"I Think You Might Like This": Exploring Effects of Confidence Signal Patterns on Trust in and Reliance on Conversational Recommender Systems

MARISSA RADENSKY*, University of Washington, USA
JULIE ANNE SÉGUIN, Google, USA
JANG SOO LIM†, Creative Circle, USA
KRISTEN OLSON, Google, UK
ROBERT GEIGER, Google, USA

	Low-Confidence Recommendations (#1-6)	Medium-Confidence Recommendations (#7-12)	High-Confidence Recommendations (#13-18)
Dynamic-Accurate	I think you might like this song.	I think you would like this song.	I think you will like this song.
Dynamic-Random	I think you might/would/will like this song.	I think you might/would/will like this song.	I think you might/would/will like this song.
Static-Low	I think you might like this song.	I think you might like this song.	I think you might like this song.
Static-High	I think you will like this song.	I think you will like this song.	I think you will like this song.
None (Baseline)	*No confidence signal*	*No confidence signal*	*No confidence signal*

Fig. 1. Left: Wizard of Oz prototype. Right: confidence signal pattern conditions over the three batches of recommendations.

With the rapid growth of large language models, conversational recommender systems (CRSs) are on the rise. In receiving a CRS recommendation, one may encounter confidence signals of varying levels such as "you might like…" or "I think you will like…," but to the best of our knowledge, no work has investigated how the pattern of confidence signals from one recommendation to the next affects trust in and reliance on CRSs. In a mixed-methods user study, we explore how 30 participants interact with two Wizard of Oz music CRSs that grow increasingly confident, but only one communicates a confidence (accurate, random, always low, or always high) using natural language and color-coding. Through semi-structured interviews, survey responses, and recommendation ratings, we find evidence suggesting that an accurate confidence signal generates the greatest increase in trust-related metrics without encouraging over-reliance but potentially under-reliance. We also identify design guidelines for CRS confidence signals associated with each trust-related metric: desire to use, perceived transparency, perceived ability, perceived benevolence, and perceived anthropomorphism.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Recommender systems**.

Additional Key Words and Phrases: human-AI interaction, conversational recommender systems, trust, reliance, AI confidence

1 INTRODUCTION AND RELATED WORK

Conversational recommender systems (CRSs) are on the rise, as users increasingly ask conversational AI systems for recommendations spanning everything from movies to travel [17]. With the rapid growth in large language models such

*Work done as student researcher at Google.
†Work done at Google via Creative Circle.

as GPT-3 [13] and LaMDA [31], interest in conversational AI will likely only grow. We adopt Jannach et al.'s definition of a CRS: "*a software system that supports its users in achieving recommendation-related goals through a multi-turn dialogue*" [17]. Given the widespread use of CRSs, it is important to determine how best to support users in developing trust in and appropriate reliance on CRSs. Trust in automation was defined by Lee and See as "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability*" [21], while appropriate reliance refers to the ability to know when and when not to go along with a system's suggestions [21].

We explore how different confidence signal patterns, communicated through a combination of natural language and color-coding, impact trust in and reliance on CRSs. A number of works have investigated whether or not communicating confidence in **recommender systems** may improve trust and appropriate reliance. Although these works have had conflicting results [10, 15, 25, 30, 32], none were in the context of a *conversational* recommender. A CRS may more naturally incorporate confidence (using natural language or other means) because a conversation is by definition expected to involve back-and-forth information exchange. Moreover, there has previously been little investigation of how the *pattern* of confidence conveyed from recommendation to recommendation impacts users' trust and reliance. Prior works have also explored how **natural-language confidence** may impact trust and reliance but not in the context of a *recommender* [19, 28, 37]. With the personal evaluation of recommendations, natural-language confidence may differently influence participants' trust and reliance. Furthermore, unlike other works with natural-language confidence, this work uses color-coding to highlight confidence changes, aligning with Dubiel et al.'s design guideline for conversational agent trust calibration, which suggests using multiple modalities to facilitate output evaluation [12].

We use semi-structured interviews, survey responses, and recommendation ratings to explore how a CRS' confidence signal pattern impacts different **trust-related metrics** (**RQ1**) and **reliance** (**RQ2**). **Our contributions are:** 1) a mixed-methods Wizard of Oz [11] user study (N=30) investigating the impact of different confidence signal patterns (accurate, random, always high, always low, none) on trust in and reliance on a music CRS, 2) evidence suggesting that an accurate pattern leads to the greatest increase in trust-related metrics, without encouraging over-reliance but potentially under-reliance, and 3) design guidelines for CRS confidence signals associated with each trust-related metric.

2 STUDY DESIGN

2.1 Participants and Condition Design

Thirty English-speaking adults (17 women, 12 men, 1 non-binary) were recruited from a US-wide participant pool by L&E Research and received $75. Participants did not work in human-computer interaction, linguistics, technology, psychology, marketing, or the government. All stream music, listen to music at least a few times per week, and seek out music for playlists at least a few times per month. Each participant received the baseline and one other condition. A summary of the wording and color-coding of confidence signals for each condition is shown in Figure 1. The accurate confidence signal was the only one aligned with actual recommendation confidence. The confidence signals are partially expressed through words, as in prior work [28, 37]. Based on Hyland et al.'s epistemic categorization of various words related to certainty, we use the possibility term "might" to indicate low confidence, the probability term "would" to indicate medium confidence, and the certainty term "will" to indicate high confidence [16]. We ran two Google Surveys to confirm this semantic hierarchy (without any color-coding). Comparing "might" and "would", 59.5% of 401 respondents thought that "would" was more confident, and comparing "would" and "will", 57.7% of 403 respondents thought that "will" was more confident. These results are weighted according to age, gender, and geography for the United States' internet population. Though the surveys reinforced the established confidence hierarchy, the differences were somewhat

narrow. In pilot study sessions, we also found that participants often did not notice the word-based confidence levels. We thus decided to add traffic-light color-coding to the three levels of confidence: red for low confidence, yellow for medium, and green for high [23]. The specific color hues were selected to be accessible for people who are color-blind.

We used YouTube Music to obtain personalized recommendations of low, medium, and high confidence for each participant prior to their study session. In their screening, participants selected the most common situation in which they would listen to a playlist for 30 minutes or more. They also provided one genre, three artists, and two songs per artist that they would want to hear in that situation. For the low-confidence songs, we used songs recommended before any interaction with the system. In order for these songs to be the same for all participants, participants needed to be around the same age because a birth date was required to generate recommendations and could influence recommendations. Thus, we recruited participants aged 18 to 24 and set the birth date to the average birth date of their age group. For the medium-confidence songs, we utilized participants' preferred genre, and for the high-confidence songs, we utilized their preferred artists and songs. For more detail on the recommendation selection and how they were rated, see Appendix A.

2.2 Procedure

Each one-hour study session was conducted and recorded over Google Meet. Participants were first reminded of the consent form's contents and guided through a tutorial. They were told that they may or may not receive a comment from the system alongside recommendations. Participants assigned to the accurate or random condition were also shown the three-level confidence scale associated with the comment. Participants were not informed as to whether or not the system would update based on feedback. They then received the baseline and one other condition in randomized order. The non-baseline condition's confidence was always high (5 participants), always low (5 participants), accurate (10 participants), or random (10 participants). The Wizard of Oz CRS prototype for each condition was presented through Google Slides. Each condition consisted of 18 recommendation chats. In every condition, the actual confidence was low for the first 6 chats, medium for the next 6, and high for the last 6. (Due to time constraints, P2 skipped two chats per batch under the baseline.) For each recommendation, a chat within a smartphone outline appeared (Figure 1). The chat always started with a pre-filled request from the participant: "What's a good song to add to my [situation provided in screening] playlist?" The baseline prototype would then respond with a recommendation, including its title, artist, and YouTube video. In the other conditions, the prototype would first also provide a confidence signal. Once the recommendation appeared, the first 15 seconds of the song played. Participants then rated their agreement with the 7-point Likert-type statement "I would enjoy listening to this song while [in Situation X]." Next, the participant rated the song's familiarity. Their options were "not at all familiar," "somewhat familiar," and "very familiar." If the song was not very familiar, they heard more and, after hearing at least 15 and up to 45 more seconds, provided a final 7-point Likert-type recommendation rating. Participants listened to at least 30 seconds before providing their final rating, as a review of Spotify music listeners showed that most users decide whether or not to skip a song within this time [20]. We did not collect final ratings for songs with which participants were very familiar, as prior work indicated that results for a movie recommendation explanation were substantially different when users had already seen the movie [3].

After the ninth and eighteenth recommendations in each condition, participants briefly provided thoughts on their experience with the recommender. Once they had provided thoughts following the eighteenth recommendation, they were asked to answer 6 survey questions adapted from prior work that correspond to different metrics that may affect users' trust (Table 1). Mayer et al.'s ABI trust framework indicates that benevolence and ability are important factors for establishing trust [24]. Though it is a significant area of interest in conversational AI [1, 2, 29], anthropomorphism has had a more complex relationship with trust in AI, as it can sometimes give rise to creepiness or disappointment when

the system is not as capable as it seemed [14]. Increasing the transparency of AI systems has had mixed results with respect to trust too, sometimes leading to inappropriate trust [4, 18, 27]. As they completed the survey questions, a list of their recommendations and any associated confidence statements were provided for reference. Upon completing both conditions, participants engaged in a semi-structured interview. The interview questions (Appendix B) focused on their experience with the confidence signals provided in the non-baseline condition and what they thought of other potential scenarios involving confidence signals, including other non-baseline conditions that they did not encounter.

Table 1. 7-point Likert-type survey questions for different trust-related metrics, with sources from which questions were adapted.

Trust-Related Metric	Question	Source
Overall Trust (Desire to Use)	I would use this music recommender again.	[35]
Perceived Benevolence	The recommender seemed sincere during our interaction.	[34]
Perceived Ability	I believe the recommender is competent.	[34]
Perceived Anthropomorphism	I felt a sense of human contact when interacting with the recommender.	[34]
Perceived Transparency I	The song recommendations made sense to me.	[26]
Perceived Transparency II	I understood what the recommendations were based on.	[3]

3 RESULTS

The semi-structured **interviews** were open coded using inductive thematic analysis [6]. The first author generated a codebook based on a review of the recordings. The first and fourth author then coded 4 transcripts, and the first author iterated on the codebook based on any disagreements, with input from the second author. The first author used the final codebook (Appendix C) to code the transcripts. Quotes are attributed to participants based on their ID number and non-baseline condition (A: accurate, L: low, H: high, R: random). To analyze the **trust** results, for each 7-point Likert-type question corresponding to a trust-related metric (Table 1), we subtracted the participant's baseline rating from their non-baseline rating to observe their change in response for that metric (Figure 2). We measure **reliance** on a recommendation by taking its final rating as ground truth and comparing that to its initial rating, as in prior work [5, 7]. To determine whether a participant over- or under-relied on the system under a particular condition, we first obtained the average difference between their final and initial ratings for recommendations in the low-confidence batch. We did the same for the medium- and high-confidence batches. Then, we calculated the average of those averages to determine the average recommendation rating update for that condition. We plotted the baseline and non-baseline average rating updates for each participant in Figure 3. Since any very familiar recommendations did not have final ratings, if a participant encountered a batch of recommendations in which all the songs were very familiar to them, we did not include them in this analysis. Still, for each condition, at least 60% of the participants were included.

3.1 RQ1 - Desire to Use Results

For desire to use, the accurate condition had the only positive median (1.00), the highest first quartile (-0.50), and the highest third quartile (1.75) (Figure 2a). Its median was also equal to or above the third quartile of the other non-baseline conditions. Meanwhile, the high condition had the worst performance with the lowest median shared with the low condition (-1.00), third quartile (0.00), and first quartile shared with the low and random conditions (-2.00).

Twenty-four of the 30 participants expressed interest in having some version of a dynamic confidence signal. P23-A shared, "*I do like the red, yellow, and what will eventually be green.... I like this [recommender] more than the last one, because I feel like it's learning I guess and more tailored.*" It should be noted that 5 of the participants in favor of confidence

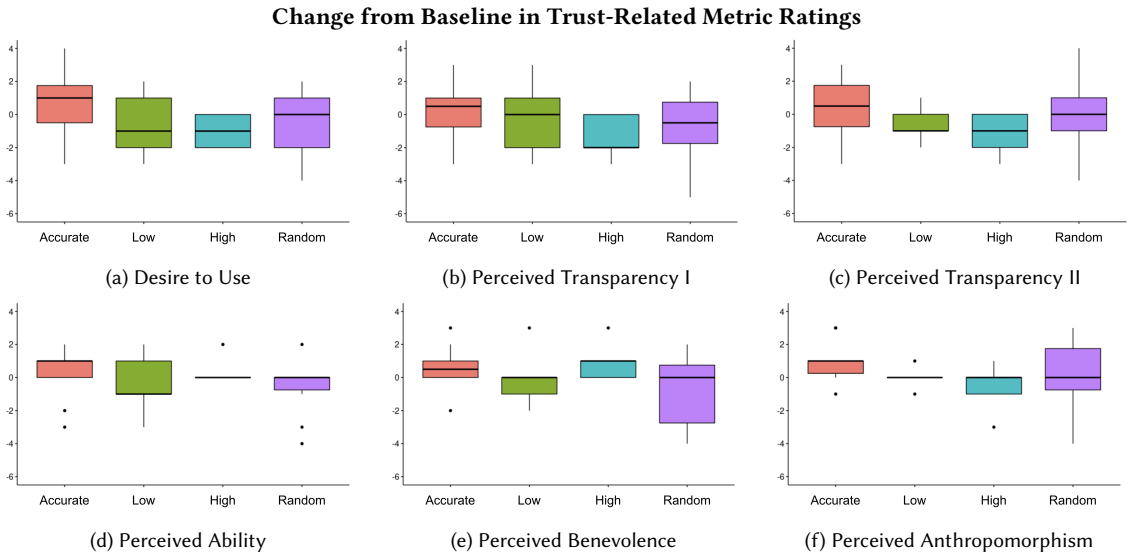


Fig. 2. For each trust-related metric and confidence signal pattern, the change in 7-point Likert-type rating from the baseline.

levels desired occasional rather than constant confidence signals. Over the course of the interviews, we realized that the idea of occasional confidence signals may be interesting to investigate, so we asked 22 participants for their thoughts on such signals for a category of recommendations, such as exploratory or high-confidence recommendations. P28-R said of an occasional high-confidence signal, "*I like that because I would expect that it would only send that when it's very sure that you would like it.*" Discussing an occasional exploratory confidence signal, P26-H commented, "*I think that every once in a while if it checked in on you and then used that feature in order to build your music profile for future song recommendations, I think that that would be a good idea.*" We also note that 4 of the participants who showed interest in confidence signals indicated that they would also be okay without them. Discussed in the next sections, the factors that may have contributed to the conditions' disparate effects on desire to use the system include perceived transparency, ability, benevolence, and anthropomorphism. Of the 6 participants who did not express interest in non-static confidence signals, one did not get a chance to voice their opinion. The rest preferred only to know the reason (e.g.,"because you like Artist X") behind a recommendation [33]. P4-R asserted, "*If it's not going to tell me why, then…I don't need a blurb.*" This point arose because we asked 17 participants whether confidence, reasons, or both would be most favorable, after realizing it was an interesting point to explore. Twelve preferred the combination. P30-L observed, "*I feel like having both is helpful, and it would make me more excited to listen to the recommendations, getting more of a full feedback.*"

Design Guideline 1: Limited wording changes may not be sufficient to convey confidence changes, so consider additional means of confidence communication such as color-coding. Though the tutorial explained the confidence levels, 4 random-condition and two accurate-condition participants reported not noticing or paying attention to the confidence levels. Four of them did not like the idea of using a full confidence scale, so it may be that they simply did not find it useful. That said, a solution for increasing awareness of confidence levels may be to emphasize their color-coding. Five participants who received the accurate or random condition noted that the color-coding was more helpful to their understanding than the wording changes. P18-A revealed, "*I think it was more about the color than it was about this statement. What were actually the three different [statements]…?*" However, if the wording changes had

been more drastic, they may have had more of an effect. P8-A noted, "*I don't really notice the changing in the words at all because it's one word changing.... When it says 'I think' it's already implying that it's not 100% sure.*"

3.2 RQ1 - Perceived Transparency Results

For perceived transparency, the accurate condition performed best (Figures 2b and 2c). For perceived transparency I, it had the only positive median (0.50), shared the highest third quartile with the low condition (1.00), and had the highest first quartile (-0.75). For perceived transparency II, it had the only positive median (0.50), the highest third quartile (1.75), and the highest first quartile (-0.75). However, its median was below the third quartile of the random condition for both metrics and below the third quartile of the low condition for the first metric. Meanwhile, the poorest performance came from the high condition. For perceived transparency I, it had the lowest median (-2.00), third quartile (0.00), and first quartile shared with the low condition (-2.00). For perceived transparency II, it had the lowest median shared with the low condition (-1.00), third quartile shared with the low condition (0.00), and first quartile (-2.00).

Design Guideline 2: Consider both quality and novelty expectations when designing confidence levels. Of the 24 participants who desired some version of confidence levels, 15 expressed that the confidence levels did or would impact their expectations. Six of the 15 had received and noted a dynamic confidence signal. Intuitively, participants generally expected that recommendations associated with higher confidence would have higher quality than those associated with lower confidence. To illustrate, when P18-A encountered a high-confidence signal, they "*felt there was a high probability that I was going to add [the] song to my playlist or that it would be something that I would at least like...*." Describing their experience with a medium-confidence signal, P13-A said, "*... that was a good job because I do like the song but... just not for that specific playlist. So I thought it was like a forewarning.*" Discussing their expectations for recommendations with low confidence, P29-A relayed, "*I felt like it would be more random, and it kind of was random.*" Because participants in the random and static conditions were not provided confidence signals that help set expectations, it makes sense that those conditions (particularly the high one which could not signal a bad recommendation) fared worse in terms of perceived transparency. Still, three participants noted that a static confidence signal gave them higher expectations than no signal or that they would be more willing to listen to a recommendation with a confidence signal.

Five participants assumed that recommendations with lower confidence signals should more likely be new to them. When faced with a low-confidence recommendation, P22-R "*figured I might not be as familiar so it might be a newer song to try.*" This may be an important expectation to factor into providing low-confidence signals, which may not always indicate novelty but rather, for example, generic popularity. For those who do not automatically assume that low confidence implies novelty, low-confidence recommendations may elicit reduced interest from users, as noted in prior work [32]. P26-H raised this concern: "*I think [a confidence scale is] probably a good idea, but you have to wonder at that point what the usefulness is of it in the first place? Like why would it even recommend me a song that's like 'I think you might like the song.'* " On a related note, two participants stated that confidence levels would not adjust their expectations much. P3-A explained, "*... If the confidence level is high, I expect it to mesh well, but [if] the confidence level is low... you know, either way, I expect it to mesh well with the rest of the playlist.*" In order to help users understand the benefits of low-confidence recommendations, it may be helpful to lean into setting a novelty expectation. By framing low-confidence recommendations as opportunities for exploring the users' preferences, users may be more willing to engage with such recommendations. Though P9-R did not think the presented confidence levels would affect their expectations, when they heard the idea of an occasional exploratory confidence signal, they reacted more positively: "*... if you begin it with 'hey, you could hate this, but I just want to see what you think,' I feel like that's better than just 'you might like this' because... there was a few on there that said, 'I think you might like the song,' and I absolutely would not.*' "

3.3 RQ1 - Perceived Ability Results

For perceived ability, the accurate condition had the only positive median (1.00), shared the highest third quartile (1.00) with the low condition, and shared the highest first quartile (0.00) with the high condition (Figure 2d). Also, its median was equal to or greater than the third quartile of the other non-baseline conditions. None of the remaining conditions performed worse overall than the others, considering their medians, first quartiles, and third quartiles.

Design Guideline 3: Differentiate between confidence and reliability signals. Four participants, three of whom had received a static confidence signal, mentioned that confidence levels would be useful for recognizing the system's current ability in terms of how well it understood them. Participants who received a static condition did not have the opportunity to experience this benefit, while the participants who received the random condition were unable to gain proper insight into the system's improving ability. For example, P17-H stated, "*I think [having confidence levels] kind of gives you a better idea of where it's at, especially because it's always a learning process for something like this to get to know you and your taste…But I think once it was something I was very familiar with, it wouldn't really matter.*" The second half of P17-H's statement illustrates an issue with viewing the confidence signal as a reliability metric. Reliability is similar to confidence in that they both describe how well a system is likely to perform. However, whereas confidence communicates the predicted accuracy of an individual output, reliability communicates the system's overall predicted performance [9, 22, 36]. As a recommender improves in understanding a user (which this study mimicked), its confidence signals could resemble a reliability metric, as they should give an indication of how much the system has learned. However, the user or system may regularly explore new potential areas of interest. The system would not be confident about recommendations in these new areas, but the system's reliability would not necessarily have decreased.

After determining it would be interesting to investigate, we asked 12 participants what they would expect or prefer the confidence trajectory to look like and gave various examples such as the confidence increasing or oscillating. The largest plurality (5) of those participants expected or preferred the confidence to increase over time. P24-R explained, "*I would prefer it to be…a higher confidence as you listen to more music.… having it go on-and-off I feel like doesn't really do anything, like you could listen to a radio and it could do that too.…*" The next most common response (three participants) was an expectation or preference that the confidence would generally increase but occasionally be lower when providing exploratory recommendations. It is possible that, had we always directly proposed the idea of such a confidence trajectory, which we did with the last participant, more participants would have agreed that that made sense. However, because it was not their initial assumption, it appears important to communicate what a confidence signal means and how it is different from a reliability metric in order for users to more appropriately utilize it.

3.4 RQ1 - Perceived Benevolence Results

The high condition followed by the accurate one performed best in perceived benevolence (Figure 2e). The two conditions shared the highest third (1.00) and first (0.00) quartiles. However, the median of the accurate condition (0.50) is below the median of the high condition (1.00) and below the third quartile of the random condition. Neither of the remaining conditions performed overall worse than the other, considering their medians, first quartiles, and third quartiles.

Design Guideline 4: Mitigate concerns about biasing towards high-confidence recommendations by showcasing exploratory benefit of engaging with lower-confidence recommendations. Two random-condition participants made the important observation that the confidence levels may influence their music preferences. P16-R commented, "*I think it's better without a comment saying whether I'd like it or not almost because I feel like some biases could come out.… almost like trick me into thinking I'd like it more if it says I'm gonna like it.*" Similarly, P24-R said of the

option for a static confidence signal instead, "*I feel like that could be a good thing because sometimes if a particular system thinks that you like this song, right? And they show more songs similar to that type of song... it doesn't really give you much variety. So maybe if it was worded like that, I would be more open to trying other songs as well.*" As with recommendation itself, it is critical to consider how confidence signals may lead to a feedback loop in which participants see narrower recommendation selections over time [8]. Nonetheless, the accurate and high conditions performed best in terms of perceived benevolence, suggesting that the presence of high-confidence signals did not generally beget concerns of malevolence. As the two participants who raised bias concerns received the random condition, they may have been more aware of the potential for bias, given the combination of a dynamic and inaccurate confidence signal.

Nine participants shared that they would be more likely to listen to a recommendation or listen for longer if its associated confidence were higher, indicating some bias towards higher-confidence recommendations. For example, if faced with a high-confidence signal, P23-A was "*more willing to listen to more of a song... because it's hard to tell in just the first couple of seconds.*" On the other hand, P22-R explained that "*if I saw [a recommendation] of the lowest [confidence] level then, depending on my mood, I might just not listen to it.*" Even the general presence of confidence signals led 4 participants to be more willing to listen to recommendations. P30-L observed, "*I feel like somewhere subconsciously [the static low-confidence signal] does kind of make you more like... 'oh, I want to listen to this song now.'* " To reduce concerns around bias towards high-confidence recommendations and encourage a more open-minded view of low-confidence ones, we may present low-confidence signals as exploratory (see Section 3.2). Reasons behind recommendations (see Section 3.1) may also improve perceived benevolence by clarifying how recommendations are relevant.

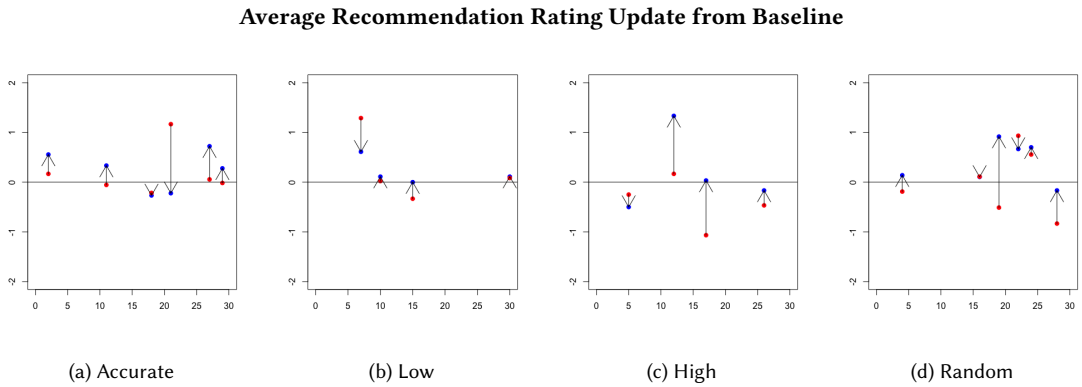
3.5 RQ1 - Perceived Anthropomorphism Results

For perceived anthropomorphism, the accurate condition had the only positive median (1.00) and the highest first quartile (0.25), but it lost the highest third quartile to the random condition (1.00 versus 1.75) (Figure 2f). The high condition performed worst with the lowest median shared with the low and random conditions (0.00), third quartile shared with the low condition (0.00), and first quartile (-1.00).

Design Guideline 5: Provide accurate confidence signals for natural interaction that does not downplay system limitations. Six participants noted that, compared to no confidence signals, dynamic ones did or would make the recommender appear more friendly or human-like. P18-A shared, "*Even if it's completely algorithmically generated, it feels like it's more akin to... like your friend showing up and being like... 'you should check this artist out.*"' Despite no quantitative evidence of an increase in perceived anthropomorphism under the static conditions, five participants mentioned feeling that static confidence signals would make the experience more human-like as well. P15-L recounted, "*When I read that [static low confidence signal], I was like, 'oh, that makes me want to go and listen to [the song] like it's recommended to me personally.*" That said, prior work highlights how anthropomorphism can be interpreted as setting users up for disappointment [14]. As described in Section 3.2, our results indicate that static as opposed to dynamic confidence signals do not provide the same sense of transparency that can help users avoid disappointment with bad recommendations. Dynamic confidence signals, if presented accurately, may make the interaction more natural without causing over-reliance, as discussed further in Section 3.6. P29-A summarized, "*I feel like [the dynamic confidence] gives it more of a sense of being human because it's... not always perfect, and it makes you know that it's not always perfect.*"

3.6 RQ2 - Reliance Results

Looking at Figure 3, we can see little indication that any condition led to more over-reliance than the baseline. In the accurate condition, only one participant moved slightly in the direction of more over-reliance, as opposed to more



(a) Accurate (b) Low (c) High (d) Random

Fig. 3. For each condition and participant ID (x-axis), the baseline (red) and non-baseline (blue) average recommendation rating update. Above zero implies under-reliance because the final rating was higher than expected, while below zero implies over-reliance.

appropriate or under-reliance. Meanwhile, 4 of the 6 of participants in the accurate condition became more under-reliant in comparison to the baseline. The accurate condition may have increased participants' under-reliance compared to the baseline because they were more aware of the fact that the system was imperfect and still learning about them. However, as noted in Section 3.1, participants' trust in and desire to use the system appeared not to suffer but rather to improve under the accurate condition. Indeed, 8 participants commented that a lower confidence level did or would allow them to be more understanding of bad recommendations. With the willingness to continue to use the system, users may build their mental models of the recommender and develop appropriate reliance over time. Nonetheless, in other domains, there may not be enough opportunities for users to properly calibrate their reliance before a significant amount of time or resources is lost. For example, if a user is to adjust their expectations for confidence levels for movie recommendations, one sample requires one or two hours to ingest as opposed to a couple minutes. Moreover, if the recommendation domain has more serious consequences (e.g., healthcare), then the user may not be able to afford experiencing unanticipated bad recommendations before arriving at a better understanding of the confidence levels.

With the exception of one participant under the high condition, participants in the random and static conditions also appeared to only develop more appropriate or under-reliance compared to the baseline. For these conditions, the lack of over-reliance could have been due to participants noticing that the received confidence signals were not aligning with their experience. As an example, P19-R commented, "*I like the idea of it saying the different levels of how well you'll like it, but it didn't seem accurate.*" Despite avoiding over-reliance, these conditions were weaker than the accurate one in terms of promoting trust and desire to use the system. Therefore, users may not be as encouraged to use the system long-term, and benefits of avoiding over-reliance or increasing appropriate reliance may be more difficult to realize.

4 DISCUSSION

Our results suggest that, in comparison to the other confidence signal patterns tested, **accurate confidence signals provide the best chance for increasing trust over no confidence signal**. While previous work on recommender confidence and natural-language confidence has had mixed results in terms of confidence improving trust [10, 19, 28, 30], our accurate confidence signal may have more easily increased trust due to the doubly personal nature of conversation and recommendation. However, simply increasing trust can lead to over-reliance [4, 8, 21]. The goal behind the accurate confidence signals is not to encourage blind trust but rather to help users calibrate their expectations for each

recommendation. Our preliminary results indicate that **the accurate condition as compared to the baseline did not encourage over-reliance but potentially under-reliance**. The transparency of the accurate condition may prompt users to be more understanding of the system's flaws, while also having lower expectations for its recommendations. Compared to related work on natural-language confidence [37], this work may have had less success with improving appropriate reliance due to its recommendation setting. More time may be required to calibrate understanding of confidence levels for recommendations in comparison to objective output, for which the levels may be more clear-cut. Although a number of participants under the accurate condition under-estimated their recommendations, participants with this condition also had more desire to use the system again, so they would have the opportunity to adjust their mental model of the confidence scale over time and potentially learn to have more appropriate reliance.

Participants overall had less trust in the random condition as compared to the accurate one. The random condition did not demonstrate much improvement in trust with respect to the baseline either; the median change in each trust-related metric was no more than zero. This suggests that **participants did not have higher trust in the accurate confidence signals simply because the confidence was dynamic**. Furthermore, participants generally had more trust in the accurate condition than the high and low conditions, which suggests that **accurate confidence signals may improve upon static confidence signals to augment the user experience**. While many recommender systems today respond to recommendation requests with unchanging confidence signals such as "you may like this" or "we think you'll like this," our results suggest that that space may be better suited for accurate confidence signals that help users to trust the recommender system without over-relying on it. Though the random and static confidence signals, like the accurate ones, mainly led to more appropriate or under-reliance in comparison to the baseline, because they did not prompt the same amount of trust and desire to use the system, users would likely not benefit as much from these confidence signals.

5 LIMITATIONS AND FUTURE WORK

This study had a small sample size, which was useful for gathering in-depth qualitative and exploratory quantitative insights, but more quantitative data is needed to compliment our results. Also, it would be interesting to see how different confidence signal patterns interact with different actual confidence trajectories, such as one that oscillates. Additionally, the wording for each confidence level did not vary, but these wordings could have significant effects on how people perceive each confidence level. Compared to the text-based interaction here, a voice-based one may surface unique results. Furthermore, a number of participants appreciated the idea of occasional confidence signals; future work may investigate how frequently confidence signals should be presented in CRSs. Participants also expressed interest in combining confidence with reasons for a recommendation. Future work may compare participants' experiences when receiving confidence, reasons, or both. Because we noted that reliability and confidence may be confused, future work may explore how to communicate the difference. Moreover, future work may focus on different recommendation domains and user groups other than just adults aged 18 to 24 years. Future work may investigate how our design guidelines extend to non-conversational recommenders as well. Finally, adapting this study to be longitudinal would allow us to understand how users' mental models adjust to different confidence signal patterns over time.

ACKNOWLEDGMENTS

The authors thank Mark Bowers for his assistance with the prototype design, the many people at Google and anonymous reviewers who provided helpful feedback on this work, and the participants who made this work possible.

REFERENCES

[1] Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. 2021. Alexa, Google, Siri: What are your pronouns? Gender and anthropomorphism in the design and perception of conversational assistants. *arXiv preprint arXiv:2106.02578* (2021).

[2] Theo Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior* 85 (2018), 183–189.

[3] Krisztian Balog and Filip Radlinski. 2020. Measuring recommendation explanation quality: The conflicting goals of explanations. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 329–338.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[5] Mustafa Bilgic and Raymond J Mooney. 2005. Explaining recommendations: Satisfaction vs. promotion. In *Beyond personalization workshop, IUI*, Vol. 5. 153.

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[7] Shuo Chang, F Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 175–182.

[8] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and debias in recommender system: A survey and future directions. *ACM Trans. Inf. Syst.* 41, 3 (2020), 39 pages.

[9] Jing Chen, Scott Mishler, and Bin Hu. 2021. Automation error type and methods of communicating automation reliability affect trust and performance: An empirical study in the cyber domain. *IEEE Transactions on Human-Machine Systems* 51, 5 (2021), 463–473.

[10] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction* 18, 5 (2008), 455–496.

[11] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies—why and how. *Knowledge-based systems* 6, 4 (1993), 258–266.

[12] Mateusz Dubiel, Sylvain Daronnat, and Luis A Leiva. 2022. Conversational Agents Trust Calibration: A User-Centred Perspective to Design. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–6.

[13] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30, 4 (2020), 681–694.

[14] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[15] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 241–250.

[16] Ken Hyland and John Milton. 1997. Qualification and certainty in L1 and L2 students' writing. *Journal of second language writing* 6, 2 (1997), 183–205.

[17] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–36.

[18] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[19] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. *arXiv preprint arXiv:2202.01875* (2022).

[20] Paul Lamere. 2014. *The Skip*. https://musicmachinery.com/2014/05/02/the-skip/

[21] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[22] Yidu Lu and Nadine Sarter. 2019. Feedback on system or operator performance: Which is more useful for the timely detection of changes in reliability, trust calibration and appropriate automation usage?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 312–316.

[23] Alan M MacEachren, Robert E Roth, James O'Brien, Bonan Li, Derek Swingley, and Mark Gahegan. 2012. Visual semiotics & uncertainty visualization: An empirical study. *IEEE transactions on visualization and computer graphics* 18, 12 (2012), 2496–2505.

[24] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[25] Sean M McNee, Shyong K Lam, Catherine Guetzlaff, Joseph A Konstan, and John Riedl. 2003. Confidence displays and training in recommender systems. In *Proc. INTERACT*, Vol. 3. 176–183.

[26] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[27] Philipp Schmidt, Felix Biessmann, and Timm Teubner. 2020. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 29, 4 (2020), 260–278.

[28] Anuschka Schmitt, Thiemo Wambsganss, and Andreas Janson. 2022. Designing for Conversational System Trustworthiness: The Impact of Model Transparency on Trust and Task Performance. *European Conference on Information Systems Research Papers* (2022).

[29] Anna-Maria Seeger, Jella Pfeiffer, and Armin Heinzl. 2017. When do we need a human? Anthropomorphic design and trustworthiness of conversational agents. *SIGHCI 2017 Proceedings* (2017).

[30] Guy Shani, Lior Rokach, Bracha Shapira, Sarit Hadash, and Moran Tangi. 2013. Investigating confidence displays for top-N recommendations. *Journal of the American Society for Information Science and Technology* 64, 12 (2013), 2548–2563.

[31] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

[32] Nava Tintarev. 2007. Explanations of recommendations. In *Proceedings of the 2007 ACM conference on Recommender systems*. 203–206.

[33] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.

[34] Diana-Cezara Toader, Grațiela Boca, Rita Toader, Mara Măcelaru, Cezar Toader, Diana Ighian, and Adrian T Rădulescu. 2019. The effect of social presence and chatbot errors on trust. *Sustainability* 12, 1 (2019), 256.

[35] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[36] Douglas A Wiegmann, Aaron Rich, and Hui Zhang. 2001. Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science* 2, 4 (2001), 352–367.

[37] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *CHI Conference on Human Factors in Computing Systems*. 1–28.

Appendix A PERSONALIZED RECOMMENDATION SELECTION

When potential participants signed up for the studies, they were asked to select and describe the most common situation in which they would listen to a playlist for 30 minutes or more. They were given multiple options including an "Other" option to fill in themselves. In addition, they were asked for one genre, three artists, and two songs per artist that they would want to hear in that situation. For the genre, they had several options including an "Other" option to fill in themselves. If they selected "Other" for the genre, they were screened out of the study, as we did not have an associated top-50 chart necessary for selecting their song recommendations. If they could not think of answers for any of the above questions, they were also screened out. Using this information and YouTube Music, we selected personalized

recommendations of low, medium, and high confidence for each participant in advance of their study session. For each participant, we made a new Gmail account with which to access YouTube Music. For the low-confidence songs, we used 12 songs recommended before any interaction with the system. In order for these songs to be the same for all participants, participants needed to be around the same age because Gmail requires a birth date, which could influence the recommendations. We therefore recruited participants in the age range of 18 to 24 and set Gmail's mandatory fields of gender to "Rather not say" and birth date to February 15, 2000, the average birth date of an 18- to 24-year-old. To obtain the medium-confidence songs, we created two playlists. One consisted of the top 6 songs from the YouTube Music top-50 chart for the participant's preferred genre; the other consisted of the top 20 songs from the same chart. The top 6 song suggestions for each playlist made up the 12 medium-confidence songs. We used both playlists because both appeared to receive similar ratings in a pilot study. To acquire the high-confidence songs, we created 4 playlists. The participant's 6 preferred songs comprised one playlist, with its top 6 suggestions used as high-confidence songs. The other three playlists corresponded to the three preferred artists; each playlist consisted of the two preferred songs associated with that artist. The top two song suggestions for each of these three playlists were also used as high-confidence songs. Again, we used two methods to generate high-confidence recommendations because they received similar ratings in piloting. To control for familiarity, we did not provide song recommendations by any of their preferred artists. If a recommendation appeared under more than one playlist, we assigned it to the highest confidence level under which it appeared. The songs for each confidence level and participant were divided randomly between the participant's two conditions.

After the study, we ran a two-sample paired sign test (rather than a Wilcoxon signed-rank test due to the violation of the symmetric distribution assumption) and found that there was indeed a significant difference between participants' average final ratings for the low-confidence recommendations versus the high-confidence recommendations ($p<.05$). Wilcoxon signed-rank tests also showed a significant difference in average final ratings between the medium- and high-confidence recommendations ($p<.05$) and a trend in the difference between the low- and medium-confidence recommendations' average final ratings ($p<.1$).

Appendix B INTERVIEW QUESTIONS

Table 2. Interview questions that were covered consistently and how many participants received each. A number of participants did not receive all questions due to limited time or the fact that some questions were added based on ideas from later interviews.

Interview Question	# of Participants
Compare the two recommenders	30
Likes/dislikes regarding received confidence signal (referred to as 'comment')	30
Impact of received comment on expectations	30
Other impacts of received comment on perception of recommender	30
Thoughts on dynamic or static confidence signals (whichever was not received)	30
Expectations for low, medium, or high confidence	29
Thoughts on receiving reasons for recommendations	23
Thoughts on whether reasons, confidence signals, or the combination is most helpful	17
Thoughts on an occasional confidence signal (exploratory, high-confidence, or low-confidence)	22
Expectations for confidence trajectory from one recommendation to the next	12
Any other thoughts on if/how music recommender should share its confidence	26

Appendix C CODEBOOK

Table 3. For interview analysis, codes and their descriptions.

Code	Description
RECOMMENDATION EXPECTATIONS AND REACTIONS	
quality-expectations	Confidence signal's effect on expectations for recommendation quality.
novelty-expectations	Confidence signal's effect on expectations for recommendation familiarity/novelty.
bad-rec-reaction	Confidence signal's effect on how would react to bad recommendation.
good-rec-reaction	Confidence signal's effect on how would react to good recommendation.
listening-reaction	Confidence signal's effect on how much willing to listen to a recommendation.
bias-reaction	How confidence signals may bias towards liking or disliking a recommendation.
wording-little-effect	How confidence wording doesn't impact expectations or reaction to recommendations. (This includes noting that color-coding is more helpful than wording for recognizing confidence levels.)
other-expectation-reaction	Confidence signal's other effects on expectations or reactions.
PREFERENCES	
ignore-or-doesn't-like-confidence-given	Ignores, does not notice, dislikes, or is neutral on confidence signals given.
likes-confidence-given	Likes confidence signals given.
doesn't-like-confidence-general	Dislikes or is neutral on confidence signals in general.
likes-confidence-levels	Likes idea of comments that adjust depending on recommender confidence.
doesn't-like-confidence-levels	Dislikes or is neutral on idea of confidence signals that adjust depending on recommender confidence.
trajectory-improve	Preference or expectation for confidence generally improving over time rather than another confidence trajectory (e.g., oscillating).
trajectory-other	Preference or expectation for another confidence trajectory (e.g., oscillating) other than confidence generally improving over time.
likes-occasional-idea	Likes an occasional confidence signal idea.
doesn't-like-occasional-idea	Dislikes or is neutral on an occasional confidence signal idea.
likes-reasons	Likes idea of receiving reasons behind recommendations provided (e.g., "because you like Artist X").
doesn't-like-reasons	Dislikes or is neutral on idea of receiving reasons behind recommendations provided (e.g., "because you like Artist X").
reason-plus-confidence	Combination of reason and confidence (over either alone) can be useful.
reason-over-confidence	Just reason (over just confidence or both) can be useful.
confidence-over-reason	Just confidence (over just reason or both) can be useful.
RECOMMENDER CHARACTERISTICS	
transparency	Confidence signal's effect on how open/honest recommender appears.
ability	Confidence signal's effect on how competent recommender appears.
benevolence	Confidence signal's effect on how well-intentioned/sincere recommender appears.
anthropomorphism	Confidence signal's effect on how human-like recommender appears (e.g., robotic, friendly, personal, natural).