderates human-robot trust-reliance relationship and improves stress coping. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 471–472.

[41] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research* 30 (2007), 457–500.

[42] Robert R McCrae and David M Greenberg. 2014. Openness to experience. *The Wiley handbook of genius* (2014), 222–243.

[43] James C McCroskey and Thomas J Young. 1981. Ethos and credibility: The construct and its measurement after three decades. *Communication Studies* 32, 1 (1981), 24–34.

[44] William J McGuire. 1985. Chapter attitudes and attitude change. *Handbook of social psychology* (1985), 233–346.

[45] Brinda Mehra. 2021. Chatbot personality preferences in Global South urban English speakers. *Social Sciences & Humanities Open* 3, 1 (2021), 100131.

[46] Jingbo Meng and Yue Dai. 2021. Emotional support from AI chatbots: Should a supportive partner self-disclose or not? *Journal of Computer-Mediated Communication* 26, 4 (2021), 207–222.

[47] Joonas Moilanen, Aku Visuri, Sharadhi Alape Suryanarayana, Andy Alorwu, Koji Yatani, and Simo Hosio. 2022. Measuring the effect of mental health chatbot personality on user engagement. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia*. 138–150.

[48] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied* 7, 3 (2001), 171.

[49] Robert N Oddy. 1977. Information retrieval through man-machine dialogue. *Journal of documentation* 33, 1 (1977), 1–14.

[50] Charles A O'Reilly and Karlene H Roberts. 1976. Relationships among components of credibility and communication behaviors in work units. *Journal of Applied Psychology* 61, 1 (1976), 99.

[51] Babajide Osatuyi. 2013. Information sharing on social media sites. *Computers in human behavior* 29, 6 (2013), 2622–2631.

[52] John K Rempel, John G Holmes, and Mark P Zanna. 1985. Trust in close relationships. *Journal of personality and social psychology* 49, 1 (1985), 95.

[53] Soo Young Rieh and David R Danielson. 2007. Credibility: A multidisciplinary framework. (2007).

[54] David Robins and Jason Holmes. 2008. Aesthetics and credibility in web site design. *Information Processing & Management* 44, 1 (2008), 386–399.

[55] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[56] Michael Shumanov and Lester Johnson. 2021. Making conversations with chatbots more personalized. *Computers in Human Behavior* 117 (2021), 106627.
[57] Richard Sutcliffe. 2023. A Survey of Personality, Persona, and Profile in Conversational Agents and Chatbots. *arXiv preprint arXiv:2401.00609* (2023).
[58] Ulrich Trautwein, Oliver Lüdtke, Brent W Roberts, Inge Schnyder, and Alois Niggli. 2009. Different forces, same consequence: conscientiousness and competence beliefs are independent predictors of academic effort and achievement. *Journal of personality and social psychology* 97, 6 (2009), 1115.
[59] Shawn Tseng and Brian J Fogg. 1999. Credibility and computing technology. *Commun. ACM* 42, 5 (1999), 39–44.
[60] Sarah Theres Völkel and Lale Kaya. 2021. Examining user preference for agreeableness in chatbots. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–6.
[61] Sarah Theres Völkel, Samantha Meindl, and Heinrich Hussmann. 2021. Manipulating and evaluating levels of personality perceptions of voice assistants through enactment-based dialogue design. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–12.
[62] Sarah Theres Völkel, Ramona Schoedel, Lale Kaya, and Sven Mayer. 2022. User perceptions of extraversion in chatbots after repeated use. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
[63] H Von der Ohe, N Martins, and M Roode. 2004. The influence of credibility on employer-employee trust relations. *South African Journal of Labour Relations* 28, 2 (2004), 4–32.
[64] C Nadine Wathen and Jacquelyn Burkell. 2002. Believe it or not: Factors influencing credibility on the Web. *Journal of the American society for information science and technology* 53, 2 (2002), 134–144.
[65] Yusuke Yamamoto and Katsumi Tanaka. 2011. Enhancing credibility judgment of web search results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1235–1244.
[66] Michelle X Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 2-3 (2019), 1–36.