# Evaluating Human-AI Hybrid Conversational Systems with Chatbot Message Suggestions

Zihan Gao*
University of Wisconsin-Madison
Madison, Wisconsin, USA
zihan.gao@wisc.edu

Jiepu Jiang
University of Wisconsin-Madison
Madison, Wisconsin, USA
jiepu.jiang@wisc.edu

## ABSTRACT

AI chatbots can offer suggestions to help humans answer questions by reducing text entry effort and providing relevant knowledge for unfamiliar questions. We study whether chatbot suggestions can help people answer knowledge-demanding questions in a conversation and influence response quality and efficiency. We conducted a large-scale crowdsourcing user study and evaluated 20 hybrid system variants and a human-only baseline. The hybrid systems used four chatbots of varied response quality and differed in the number of suggestions and whether to preset the message box with top suggestions.

Experimental results show that chatbot suggestions—even using poor-performing chatbots—have consistently improved response efficiency. Compared with the human-only setting, hybrid systems have reduced response time by 12%–35% and keystrokes by 33%–60%, and users have adopted a suggestion for the final response without any changes in 44%–68% of the cases. In contrast, crowd workers in the human-only setting typed most of the response texts and copied 5% of the answers from other sites.

However, we also found that chatbot suggestions did not always help response quality. Specifically, in hybrid systems equipped with poor-performing chatbots, users responded with lower-quality answers than others in the human-only setting. It seems that users would not simply ignore poor suggestions and compose responses as they could without seeing the suggestions. Besides, presetting the message box has improved reply efficiency without hurting response quality. We did not find that showing more suggestions helps or hurts response quality or efficiency consistently. Our study reveals how and when AI chatbot suggestions can help people answer questions in hybrid conversational systems.

## KEYWORDS

conversational agents; chatbots; question-answering; reply suggestion; predictive text suggestion; human-AI hybrid systems

*This work was done while the author was an undergraduate research intern at the University of Wisconsin-Madison.


*CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia*

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482340

**ACM Reference Format:**
Zihan Gao and Jiepu Jiang. 2021. Evaluating Human-AI Hybrid Conversational Systems with Chatbot Message Suggestions. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3459637.3482340

## 1 INTRODUCTION

Conversational agents such as chatbots are a popular way of communicating and providing information. Previous research has made remarkable progress in designing chatbots [11, 16, 22, 32, 34, 41, 47, 49, 50]. However, AI chatbots cannot replace humans in many scenarios. For example, many companies use hybrid customer service systems combining human representatives and AI chatbots to ensure conversation quality.

We study a human-chatbot hybrid system design where we offer users chatbot reply suggestions. Users may directly adopt a suggestion to reply or edit responses on top of a suggestion. Many email [20, 35] and messaging [14] apps have already used such reply suggestions. However, they focus on facilitating communication—the reply suggestions are primarily short and functional, such as expressing gratitude ("thank you"), confirmation ("got it"), or accepting/declining others' proposals. Previous studies found that such reply suggestions may improve response efficiency [20].

In contrast, we examine using chatbot suggestions to facilitate human agents providing information. Specifically, we evaluate if chatbot suggestions help people answer knowledge-demanding questions during a conversation, where chatbot suggestions may help in two ways. First, they provide essential relevant knowledge for answering questions, especially when humans do not know or have difficulty recalling relevant information. Second, they may improve response efficiency because users may only need minor edits on top of a suggestion. Previous work of text suggestion in information seeking has only focused on supporting seekers (e.g., query suggestion [21] and autocompletion [27], and user request clarification [1, 52]), while we support information providers.

We use a crowdsourcing user study to examine this matter. We assigned crowd workers to answer knowledge-demanding questions using systems with different reply suggestion supports. We evaluated 20 hybrid system variants. They differ regarding the chatbots used for providing suggestions, the number of displayed suggestions, and whether to preset users' message input box. We recorded the time and keystrokes needed to finish the answers and recruited other crowd workers to assess response quality. We compare the hybrid systems (human+AI chatbot suggestions) with a baseline without any suggestions (a human-only system) and the top-ranked chatbot responses (an AI-only system). The rest of the article introduces our experiment and findings.

## 2 RELATED WORK

### 2.1 Text Suggestion

Our study is closely related to previous work on text suggestions in various scenarios. Text suggestion is a widely-used technique for assisting text entry. For example, many mobile devices offer word suggestions while users type on on-screen keyboards [18, 30]. Search engines also provide query suggestions [21] and autocompletion [27]. Besides, email [20, 35] and messaging [14] apps today also provide short reply suggestions. The presented text suggestions can be static (e.g., showing selectable quick reply suggestions [20, 35]) or interactive (e.g., showing text completions while users type [2, 4]). Also, the unit of suggestion can be a word [30], multiple words [2, 4], short sentences [20, 35], etc. Previous studies found that such text suggestions may improve text entry efficiency, although some other factors may hurt its usefulness [2, 14, 36].

Our study differs from previous work in several ways. First, we evaluate text suggestions in question-answering conversations, and we specifically examine how well they support the answerers. Such tasks are more knowledge-demanding than previous ones, and the text suggestions are also longer (about 70–150 characters) than those in previous studies. Second, in addition to text entry efficiency (e.g., messaging time and keystrokes), we also look into answer quality from various dimensions.

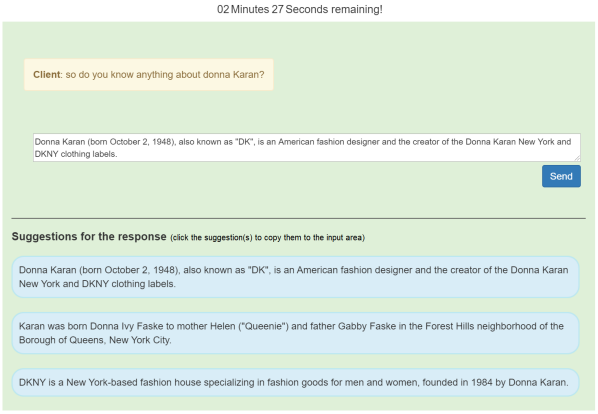
### 2.2 Conversational Systems and Chatbots

We study text suggestions in conversational systems. Conversational systems are agents designed to converse with humans in text, speech, or combined [19], e.g., dialog systems and chatbots. Dialog systems mainly help users solve tasks such as giving directions, finding restaurants, booking flights, and controlling smartphone functionalities. Most dialog systems model human-agent conversation as a sequence of acts and states [46, 51, 55] and offer template-based outputs. In contrast, chatbots aim to respond in natural language, unstructured conversations in a human-like manner [11, 16, 22, 32, 34, 41, 47, 49, 50]. These conversations can be chit-chats or informational, where system responses are not restricted to predefined templates. Today many commercial products such as Siri, Google Now, and Cortana combine dialog systems, chatbots, and other functionalities (e.g., mobile web search). Also, many other systems use conversation-like interaction, e.g., conversational search [28, 29, 31, 42–44] and recommendation systems [6, 54] and sequential question answering [16, 38].

Our hybrid systems' scope is similar to chatbot systems. However, it differs because we use a hybrid model and let humans work with an AI chatbot to converse with another human in question-answering conversations. We provide users with chatbot outputs as reply suggestions while they compose responses. Such hybrid systems have many potential applications, e.g., improving online customer service quality and efficiency [12, 48].

We also examine hybrid systems equipped with different chatbots. Current chatbots use retrieval-based or generation-based methods trained on conversation corpora to select or synthesize responses. Retrieval-based chatbots [16, 47, 49, 50] search for texts that are most likely appropriate responses in existing corpora, making the response generation task similar to sentence and paragraph-level text retrieval problems. The effectiveness of retrieval-based approaches largely depends on whether the utterance corpora include an ideal response. This assumption is reasonable in application scenarios where similar dialogues repeat but may not hold in open-domain discussions. In contrast, generation-based chatbots [11, 22, 32, 34, 41] synthesize new utterances based on response patterns learned from the training corpora. Here we evaluate hybrid systems using different chatbots of varied response quality, including retrieval-based and generation-based ones.

**Figure 1: A screenshot of the hybrid conversational system.**




Evaluating chatbot response quality is challenging. On the one hand, many existing datasets are not using natural online conversations, but online social media discourses or movie lines [5, 7, 10, 25, 33, 37, 45]. On the other hand, automatic evaluation methods for chatbot responses mostly use impractical settings or have low consistency with human judgments [23]. For example, many studies evaluated retrieval-based chatbots by whether they can rank the actual response to the top without examining the critical assumption of retrieval-based chatbots (the coverage and generalizability of the response corpora to new tasks). In contrast, we use large-scale crowdsourcing judgments to evaluate the quality of responses in our hybrid systems.

## 3 EXPERIMENT

We use online crowdsourcing experiments to study hybrid conversational systems with chatbot suggestions. Our primary purpose is to compare hybrid systems with human-only and chatbot-only baselines regarding response quality and efficiency. Also, we examine the influence of chatbot suggestion quality and two interface design factors—the number of displayed suggestions and message box initialization. We assign crowd workers to finish simulated question-answering tasks using different systems and record their responses for analysis.

### 3.1 Conversation Tasks

Human conversations are complex activities and may have different intents. Here we focus on answerers in question-answering conversations, where the task is to answer informational questions. We have created a task pool based on the Wizard of Wikipedia dataset [9]. We chose this dataset because the included conversations are informational and knowledge-demanding but do not require extensive domain expertise (to ensure that crowd workers

**Figure 2: An example conversation task and its chatbot suggestions.**

| Conversation Context |
|---|
| **Person 1**: Sushi is so amazing! I love the taste of the different sea foods. Do you like sushi? |
| **Person2**: Love sushi. I haven't had it in a while Would love some right now since it's dinnertime for me. I prefer it with white rice, instead of brown. |
| **Person 1**: Do you know where sushi originally comes from? |

| | "RetWiki" Suggestions | "RetBest" Suggestions | "RetWorst" Suggestions | "Gen" Suggestion |
|---|---|---|---|---|
| 1 | Sushi originates in a Southeast Asian dish, known today as 'narezushi' (–'salted fish'), stored in fermented rice for possibly | **In Japan around the Mayor period. It's known there as narezushi.** | Hi I am not sure but if it's true it sounds funny. Don't you think? | It is prepared with raw seafood, but it is often prepared with raw seafood. |
| 2 | The lacto-fermentation of the rice prevented the fish from spoiling; the rice would be discarded before consumption of the fish. | Hi I am not sure but if it's true it sounds funny. Don't you think? | I can't say that I do. | |
| 3 | This early type of sushi became an important source of protein for its Japanese consumers. | I can't say that I do. | I didn't know that at all. That's very interesting. | |
| 4 | The term 'sushi' comes from an antiquated grammatical form no longer used in other contexts, and literally means 'sour-tasting'; the overall dish has a sour and Miami or savory taste. | I didn't know that at all. That's very interesting. | That sounds excellent. Where did it originate? | |
| 5 | Narezushi still exists as a regional specialty, notably as 'funa-zushi' from Shiva Prefecture. | That sounds excellent. Where did it originate? | Yeah I found out that the stat of having lost someone is called 'widowhood'. | |

can handle). Also, this dataset is more similar to real-world conversations than many others [10, 25, 45] because the messages came from a user study's chat log.

The task pool included 90 conversations with one, two, or three rounds of existing messages (30 for each case). For each task, we show participants an existing conversation, and they need to respond to the most recent message. The participants only play the role of an answerer. For example, Figure 2 shows a task with two existing rounds, where the participant plays the role of Person 2 and responds to Person 1's last message (a question). Our systems do NOT further reply to the participant's response—participants only have one-shot interaction with the system. Such one-shot interaction is similar to batch-mode online customer support, where human agents can take over any ongoing conversation, respond to it, and move on to the next one. We randomly sampled conversations from the dataset and manually removed unqualified ones (e.g., chit-chats or those that need domain expertise).

## 3.2 Hybrid Conversational Systems

Our hybrid systems show users chatbot responses for the same conversation as suggestions to help them compose answers. Figure 1 shows a screenshot of a hybrid system with three suggestions. Clicking on a suggestion will append its content to the message box. Users can make further edits and select other suggestions. Clicking on the "Send" button will finish a response (and the task).

We also examine two system design considerations. The first one is to initiate the message box—to leave it blank or preset it with the top-ranked suggestion. We suspect presetting the text box may encourage users to compose responses based on chatbot suggestions (as it requires extra effort to remove the preset content and write from scratch). The second one is the number of displayed suggestions. Showing more suggestions increase the chances of providing high-quality suggestions but may cost users more effort to read the suggestions.

## 3.3 Chatbot Systems

We suspect that the chatbot used for providing suggestions plays a critical role in the success of a hybrid system. Thus, we compare hybrid systems using different chatbots. Note that we are less interested in knowing any specific chatbot is a better choice than others. Instead, we hope that our chatbot selections can provide diverse samples of suggestions with varied quality and characteristics to help draw more generalized conclusions regarding the influence of chatbot suggestion quality on hybrid systems' responses.

- **`RetBest`** is a retrieval-based chatbot retrieving messages from the Wizard of Wikipedia dataset as responses. The dataset includes the actual responses to the conversation tasks. This ensures an ideal response exists in the conversation corpus and stands for the retrieval-based chatbot's "best" possible performance.
- **`RetWorst`**, in contrast, is the retrieval-based chatbot after removing the actual responses to the selected 90 conversation tasks. `RetWorst` stands for the performance of the retrieval-based chatbot when the conversation repository does not generalize well to the target tasks.
- **`RetWiki`** is another retrieval-based chatbot retrieving sentences from an external but high-quality corpus—Wikipedia. Wikipedia is not a conversation repository but provides highly informative sentences that may provide users relevant knowledge for answering questions.
- **`Gen`** is a generation-based chatbot that can synthesize responses that do not exist in the training corpus. `Gen` provides one single response output, while the three retrieval-based chatbots offer a ranked list of messages.

We built the four chatbots using existing models implemented in ParlAI [26], an open-source conversational system toolkit. The three retrieval-based chatbots use the Retrieval Transformer Memory Network [26]. We trained the network to minimize the cross-entropy loss on the Wizard of Wikipedia dataset, excluding the 90 conversations selected into our task pool. We use the trained model to retrieve responses from the Wizard of Wikipedia dataset as results for `RetBest`. `RetWorst` simply removes the actual responses of the 90 conversations from `RetBest`'s results. `RetWiki` uses the same model to rank Wikipedia sentences. The generation-based chatbot uses the Generative Transformer Memory Network [26]. We trained the network to minimize the negative log-likelihood of response messages on the Wizard of Wikipedia dataset. Figure 2 shows example top results of the four chatbots.

## 3.4 Experimental Design

We conducted a crowdsourcing user study to compare different systems. Our experiment used a between-subjects design. We assign

**Table 1: Experimental settings (21 systems in total).**

| Chatbot Support | Message Box Initialization | Num. of Suggestions | Num. of Participants |
|---|---|---|---|
| Human (no support) | - | - | 50 |
| RetWiki | blank/preset | 1/3/5 | $50 \times 2 \times 3$ |
| RetBest | blank/preset | 1/3/5 | $50 \times 2 \times 3$ |
| RetWorst | blank/preset | 1/3/5 | $50 \times 2 \times 3$ |
| Gen | blank/preset | 1 | $50 \times 2 \times 1$ |

each participant to one of the following 21 systems to finish some conversation tasks. Table 1 summarizes the experimental settings.

- Human is a baseline system where users are only provided with a text box to input their replies without any suggestions. However, we did not prohibit or encourage crowd workers to acquire help on a search engine or other external sites.
- For each of RetWiki, RetBest, and RetWorst, we examine the hybrid system variants using the chatbot to offer top 1, 3, or 5 responses as suggestions, with or without presetting the message box. This includes 18 systems in total ($3 \times 3 \times 2 = 18$).
- We also examine two hybrid system variants using Gen for chatbot suggestion, with or without presetting the message box.

We required each participant to finish an experimental session of five minutes (excluding the time spent on instructions and a training task at the beginning of the session). The participants completed conversation tasks randomly sampled from the pool one after another until five minutes. We instructed participants to provide informative responses instead of short and uninformative replies such as "I don't know." We recorded the participants' responses and their keystrokes on the messaging interface.

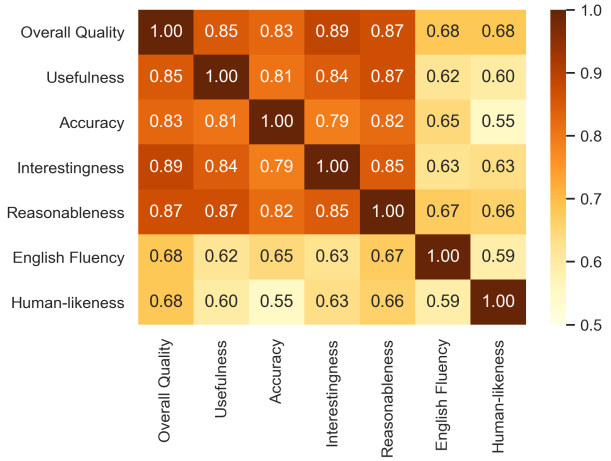
For each system, we recruited 50 participants from Amazon Mechanical Turk. We required them to have a higher than 95% HIT approval rate and at least 1,000 approved HITs, though we did not require them to be English native speakers (Amazon Mechanical Turk did not provide this qualification filter). We paid each HIT (a 5-minute session) $0.25. We instructed the participants that other human workers would assess their responses, and they needed to finish at least five conversations with informative responses. We instructed them that the top 10% performing HITs (by the number of finished conversations with informative responses) will receive a $0.25 bonus. We determined an informative response by an average overall quality rating of at least 3.0 on a 5-point scale by three different crowd workers. We have obtained institutional approval for this human subjects research.

## 3.5 Response Quality Judgments

We recruited another group of crowd workers to assess the quality of the collected responses (plus each chatbot's top responses). We showed assessors a conversation, highlighting the response to be judged, and asked them to rate the *overall quality* of the response plus its *usefulness*, *accuracy*, *English fluency*, *human-likeness*, *interestingness*, and *reasonableness*. The questions are adapted from previous studies evaluating chatbots [40, 53]:

- **Overall Quality** – How well would you rate the quality of the highlighted response? *Very Poor* (1), *Poor* (2), *Okay* (3), *Good* (4), *Very Good* (5).

**Figure 3: Pearson's correlation of response quality.**




- **Usefulness** – The highlighted response provides useful information relevant to the conversation. *Strongly Disagree* (1), *Disagree* (2), *Neutral* (3), *Agree* (4), *Strongly Agree* (5).
- **Accuracy** – The information provided by the highlighted response is correct and accurate. *Strongly Disagree* (1), *Disagree* (2), *Neutral* (3), *Agree* (4), *Strongly Agree* (5).
- **English Fluency** – The highlighted response sounds like someone who speaks fluent and natural English. *Strongly Disagree* (1), *Disagree* (2), *Neutral* (3), *Agree* (4), *Strongly Agree* (5).
- **Human-likeness** – I believe the highlighted response is from a real human instead of a bot. *Strongly Disagree* (1), *Disagree* (2), *Neutral* (3), *Agree* (4), *Strongly Agree* (5).
- **Interestingness** – The highlighted response reads interesting to me. *Strongly Disagree* (1), *Disagree* (2), *Neutral* (3), *Agree* (4), *Strongly Agree* (5).
- **Reasonableness** – The highlighted response is reasonable and logical in its context. *Strongly Disagree* (1), *Disagree* (2), *Neutral* (3), *Agree* (4), *Strongly Agree* (5).

Each judgment HIT included five conversations to be judged and a verification task. The verification task shows a response that is obviously irrelevant to the conversation and provides little information. We rejected a HIT if the worker provided an overall quality rating $\geq 3$ for the verification task. We paid each judgment HIT $0.2 and required the assessors to have a higher than 95% HIT approval rate and at least 1,000 approved HITs. We collected three assessors' ratings for each conversation response and used the mean values of the judgments as quality measures.

## 4 DATA

We have collected 1,050 experiment sessions. On average, an experiment session (5 minutes) had 9.9 completed conversations. We collected crowdsourcing judgments for user responses in all finished conversations (10,435 in total). We also collected judgments for each chatbot's top 5 results. These responses included many duplicates (5,477 unique ones in total), e.g., participants adopted a suggestion for the response without any changes.

Figure 3 shows the Pearson's correlation of response quality measures (all correlations are statistically significant at $p < 0.001$). Five measures (overall quality, usefulness, accuracy, interestingness,

**Figure 4: Distribution of response quality, length, and the time and keystrokes for composing messages ($N = 10,435$).**

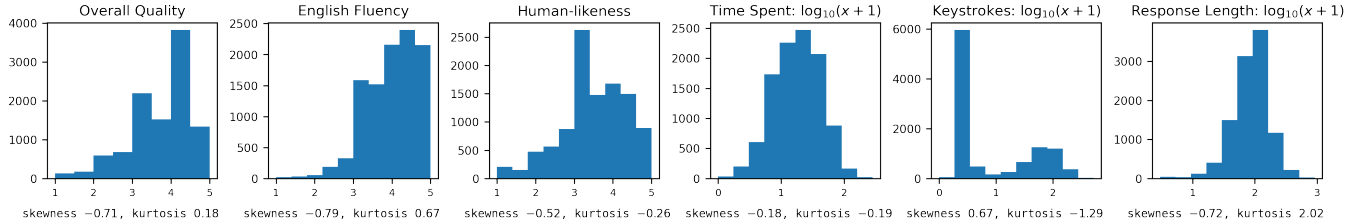



**Table 2: Comparisons between hybrid systems using different chatbots and the human-only and chatbot-only baselines.**

| | **Human** $N = 400$ | **RetWiki** $N = 2,962$ | | | **RetBest** $N = 3,292$ | | | **RetWorst** $N = 2,650$ | | | **Gen** $N = 1,131$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Chatbot | Hybrid | | Chatbot | Hybrid | | Chatbot | Hybrid | | Chatbot | Hybrid | |
| Overall Quality | 3.74 | 3.68 | 3.77 | ↑↑↑ | 3.87 | 3.89 | | 2.84 | 3.35 | ↑↑↑ | 3.10 | 3.43 | ↑↑↑ |
| Usefulness | 3.78 | 3.87 | 3.90 | | 3.90 | 3.93 | | 2.85 | 3.39 | ↑↑↑ | 3.23 | 3.53 | ↑↑↑ |
| Accuracy | 3.87 | 3.99 | 3.99 | | 3.94 | 3.96 | | 2.95 | 3.46 | ↑↑↑ | 3.36 | 3.64 | ↑↑↑ |
| Interestingness | 3.71 | 3.80 | 3.85 | | 3.90 | 3.91 | | 2.91 | 3.38 | ↑↑↑ | 3.20 | 3.48 | ↑↑↑ |
| Reasonableness | 3.91 | 3.73 | 3.86 | ↑↑↑ | 3.81 | 3.88 | ↑↑ | 2.79 | 3.40 | ↑↑↑ | 3.17 | 3.51 | ↑↑↑ |
| English Fluency | 4.06 | 3.96 | 4.01 | | 4.07 | 4.07 | | 3.54 | 3.83 | ↑↑↑ | 3.64 | 3.84 | ↑↑↑ |
| Humanlikeness | 3.77 | 3.06 | 3.31 | ↑↑↑ | 3.67 | 3.69 | | 2.92 | 3.39 | ↑↑↑ | 3.14 | 3.42 | ↑↑↑ |
| Time Spent on a Response (s) | 31.2 | - | 24.4 | | - | 21.8 | | - | 27.6 | | - | 20.5 | |
| Number of Keystrokes | 65.7 | - | 27.4 | | - | 21.4 | | - | 39.3 | | - | 29.6 | |
| % Response same as a suggest. | - | - | 55.6% | | - | 68.3% | | - | 44.2% | | - | 58.7% | |
| Response Length (characters) | 60.9 | 152.6 | 145.2 | ↓↓↓ | 99.2 | 102.9 | | 78.8 | 78.0 | | 73.2 | 75.5 | |

↑, ↑↑, and ↑↑↑: Hybrid is significantly higher than Chatbot at 0.05, 0.01, and 0.001 level. ↓, ↓↓, and ↓↓↓: Hybrid is significantly lower than Chatbot at 0.05, 0.01, and 0.001 level.
Orange shading: Hybrid > Human statistically significant at $p <$ 0.05, 0.01, and 0.001. Blue shading: Hybrid < Human statistically significant at $p <$ 0.05, 0.01, and 0.001.

and reasonableness) have strong positive correlations with each other ($r > 0.79$). In contrast, the other two (English fluency and human-likeness) have moderate positive correlations with other measures ($r$ ranges between 0.5 and 0.7). The trends of the five highly correlated measures are very similar. Thus, we only report overall quality as an example of the five measures in some analyses.

Figure 4 plots the distribution of the quality measures and other variables for the collected conversations. The response quality measures all have left-skewed bell-shaped distributions. We transform time spent and response length (the number of characters) by taking $\log_{10}(x + 1)$ where $x$ is the raw value. The transformed values also have bell-shaped distributions. The skewness and kurtosis of these five variables are within the acceptable range for data analysis requiring normal distributions, e.g., $(-2, 2)$ for skewness and $(-7, 7)$ for kurtosis [3, 13]. The number of keystrokes is still far from a normal distribution after transformation (although skewness and kurtosis are within the suggested range). Thus, we use non-parametric tests for keystrokes.

## 5 RESULTS

We examine the collected results to discuss the following questions:

- RQ1 (Section 5.1)—Do hybrid systems help users better answer questions compared with one with no suggestions?
- RQ2 (Section 5.2)—How do hybrid systems' responses compare with the provided chatbot suggestions?
- RQ3 (Section 5.3)—How do characteristics of the displayed suggestions influence users' responses in hybrid systems?
- RQ4 (Section 5.4)—Should hybrid systems preset the message box with top chatbot suggestions?
- RQ5 (Section 5.5)—How many suggestions should be provided?

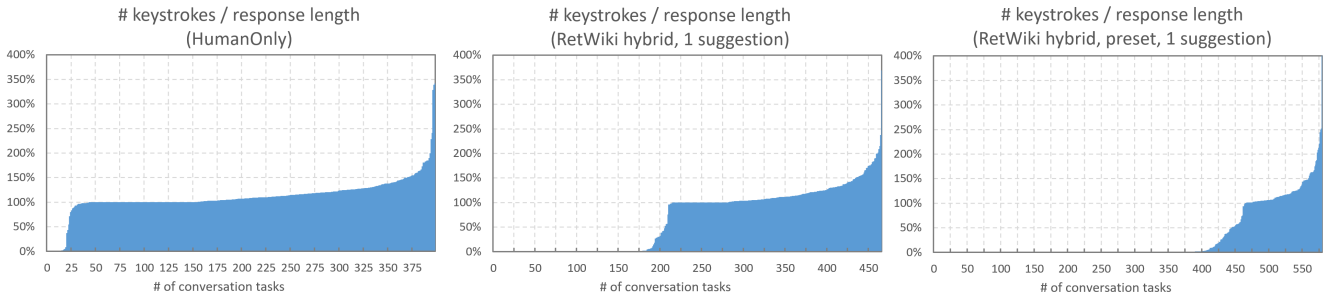
### 5.1 Hybrid vs. Human-Only Systems

Table 2 compares the hybrid systems with human-only and chatbot-only baselines. For the human-only baseline (Human), we report the mean values of each measure across all conversations. For the hybrid systems (Hybrid), we aggregate the experiment sessions using the same chatbot (ignoring the differences in the number of displayed suggestions and message box initialization) and report the mean values for hybrid systems using each chatbot.

We compare the five groups (Human and Hybrid using each chatbot) using a one-way ANOVA with participants and tasks being nested variables. The only exception is that we use a Kruskal–Wallis test for the number of keystrokes due to its skewed distribution. We test significant differences between two groups using the Bonferroni correction post-hoc test. We use different shadings in Table 2 for significant differences between Hybrid and Human at different levels. We test for significant differences using the log transformation of the time spent and response length. However, we report the raw values for the two variables in Table 2 for better interpretability.

Experimental results show that the hybrid systems have consistently improved users' response efficiency compared with the human-only baseline. On average, the hybrid systems have reduced response time by 12%–35%—the differences are significant at 0.001 level between the human-only setting and any hybrid systems. Participants using hybrid systems have also used significantly fewer—only about 33%–60%—keystrokes for a response than others with the human-only system. In contrast, the responses in the hybrid systems are also significantly longer than those from the human-only baseline. Altogether, we found that participants have used significantly shorter time and fewer keystrokes to reply with much longer responses in hybrid systems than the human-only baselines.

**Figure 5: The distribution of keystroke/response-length ratio in different conversational systems.**




We further examined why hybrid systems improve response efficiency. The primary reason is that users directly adopted a suggestion for the response—in the hybrid systems, 44.2%–68.3% of the replies are the same as a displayed suggestion. Users did also edit responses on top of an existing suggestion but with much lower frequency than adopting one with no changes. Figure 5 plots the number of keystrokes spent compared with the response's length (keystroke/response-length ratio). This ratio measures the efficiency of composing a message. For example, if a user types all the characters of a response text manually without making any mistakes, the ratio would be 1. If a user types manually but changes back and forth, the ratio would be higher than 1. The ratio would be close to 0 if the user directly adopts a suggestion for the response. As Figure 5 (the right two charts) shows, the ratio for a small proportion of responses lies between 0 and 1, indicating that users have edited their responses on top of an existing suggestion and saved some keystrokes. However, this happened much less frequently than directly adopting a suggestion in our experiments.

It worths noting that the keystroke/response-length ratio is close to 0 in about 5% of the conversations completed under the human-only condition, although we provided no suggestions. The most likely explanation is that users went to external websites and copied answer texts. We manually verified this by searching their responses in Google, and we did have found the source websites providing the copied contents. This indicates that it is natural for human agents to seek information even when their roles are to answer and provide information to others. Our hybrid systems have automated this process by suggesting relevant knowledge. We also suspect that chatbot outputs' actual "usage" should be even higher than simply looking at the keystroke/response-length ratio—users may get ideas from reading a suggestion's text without clicking on it.

In contrast, the hybrid systems did not consistently help or hurt response quality compared with the human-only baseline. The advantages of hybrid over human-only systems seem to depend on whether the chatbot can provide high-quality suggestions. Table 2 also reports the average quality of each chatbot's top responses. RetWiki and RetBest are two well-performing chatbots, and their top responses' quality is comparable to humans', while RetWorst and Gen are two poor-performing ones.

On the one hand, participants using hybrid systems with the two well-performing chatbots did outperform the human-only baseline in several (but not all) aspects of response quality. However, the magnitude of the improvements only lies between 0.1 to 0.2 on a five-point Likert scale. Besides, the hybrid systems' responses did not significantly improve in a few measures, e.g., the RetWiki hybrid systems' responses are significantly less human-like than those by the human-only baseline. On the other hand, hybrid systems using poor-performing chatbots consistently underperformed the human-only baseline in all quality measures. The magnitude of differences is evident (between 0.2 to 0.5), although these hybrid systems had still reduced response time and keystrokes significantly. It seems that users would not simply ignore poor suggestions and respond as they could without seeing the suggestions. Instead, they adopted some suggestions directly even though they might write better ones themselves.

To conclude, we found that hybrid systems with chatbot suggestions can consistently reduce the time and keystrokes needed for composing responses than the human-only baseline. The hybrid systems may improve response quality, provided that the chatbot provides high-quality suggestions. However, poor-performing chatbots' suggestions may reduce answer quality, though still help with response efficiency.

## 5.2 Hybrid vs. Chatbot-Only Systems

Table 2 also compares the hybrid systems with their corresponding chatbot-only baselines. We created a set of comparable observations for the chatbot-only baselines (`Chatbot`) to "simulate" a mixed-design experiment. For each session using a hybrid system, we create a corresponding chatbot-only "session" on the same sequence of tasks using the chatbot's top response as the answer. This allows us to compare hybrid systems with chatbot baselines on precisely the same tasks. We compare hybrid and chatbot-only systems using a two-way mixed-design ANOVA—the chatbot choice is a between-subject factor, and hybrid vs. chatbot-only systems is a within-subject factor. We compare each chatbot-hybrid pair using a Bonferroni post-hoc test and report significant differences by arrows with different directions.

The two-way tests have found significant overall differences and improvements (main effects) of hybrid systems over chatbot-only ones on all measures except response length. Participants' responses have higher quality ratings than the corresponding chatbot baselines in all hybrid systems, and their differences are statistically significant in most cases. Overall, this suggests that the hybrid systems improve the quality of responses over their chatbots' top-ranked suggestions. However, the magnitude of response quality improvements varies by the chatbots used in the hybrid systems. In those using the two poor-performing chatbots, participants have improved response quality by 0.2 to 0.5 on a five-point Likert scale

**Table 3: Multilevel regression analysis—the fixed effects of system settings and top suggestion's characteristics.**

| Independent Variables | Dependent Variables: Hybrid System Responses | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Overall Quality | Δ Overall Quality | English Fluency | Δ English Fluency | Human-likeness | Δ Human-likeness | Time Spent $\log_{10}(y{+}1)$ | Response same as a suggestion (binary) |
| (Intercept) | 2.177 | 2.183 | 2.269 | 2.269 | 2.222 | 2.220 | 1.454 | −4.447 |
| RetWiki vs. RetBest (1=true;0=false) | −0.066 | −0.070 | 0.001 | 0.001 | −0.113 | −0.117 | 0.047 | −1.102 |
| RetWorst vs. RetBest (1=true;0=false) | −0.115 | −0.123 | −0.018 | −0.018 | −0.014 | −0.020 | 0.083 | −1.314 |
| Preset (1=true;0=false) | 0.066 | 0.071 | 0.038 | 0.038 | 0.025 | 0.027 | −0.078 | 0.742 |
| Showing 1 suggestion (1=true;0=false) | −0.006 | −0.007 | −0.005 | −0.005 | 0.104 | 0.105 | −0.060 | −0.430 |
| Showing 5 suggestions (1=true;0=false) | −0.029 | −0.028 | −0.041 | −0.041 | 0.005 | 0.006 | 0.004 | 0.266 |
| Top Suggestion's Overall Quality | 0.362 | −0.638 | 0.001 | 0.001 | −0.052 | −0.053 | −0.034 | 0.424 |
| Top Suggestion's English Fluency | 0.046 | 0.044 | 0.405 | −0.595 | 0.009 | 0.011 | −0.008 | 0.196 |
| Top Suggestion's Human-likeness | −0.006 | −0.006 | 0.009 | 0.009 | 0.403 | −0.598 | 0.007 | 0.024 |
| Top Suggestion's Length: $\log_{10}(x+1)$ | 0.070 | 0.069 | 0.059 | 0.059 | 0.064 | 0.065 | 0.050 | 0.969 |

Orange shading: positive coefficient significant at 0.05, 0.01, and 0.001 level. Blue shading: negative coefficient significant at 0.05, 0.01, and 0.001 level.

in different measures. In contrast, most of the improvements in systems using the two well-performing chatbots are small (< 0.1) and sometimes not statistically significant.

To conclude, we found that the hybrid systems can consistently enhance overall response quality than their chatbots' top results—not always have evident improvements, but rarely hurt on average. This suggests there is little risk of replacing chatbots with hybrid settings concerning response quality.

## 5.3 Influence of the Top Displayed Suggestion

We further examine the influence of three factors—chatbot suggestions, message box initialization, and the number of displayed suggestions—on the hybrid systems. Here we exclude the hybrid systems using Gen because they can only display one suggestion. The rest 18 settings (8,904 sessions in total) by RetWiki, RetBest, and RetWorst create a 3×3×2 factorial design—chatbot (3 levels), message box initialization (2 levels), and the number of suggestions (3 levels). We examine the three factors using a three-way ANOVA. Figure 6 reports the interaction between message box initialization and the number of displayed suggestions on the hybrid systems using the two well-performing chatbots.

Also, we suspect a chatbot choice is insufficient to characterize its suggestions' characteristics as the same chatbot's suggestions vary across different tasks. Thus, we also use regression analysis to examine the influence of chatbot suggestions' attributes on the hybrid systems. Here we only focus on the top displayed suggestion. We use multilevel regression instead of an ordinary one because our observations are nested—multiple finished conversations' responses (level 1) are nested within the same participant (level 2), and multiple participants are nested within the same experiment setting (level 3).

The level 1 independent variables include the top displayed suggestion's overall quality, English fluency, human-likeness, and length. We do not include level 2 variables to characterize the participants. The level 3 variables are the experiment setting factors. We use two dummy variables (RetWiki and RetWorst) for chatbot choice (RetBest is the reference category). We use two dummy variables (showing 1 or 5 suggestions) for the number of displayed suggestions because its effects do not seem linear in Figure 6. We use a "random slope" model to allow level 1 variables' effects to vary by participants (level 2) and experiment settings (level 3). The level 3 variables have fixed slopes. Table 3 reports the fixed effects of each independent variable on each dependent variable, and we label statistically positive and negative effects by different colors. The fixed effects can be interpreted similarly to the coefficients in an ordinary regression, except that the level 1 variables' effects have factored out the variation by different participants and experiment settings.

The effects of the level 1 variables in Table 3 confirmed our prior conjecture that the top displayed suggestion has a salient influence on response quality and composing behavior (after factoring out the experiment setting and individual differences).

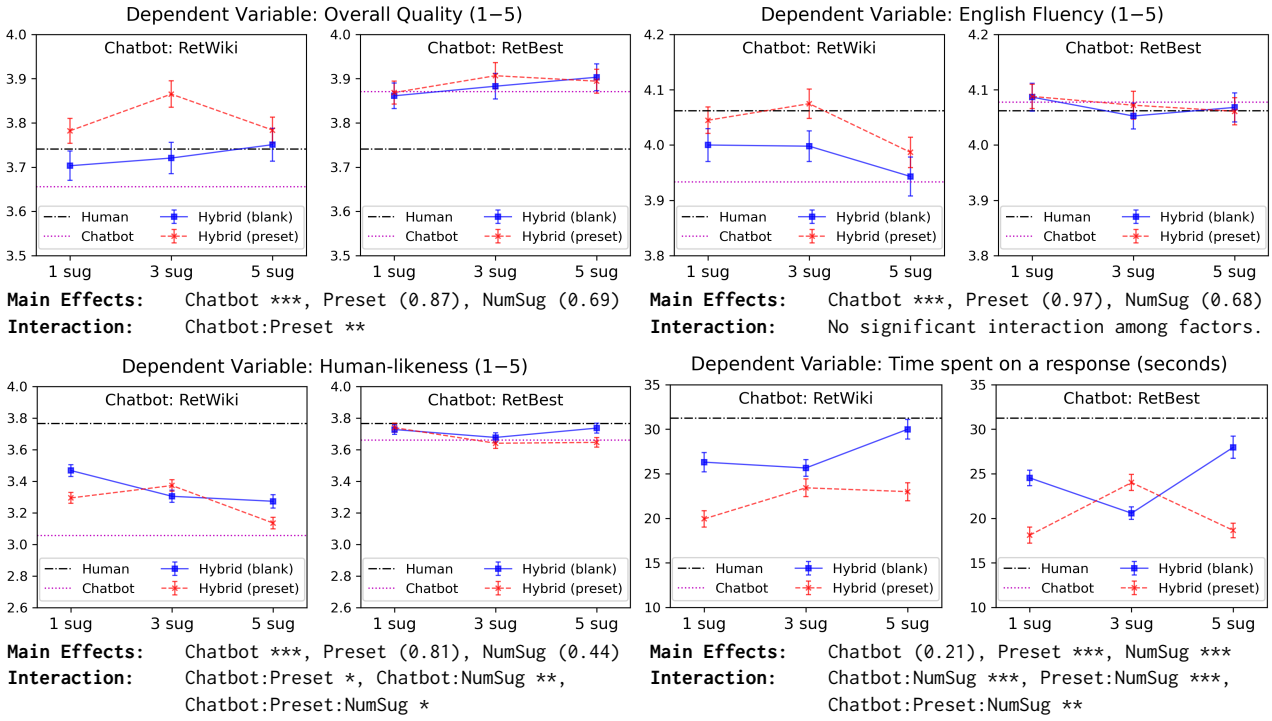
First, the top suggestion's quality on a particular dimension significantly and positively affects the final response's quality on the same dimension. For example, the model's coefficients suggest that, while other factors are equal, a 1-unit increase in the top suggestion's overall quality will enhance the hybrid system response's overall quality by 0.362. This indicates that providing better top suggestions can directly increase response quality in hybrid systems. This also explains why hybrid systems using well-performing chatbots have better quality responses than those using poor-performing bots.

Second, the top suggestion's quality significantly but negatively affects the quality improvement of hybrid systems' responses over the chatbot baselines (the dependent variables starting with Δ in Table 3). This explains why participants' responses in hybrid systems using different chatbots improved over the chatbot baselines by different magnitudes. Practically, the negative effects suggest that users may not improve a suggestion if it is already good enough.

Third, having a top suggestion with better overall quality makes users more prone to adopt a suggestion directly for the final response. The last column in Table 3, "response same as a suggestion (binary)," is a binary variable for whether the final response is the same as a displayed suggestion. The reported effects are the raw coefficients for the multilevel logistic regression. We found that the top suggestion's overall quality has a positive influence—increasing the top suggestion's overall quality will increase the chances of directly adopting a suggestion for the response.

Moreover, providing top suggestions with better overall quality also saves users' response time—we observed a negative effect of top suggestion's overall quality on time spent (log transformation). While other factors are equal, a 1-unit increase in the top suggestion's overall quality will reduce the $\log_{10}(x+1)$ transformation

**Figure 6: Interaction between message box initialization and the number of displayed suggestions in hybrid systems.**




value of the time spent by 0.034—this is roughly to reduce the time (more precisely, time + 1) to $10^{-0.034} = 92.5\%$ of the original value. We believe this is because users are more prone to adopt a chatbot suggestion with no changes or minor edits when the suggestion is good enough.

In addition to the top suggestion's quality, we found that the top suggestion's length (log transformation) significantly and positively affected the time spent on a response (log transformation). We suspect it is because longer suggestions take users longer to read. The practical implication is that while two chatbots provide suggestions with the same quality, a hybrid system may use the one that gives shorter suggestions due to user efficiency concerns. The top suggestion's length also significantly and positively affects the odds ratio to adopt a suggestion without changes—it requires future study to explain why.

To conclude, providing better-quality top suggestions in hybrid systems can improve response quality, encourage users to adopt suggestions, and reduce response time. Also, providing longer suggestions increases hybrid systems' response time.

## 5.4 Whether to Preset the Message Box

We found that presetting the message box with the chatbot's top suggestion can significantly reduce the time spent and encourage users to adopt a suggestion without changes. The coefficients in Table 3 suggest that, while other factors are equal, presetting the message box will shorten the time spent to roughly $10^{-0.078} = 84\%$ of that without presetting and increase the odds ratio of adopting a suggestion with no changes to $e^{0.742} = 210\%$ of the ratio without presetting the message box.

The influence of message box initialization on response quality seems inconsistent across ANOVA and multilevel regression results. ANOVA tests found no significant main effects but some significant interactions between message box initialization and chatbot choices. The multilevel regression suggests that presetting significantly and positively affects responses' overall quality and English fluency, but the practical impacts are limited based on the coefficients' values. Here we believe it requires future research to confirm whether presetting the message box affects response quality.

## 5.5 Number of Displayed Suggestions

Findings regarding the number of displayed suggestions are mostly inconclusive, suggesting that we need further work to understand its influence on hybrid systems. Figure 6 suggests that we should not explain the effects of the number of displayed suggestions as simply ordinal. In many cases, the trends from showing one suggestion to three are inconsistent with those from showing three to five. Besides, the ANOVA tests found some significant interactions between the number of displayed suggestions and other factors, suggesting its effects are dependent on other factors. The multilevel regression found significant effects on responses' English fluency and human-likeness, but we do not interpret them further. At least, our experimental results did not show evident benefits or drawbacks for providing more suggestions in hybrid systems.

# 6 DISCUSSION AND CONCLUSION

## 6.1 How and Why Chatbot Suggestions Help

Our study reveals that well-designed hybrid systems can improve both response quality and efficiency (but mainly efficiency) in question-answering conversations because of two reasons.

First, chatbot suggestions can provide critical relevant information for completing knowledge-demanding questions like those we used in our experiments. Also, as Figure 5 shows, it seems a natural need to seek information when answering knowledge-demanding questions, even though the human agents' role is to provide information. Hybrid systems' suggestions can address such information-seeking needs.

Second, users can efficiently reuse the chatbot suggestions' texts, in part or whole, when composing their responses. It seems that adopting a suggestion directly or making edits on top of it costs users much fewer keystrokes and a shorter time than writing a response entirely from scratch. We note that this observation holds in all hybrid systems we evaluated, even though some used poor-performing chatbots for suggestions.

However, the primary benefits of hybrid systems over human-only ones lie in their efficiency. According to our results, even the best-performing systems had only slightly improved response quality in some but not all dimensions. In contrast, hybrid systems had shortened response time consistently.

## 6.2 Hybrid System Design Suggestions

Our study provides suggestions for designing hybrid systems to support humans in answering informational questions.

First, providing high-quality suggestions is key to the benefits of a hybrid system. This may seem obvious, but what is striking is that our results showed how poor-performing chatbots' suggestions might hurt response quality in hybrid systems. There is likely a quality-efficiency trade-off between human-only and hybrid systems if we cannot guarantee the quality of chatbot suggestions. In a pilot study [17], we were not able to identify this risk of hybrid systems due to the limited chatbot choice we evaluated. It is essential to decide whether and when to suggest replies in hybrid systems.

Our study also provides insights regarding the choice of chatbots for hybrid systems. We have found that the generation-based chatbot did not perform well compared with the retrieval-based ones. However, we note that retrieval-based chatbot's actual effectiveness may lie between `RetBest` and `RetWorst`, and the comparison between retrieval and generation-based approaches may differ by datasets and models. Interestingly, `RetWiki` supported users rather well, even though previous studies did not extensively explore retrieval-based bots searching over non-conversational datasets. We also suspect that the criteria of whether a chatbot well supports a hybrid system may differ from those for responding to users alone.

Finally, our study suggests that one needs to be cautious about how many suggestions to display in hybrid systems. As our experimental results show, offering more chatbot suggestions did not consistently improve response quality or efficiency (though not consistently hurt either). Also, its significant interaction with other factors suggests that the optimal number of suggestions is highly system-dependent and may vary by other factors.

## 6.3 Limitations and Future Work

We also acknowledge some limitations in our study.

First and most important, we have only examined question-answering conversations for providing information. We made this choice due to the lack of conversational datasets of other types. The findings may vary in other types of conversations, especially communicative (e.g., an agent for amusing people) and transactional ones (e.g., online customer support helping a traveler change a flight). Also, our users are "answerers" in these tasks, and their role is to provide information to others. It is unclear how suggestions affect message quality for other roles of users.

Second, we have relied on crowdsourcing response quality judgments to evaluate the hybrid systems. It remains unclear how well the collected assessments and the six quality dimensions represent the experience of people who received these messages in a conversation. For example, we did not use dimensions such as empathy and ethics while judging the responses, although some previous studies suggested a need to make more empathetic and ethical chatbots [8, 15, 24, 39, 48]. We also expect that user experience with chatbot responses will vary a lot in conversations of various types.

Third, we used only a simple interaction design for our hybrid systems, e.g., we only provided static whole-message suggestions and did not interactively suggest while users type in the message box. It requires further work to examine how different interaction designs influence response quality and efficiency.

We suggest future work to understand the effectiveness of hybrid systems in a more natural interactive multi-round conversation setting, using more realistic tasks, and further understand the influence of the number of suggestions. Also, our experiments asked crowd workers to answer questions in a batch mode using hybrid systems. This suggests future work may explore if we can address similar applications (such as online customer services) using human computation combining hybrid systems and crowdsourcing.

## 6.4 Conclusion

This paper presents a large-scale crowdsourcing user study evaluating hybrid conversational systems that allow users to compose messages based on AI chatbot suggestions. We have specifically focused on using hybrid systems for supporting people to finish knowledge-demanding question-answering tasks. First, we found that hybrid systems primarily help users compose responses more efficiently—occasionally, they also help improve response quality slightly. Second, we illustrated the critical role of chatbot suggestion's characteristics in hybrid systems—we especially note that hybrid systems need to provide high-quality chatbot suggestions to ensure answer quality. Using hybrid systems with poor-performing chatbots may face a quality-efficiency tradeoff. Third, we also identified a few optimal design choices for hybrid systems regarding presetting the message box or not. Our study provided guidance and suggestions to design and deploy hybrid chat systems in similar scenarios effectively.

## ACKNOWLEDGMENT

We thank Naman Ahuja for helping with the experiments and the anonymous reviewers for their valuable comments and suggestions. Support for this research was provided by the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

# REFERENCES


[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 475–484. https://doi.org/10.1145/3331184.3331265

[2] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. https://doi.org/10.1145/3411764.3445372

[3] Barbara M. Byrne. 2016. *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming* (3 ed.). Routledge.

[4] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail Smart Compose: Real-Time Assisted Writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2287–2295. https://doi.org/10.1145/3292500.3330723

[5] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*. Association for Computational Linguistics, Brussels, Belgium, 2174–2184. https://doi.org/10.18653/v1/D18-1241

[6] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 815–824. https://doi.org/10.1145/2939672.2939746

[7] Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 729–738. https://doi.org/10.1145/3357384.3358016

[8] Stephan Diederich, Max Janssen-Müller, Alfred Benedikt Brendel, and Stefan Morana. 2019. Emulating Empathetic Behavior in Online Service Encounters with Sentiment-Adaptive Responses: Insights from an Experiment with a Conversational Agent. In *Proceedings of the 40th International Conference on Information Systems (ICIS 2019)*.

[9] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations (ICLR 2019)*. https://openreview.net/forum?id=r1l73iRqKm

[10] Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. A dataset and baselines for sequential open-domain question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*. Association for Computational Linguistics, Brussels, Belgium, 1077–1083. https://doi.org/10.18653/v1/D18-1134

[11] Mihail Eric and Christopher Manning. 2017. A Copy-Augmented Sequence-to-Sequence Architecture Gives Good Performance on Task-Oriented Dialogue. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (EACL '17)*. Association for Computational Linguistics, Valencia, Spain, 468–473. https://www.aclweb.org/anthology/E17-2075

[12] Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2017. Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *Proceedings of the 38th International Conference on Information Systems (ICIS 2017)*.

[13] Joseph F. Hair, Bill Black, Barry J. Babin, and Rolph E. Anderson. 2010. *Multivariate Data Analysis: Global Edition* (7 ed.). Pearson Education.

[14] Jess Hohenstein and Malte Jung. 2018. AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188487

[15] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch Your Heart: A Tone-Aware Chatbot for Customer Care on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173989

[16] Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based Neural Structured Learning for Sequential Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '17)*. Association for Computational Linguistics, Vancouver, Canada, 1821–1831. https://doi.org/10.18653/v1/P17-1167

[17] Jiepu Jiang and Naman Ahuja. 2020. Response Quality in Human-Chatbot Collaborative Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1545–1548. https://doi.org/10.1145/3397271.3401234

[18] Jussi Jokinen, Aditya Acharya, Mohammad Uzair, Xinhui Jiang, and Antti Oulasvirta. 2021. Touchscreen Typing As Optimal Supervisory Control. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 720, 14 pages. https://doi.org/10.1145/3411764.3445483

[19] Dan Jurafsky and James H. Martin. 2020. Dialogue Systems and Chatbots (Draft of December 30, 2020). In *Speech and Language Processing (3rd ed. draft)*. Chapter 26. https://web.stanford.edu/~jurafsky/slp3/26.pdf

[20] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart Reply: Automated Response Suggestion for Email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 955–964. https://doi.org/10.1145/2939672.2939801

[21] Diane Kelly, Karl Gyllstrom, and Earl W. Bailey. 2009. A Comparison of Query and Term Suggestion Features for Interactive Searching. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 371–378. https://doi.org/10.1145/1571941.1572006

[22] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP '16)*. Association for Computational Linguistics, Austin, Texas, 1192–1202. https://doi.org/10.18653/v1/D16-1127

[23] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP '16)*. Association for Computational Linguistics, Austin, Texas, 2122–2132. https://doi.org/10.18653/v1/D16-1230

[24] Mingming Liu, Qicheng Ding, Yu Zhang, Guoguang Zhao, Changjian Hu, Jiangtao Gong, Penghui Xu, Yu Zhang, Liuxin Zhang, and Qianying Wang. 2020. Cold Comfort Matters - How Channel-Wise Emotional Strategies Help in a Customer Service Chatbot. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3382905

[25] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2015)*. Association for Computational Linguistics, Prague, Czech Republic, 285–294. https://doi.org/10.18653/v1/W15-4640

[26] Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A Dialog Research Software Platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP '17)*. Association for Computational Linguistics, Copenhagen, Denmark, 79–84. https://doi.org/10.18653/v1/D17-2014

[27] Alexandra Olteanu, Fernando Diaz, and Gabriella Kazai. 2020. When Are Search Completion Suggestions Problematic? *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 171 (Oct. 2020), 25 pages. https://doi.org/10.1145/3415242

[28] Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and Characterizing User Intent in Information-Seeking Conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 989–992. https://doi.org/10.1145/3209978.3210124

[29] Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. 2019. User Intent Prediction in Information-Seeking Conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 25–33. https://doi.org/10.1145/3295750.3298924

[30] Philip Quinn and Shumin Zhai. 2016. A Cost-Benefit Study of Text Entry Suggestion Interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 83–88. https://doi.org/10.1145/2858036.2858305

[31] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 117–126. https://doi.org/10.1145/3020165.3020183

[32] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks. In *International Conference on Learning Representations (ICLR '16)*. https://arxiv.org/abs/1511.06732

[33] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (March 2019), 249–266. https://doi.org/10.1162/tacl_a_00266

[34] Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Edinburgh, Scotland, UK., 583–593. https://www.aclweb.org/anthology/D11-1054

[35] Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021. "I Can't Reply with That": Characterizing Problematic Email Reply Suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 724, 18 pages. https://doi.org/10.1145/3411764.3445557

[36] Quentin Roy, Sébastien Berlioux, Géry Casiez, and Daniel Vogel. 2021. Typing Efficiency and Suggestion Accuracy Influence the Benefits and Adoption of Word Suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 714, 13 pages. https://doi.org/10.1145/3411764.3445725

[37] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of Natural Language Rules in Conversational Machine Reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*. Association for Computational Linguistics, Brussels, Belgium, 2087–2097. https://doi.org/10.18653/v1/D18-1233

[38] Amrita Saha, Vardaan Pahuja, Mitesh Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex Sequential Question Answering: Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph *(AAAI '18)*. AAAI Press, 705–713. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17181

[39] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173889

[40] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL '19)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1702–1723. https://doi.org/10.18653/v1/N19-1170

[41] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3104–3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[42] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 32–41. https://doi.org/10.1145/3176349.3176387

[43] Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a model for spoken conversational search. *Information Processing & Management* 57, 2 (2020), 102162. https://doi.org/10.1016/j.ipm.2019.102162

[44] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring Conversational Search With Humans, Assistants, and Wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2187–2193. https://doi.org/10.1145/3027063.3053175

[45] Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A Dataset for Research on Short-Text Conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP '13)*. Association for Computational Linguistics, Seattle, Washington, USA, 935–945. https://www.aclweb.org/anthology/D13-1096

[46] Jason D. Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language* 21, 2 (2007), 393 – 422. https://doi.org/10.1016/j.csl.2006.06.008

[47] Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2019. A Sequential Matching Framework for Multi-Turn Response Selection in Retrieval-Based Chatbots. *Computational Linguistics* 45, 1 (March 2019), 163–197. https://doi.org/10.1162/coli_a_00345

[48] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3506–3510. https://doi.org/10.1145/3025453.3025496

[49] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 55–64. https://doi.org/10.1145/2911451.2911542

[50] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response Ranking with Deep Matching Networks and External Knowledge in Information-Seeking Conversation Systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 245–254. https://doi.org/10.1145/3209978.3210011

[51] Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The Hidden Information State model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language* 24, 2 (2010), 150 – 174. https://doi.org/10.1016/j.csl.2009.04.001

[52] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating Clarifying Questions for Information Retrieval. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 418–428. https://doi.org/10.1145/3366423.3380126

[53] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '18)*. Association for Computational Linguistics, Melbourne, Australia, 2204–2213. https://doi.org/10.18653/v1/P18-1205

[54] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 177–186. https://doi.org/10.1145/3269206.3271776

[55] Tiancheng Zhao and Maxine Eskenazi. 2016. Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2016)*. Association for Computational Linguistics, Los Angeles, 1–10. https://doi.org/10.18653/v1/W16-3601