SearchBots: User Engagement with ChatBots during Collaborative Search

Sandeep Avula, Gordon Chadwick, Jaime Arguello, Robert Capra
University of North Carolina at Chapel Hill
asandeep@live.unc.edu, gchadwick10@gmail.com, {jarguello, rcapra}@unc.edu

ABSTRACT

Popular messaging platforms such as Slack have given rise to hundreds of chatbots that users can engage with individually or as a group. We present a Wizard of Oz study on the use of searchbots (i.e., chatbots that perform specific types of searches) during collaborative information-seeking tasks. Specifically, we study searchbots that intervene dynamically and compare between two intervention types: (1) the searchbot presents questions to users to gather the information it needs to produce results, and (2) the searchbot monitors the conversation among the collaborators, infers the necessary information, and then displays search results with no additional input from the users. We investigate three research questions: (RQ1) What is the effect of a searchbot (and its intervention type) on participants’ collaborative experience? (RQ2) What is the effect of a searchbot’s intervention type on participants’ perceptions about the searchbot and level of engagement with the searchbot? and (RQ3) What are participants’ impressions of a dynamic searchbot? Our results suggest that dynamic searchbots can enhance users’ collaborative experience and that the intervention type does not greatly affect users’ perceptions and level of engagement. Participants’ impressions of the searchbot suggest unique opportunities and challenges for future work.

ACM Reference Format:

Keywords: Chatbots; Collaborative Search; Intelligent Agents

1 INTRODUCTION

Messaging platforms such as Slack, Yammer, and Facebook Workplace have become commonplace in work environments, allowing distributed workers to communicate and collaborate on shared tasks. These new platforms are different from traditional chat interfaces in their aesthetics and ability to integrate with workplace collaboration technologies such as Github and Google Drive. The increasing popularity of such messaging platforms has also inspired the development of hundreds of third-party chatbots. User can engage these chatbots in dialogs to accomplish specific tasks such as to update a social media status or to schedule a meeting with a group of collaborators. Additionally, there are chatbots that search for information on specific topics such as news, restaurants, and weather. Users engage with these chatbots by sending specific requests (e.g., "@weatherbot New York City"), answering optional follow-up questions required by the chatbot, and then interacting with the search results provided. Much of the consideration of chatbots has focused on their use by individuals to accomplish personal tasks or to search for information. However, chatbots are also well-positioned to help groups of users who are working collaboratively on tasks that involve searching for information.

Research on collaborative search has sought to understand how people collaborate during tasks that involve searching for information and to develop tools to support such collaborations. The most prominent approach has been to develop dedicated systems for collaborative search [4, 12, 26, 28, 32, 37]. These systems have been designed with the search engine as the centerpiece component, but include additional features that allow users to communicate, share information, and become aware of each other’s search activities.

A key finding from research on real-world collaborative search practices is that while people often search in groups, they do so without the use of dedicated collaborative search systems. Instead, they search independently and coordinate using communication tools such as instant messaging, social media, email, and phone [5, 6, 24, 25]. Morris [25] and Hearst [14] highlighted these findings as a rationale to develop lightweight tools for collaborative search that are directly integrated with existing communication platforms. Our research in this paper is an answer to this call.

Little prior research has investigated how to integrate collaborative search functions into an existing messaging channel and there are many open questions about how best to do so. In this paper, we present a foray into this design space—we investigate the use of searchbots (i.e., chatbots that perform specific types of searches) during collaborative information-seeking tasks facilitated through Slack. Specifically, we investigate searchbots that intervene dynamically in the conversation in order to provide contextually relevant search results, and focus on two types of interventions: (1) the searchbot intervenes and elicits the information it needs in order to produce search results using a scripted dialogue, and (2) the searchbot intervenes and directly produces search results by “inferring” the information it needs from the ongoing conversation.

Two additional findings from prior research in collaborative search motivate the study of dynamic searchbots within messaging platforms such as Slack. First, studies have found that chat-based communication is an extremely common activity during collaborative search [37, 38]. Second, studies have found that collaborators often chat about what they are going to search for before actually doing so [37, 38]. Oftentimes, this is done to support strategies...
such as division of labor or to maintain collaborative awareness of each other’s plans and actions. In situations where collaborators are willing to let a searchbot monitor their chat, there are opportunities for a searchbot to infer information from the conversation and dynamically contribute.

We report on a Wizard of Oz study that investigates the following three research questions:

**RQ1:** In our first research question, we investigate the effects of a searchbot (and its intervention type) on participants’ collaborative experience. We address this question from two perspectives. First, we focus on participants’ self-reported perceptions about the collaboration, such as the level of awareness of each other’s activities, level of effort, and level of enjoyment. Second, we focus on objective measures of collaborative effort, such as task completion time, number of messages exchanged, and number of URLs exchanged, which suggests out-of-channel searching and sharing.

**RQ2:** In our second research question, we investigate the effects of the searchbot’s type of intervention of participants’ perceptions about the searchbot and their level of engagement with the searchbot. As in RQ1, we address this question from two perspectives. First, we focus on participants’ self-reported perceptions, such as their level of annoyance with the intervention, their confidence in the searchbot’s ability to help, and their gains obtained from the searchbot. Second, we focus on participants’ decisions to engage (or not engage) with the searchbot’s results.

**RQ3:** In our third research question, we investigate participants’ general impressions about the searchbot. To address this question, we analyzed participants’ responses to two open-ended questions: (1) If the searchbot helped you, how? and (2) If the searchbot did not help you, why not?

## 2 RELATED WORK

Our work builds on two areas of prior research: (1) collaborative search and (2) dynamic help systems and interruptions.

**Collaborative Search:** Collaborative search happens when multiple people work together on an information-seeking task. Collaborative search is often investigated with two dimensions in mind: time and space. The time dimension focuses on whether the collaboration happens synchronously or asynchronously, while the space dimension focuses on whether the collaborators are co-located or remote. A large body of prior work has focused on understanding collaborative search practices along these two dimensions [24, 25, 34, 35]. In this paper, we focus on synchronous collaborative search in situations where the collaborators can only communicate via the Slack messaging platform.

A number of different systems have been developed to support collaborative search, including SearchTogether [26], CoSense [28], Coagmento [32], CollabSearch [37], Querium [12], and ResultsSpace [4]. These systems have been designed with the traditional search engine as the centerpiece component, but include additional features for collaborators to communicate, share information, and become aware of each other’s search activities. The goal of these additional features is to allow collaborators to coordinate, learn from each other’s search paths, avoid duplicating work, and to assist with collaborative sensemaking—becoming aware of collaborators’ motivations, actions, and state of knowledge [20, 26]. Systems have also been designed to algorithmically alter the ranking of documents based on collaborators’ activities, for example, by using documents shared between collaborators as a form of relevance feedback [30].

Studies have found that these specialized systems provide different benefits during collaborative search, for example, by improving the collaborative experience compared to non-integrated tools [26], by raising the awareness of collaborators’ activities [28]; by supporting different strategies adopted by the group (e.g., agreeing on a few relevant items vs. being as exhaustive as possible) [4]; and by reducing communication and coordination efforts [33].

While many different systems have been developed to support collaborative search, these systems have not enjoyed wide-spread use [14]. A survey by Morris [25] found that while collaborative search has become increasingly common, most people use a combination of everyday search and communication technologies to collaborate on search tasks. Morris concluded by suggesting that integrating lightweight search tools into existing communication channels may be a more promising approach than developing dedicated systems for collaborative search.

Prior research has found that people often use social networks such as Facebook and Twitter to engage in asynchronous collaborative search, an activity referred to as social search [11, 27]. Efron and Winget [9] developed a taxonomy of questions posted on Twitter, and found that a large proportion request factual information that is likely to exist on the Web. This result suggests the possibility of developing search systems that can automatically respond to questions posted on social media and partly motivated the development of the SearchBuddies system [15]. SearchBuddies was designed to embed search results in response to questions posted on Facebook. The embedded search results appeared as a new post in the Facebook thread. A qualitative analysis of people’s perceptions found interesting challenges and opportunities for “socially-embedded search engines”. For example, users only reacted positively to the embedded search results when they were extremely relevant and non-obvious, or when they complemented another user’s answer to the question. To our knowledge, no prior work has investigated how people perceive search systems that intervene in synchronous instant messaging conversations.

**Dynamic Help Systems and Interruptions:** Prior research has investigated the reasons why people avoid systems that intervene to provide assistance. Users avoid help systems due to the cost of cognitively disengaging with the primary task, due to the fear of unproductive help-seeking, due to a failure to admit defeat, or because they are unaware of how the help system can provide support [8, 18].

An unwanted intervention can be viewed as an interruption. A large body of research has also focused on understanding how people respond to interruptions while engaged in a task (see Li et al. [22] for a review). Studies have found that interruptions can negatively affect task performance [2], cognitive load [16], and emotional state [1]. Research on interruptions has focused on three dimensions: the interruption protocol, timing, and relevance. Early work by McFarlane [23] investigated four interruption protocols: immediate, negotiated, mediated, and scheduled. Negotiated interruptions, which provide mechanisms for easily ignoring the interruption, were the most effective. A wide range of studies have focused on the timing of an interruption. Results consistently show that interruptions during periods of low mental workload are less
disruptive. In this respect, studies have found that interruptions are less disruptive when they occur early in the task (before the user is deeply engaged) [7] and during sub-task transitions [1, 16, 17]. Finally, studies have found that interruptions that are more relevant to the primary task are less disruptive [7, 17].

Most research on interruptions has focused on interrupting individuals, rather than collaborators working on a common task. As one exception, Peters et al. [29] investigated interruptions aimed at one individual while collaborating with another. This study compared interruptions sent at random intervals versus interruptions sent by a human “wizard” monitoring the communication channel. The wizard’s interruptions were less disruptive, suggesting that a system with access to the communication channel might be able to predict when to intervene.

3 USER STUDY
To investigate our three research questions, we conducted a Wizard of Oz laboratory study with 27 pairs of participants (34 female and 20 male). Participants were undergraduate students and were recruited in pairs. Each pair of participants collaborated on four tasks that required searching for information (Section 3.2) and were exposed to three searchbot conditions (Section 3.3). Participants used the Slack messaging system to communicate and were also provided with a Google Chrome browser in order to perform any desired searches. Similar to the protocol used by Morris and Horvitz [26], participants were seated in the same room, but did not face each other and were not to communicate outside of Slack.

3.1 Study Protocol
Before starting the experiment, the moderator outlined the study protocol, described Slack, and described the basic functionality of a searchbot. Searchbots that intervene do not currently exist in messaging platforms such as Slack. Thus, we believed it was important to explain how searchbots work to our participants. Participants were told that a searchbot is an interactive agent that may intervene in a Slack conversation to provide search results after possibly asking some questions. Participants were told that searchbots embed search results directly in the chat window and also provide a “click here for more” hyperlink that opens a pop-up browser window. Also, participants were told that the searchbots used in the study were not designed to respond to explicit requests and could only accept input in response to a searchbot-initiated question. Following these explanations, participants interacted with a simple “weatherbot” (designed by us) in order to familiarize themselves with interacting with a searchbot.

As described in more detail below, participants were assigned four tasks that required them to search for information and coordinate towards a solution. Additionally, for each task, each participant was given one “personal preference” they should try to satisfy. Each participant’s “personal preference” was not known to the other participant. The purpose of these preferences was to emulate a common situation in collaborative search in which collaborators have individual constraints that need to be expressed and accounted for in the final solution. Participants were told that they could use whatever means necessary to search for information—they could interact with the searchbot and/or conduct their own searches.

To familiarize participants with the task format, the first task was always a practice task. Participants were asked to choose three movies they would like to watch together over the weekend. Before doing so, participants were asked to write down a personal preference they would like to satisfy (e.g., I would like to watch a horror movie.). Participants were then asked to complete the practice task by communicating through Slack and searching on their own (no searchbot intervened). The practice task was the only one in which the participants chose their own personal preference. After completing each of the four tasks, participants were asked to complete a post-task questionnaire (Section 3.4). Each participant was given $20 USD for participating in the study. We used Camtasia software to capture participants’ screen activity and the Slack API to record all their activity inside of Slack. Additionally, we logged all clicks on the searchbot’s results.

3.2 Search Tasks
Participants completed three search tasks in addition to the first practice task: (1) a restaurant-finding task, (2) a local attractions-finding task, and (3) a book-finding task. Each task had a “background story” and asked participants to search for and agree on three different items. Participants were given gender-neutral first names (Jamie and Taylor). For example, the restaurant-finding task had the following background story and objective.

Background/Objective: Jaime and Taylor went to grad school together in Boulder, Colorado which is about 45 minutes outside of Denver. After graduation, Jamie moved to Denver and Taylor moved to Phoenix. Their new lives have become very hectic, which makes it difficult to keep in touch. However, they are determined to change this because Taylor is coming to Denver for a professional conference. Taylor messaged Jamie regarding meeting soon. Your goal is to pick three potential restaurants to get food.

Additionally, for each task, each participant was given a personal preference they should try to satisfy during the task. For example, for the restaurant-finding task, one participant (Jamie) was given a constraint on the type of food and the other participant (Taylor) was given a constraint about the location:

Food constraint: You recently made a life choice to go vegan. To keep yourself in line with this new lifestyle, you have decided to only eat at restaurants that provide good vegan options.

Location constraint: You live in downtown Denver (in Capitol Hill). You just sold your car and have been mostly commuting by bike. Since your workplace is about two miles from where you live, you don’t have a strong urge to buy a new car immediately. For this reason, you currently like to meet people downtown (in Capitol Hill) and not go anywhere else.

The personal constraints associated with each task were: restaurants task—location, food preference; local attractions task—location, attraction type; books task—fiction vs. non-fiction, sub-genre. For the two tasks involving location constraints, participants were provided with the constraint description (e.g., as shown above) as well as a map illustrating the location of interest. Different cities were used for these two tasks to avoid learning effects.

3.3 Searchbot Conditions
We custom-designed three different searchbots to match our three tasks: (1) a searchbot for local restaurants, (2) a searchbot for local attractions, and (3) a searchbot for books. Each searchbot required two key attributes in order to produce search results. The two key attributes required by each searchbot were designed to match the personal constraints given to participants for the corresponding
The searchbot intervened at a point that was consistent (i.e., a point of low cognitive load) in which participants would be immediately aware of the information, but does not learn from the conversation and must explicitly request what it needs in order to produce results. The goal of the bot_q condition was to mimic a searchbot that is able to intervene and provide relevant information, but does not learn from the conversation and must explicitly request what it needs in order to produce results. The goal of the bot q condition was to mimic a searchbot that is able to “learn” from the conversation and directly provide contextually relevant results.

The results provided by the searchbot in the bot q and bot auto conditions were exactly the same for each task. In other words, in the bot q condition, the search results provided by the searchbot were the same regardless of participants’ responses to the searchbot’s questions. All three searchbots returned 15 results that were pre-fetched from Google Maps for the restaurant-finding and local attractions-finding tasks, and from Goodreads for the book-finding task. Similarly, all three searchbots embedded the top-three results directly into the Slack window and provided a “click here for more” link that opened a pop-up window with all 15 results. For the restaurants and local attractions tasks, the pop-up window also included an interactive map with the search results displayed.

Figures 1a-1c illustrate the look and feel of the local attractions searchbot. Figure 1a illustrates the searchbot’s intervention in the bot q and bot auto conditions. Figure 1b illustrates the searchbot’s top-three results that were displayed inside of Slack and were visible to both participants. As shown, the results were always followed by a “click here for more” link that opened a pop-up browser window with all the search results (referred to as the landing page). Figure 1c illustrates the landing page, which always included 15 items.

The searchbot was operated by a “Wizard” who had access to the participants’ Slack channel and was sitting in a different room. The role of the Wizard was to monitor the conversation and always intervene immediately after both participants mentioned their personal preferences in the conversation. By using this point of intervention, we achieved three goals: (1) we maintained a consistent point of intervention between the bot q and bot auto conditions, (2) we used a realistic point of intervention for the bot auto condition (a point in which a searchbot would be able to infer the necessary information to produce search results), and (3) we created a situation in the bot q condition in which participants might perceive the searchbot as having “missed” information that would have enabled it to directly produce relevant results. Additionally, we believe that this point in the conversation might often mark a sub-task-transition point (i.e., a point of low cognitive load) in which participants would be less disrupted by the intervention [1, 16, 17]. It should be noted that it was possible for participants to not "trigger" the searchbot in the bot_q and bot_auto conditions if they did not mention their personal preferences in the conversation.

Our experimental design involved three search tasks and three searchbot conditions. Each participant pair completed three tasks, with each task combined with one of the searchbot conditions. We used separate Latin Squares to counterbalance the presentation order of the tasks (3 orders) and of the searchbot conditions (3 orders), and then included all 9 combinations of these in our design. Thus, across our 27 participant pairs, the 9 treatment orders were each repeated 3 times.

### 3.4 Post-task Questionnaire

After completing each of the four tasks, individual participants were asked to complete a post-task questionnaire that had two parts. The first part asked about the participants’ collaborative experience and was always given to participants. Specifically, we asked questions about the level of collaborative awareness, effort, and enjoyment (Table 1). The second part of the post-task questionnaire asked about participants’ experience with the searchbot and was only given to participants in the bot q and bot auto conditions if they actually “triggered” the searchbot during the task by mentioning their personal preferences. Specifically, we asked about the searchbot’s point and manner of intervention, the participant’s confidence in the searchbot’s results, and the gains obtained from the searchbot (Table 2). Additionally, we asked two open-ended (and optional) questions about the searchbot: (1) “If the searchbot helped you in the task, briefly explain how.” and (2) “If the searchbot did not help you in the task, briefly explain why not.” Excluding the open-ended questions, all questions were asked using agreement statements with a 7-point scale with labeled endpoints (strongly disagree (1) to strongly agree (7)).

<table>
<thead>
<tr>
<th>Theme</th>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Awareness</td>
<td>aware_browser</td>
<td>During the task, I had a pretty good idea about the information my partner was looking at.</td>
</tr>
<tr>
<td></td>
<td>aware_myprefs</td>
<td>During the task, I was confident that my partner was looking at information that satisfied my own preferences.</td>
</tr>
<tr>
<td></td>
<td>aware_pyprefs</td>
<td>During the task, I was confident that I was looking at information that would satisfy my partner’s preferences.</td>
</tr>
<tr>
<td>Effort</td>
<td>ease_share</td>
<td>It was easy to share information with my partner during the task.</td>
</tr>
<tr>
<td></td>
<td>ease_coord</td>
<td>It was easy for my partner and I to coordinate our search efforts during this task.</td>
</tr>
<tr>
<td></td>
<td>ease_comm</td>
<td>It was easy to communicate my preferences with my partner during this task.</td>
</tr>
<tr>
<td></td>
<td>ease_cons</td>
<td>It was easy for my partner and I to reach consensus during this task.</td>
</tr>
<tr>
<td>Enjoyment</td>
<td>enjoy_me</td>
<td>I enjoyed completing this task.</td>
</tr>
<tr>
<td></td>
<td>enjoy_part</td>
<td>I think my partner enjoyed completing this task.</td>
</tr>
</tbody>
</table>

Table 2: Post-task questions about the searchbot.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention</td>
<td>distracting</td>
<td>The searchbot intervened at a point that was distracting. When it intervened, the searchbot asked us questions that were annoying.</td>
</tr>
<tr>
<td></td>
<td>annoying</td>
<td>When I first saw the information provided by the searchbot, I was confident that it would be useful.</td>
</tr>
<tr>
<td>Confidence</td>
<td>cond_results</td>
<td>The searchbot saved me and my partner some time. The searchbot provided us with useful information.</td>
</tr>
<tr>
<td></td>
<td>saved_time</td>
<td>The information provided by the searchbot gave me ideas about things to search for on my own.</td>
</tr>
<tr>
<td></td>
<td>useful_info</td>
<td>The searchbot helped me to discover new information.</td>
</tr>
<tr>
<td></td>
<td>discovery_info</td>
<td>The searchbot provided us with useful information.</td>
</tr>
<tr>
<td></td>
<td>ideas</td>
<td>The searchbot helped me to discover new information.</td>
</tr>
</tbody>
</table>
3.5 Data Analysis
In our experimental design, each participant pair had an opportunity to experience all three searchbot conditions: no_bot, bot_q, and bot_auto (i.e., a within-subjects design). However, it was possible for participants in the bot_q and bot_auto conditions not to “trigger” the searchbot if they did not mention their personal preferences in the conversation. This happened for five sessions in the study (out of 27 × 2 = 54 sessions). For these sessions, the participants experienced the no_bot condition. Thus, our three searchbot conditions were not equally balanced among participant pairs. To account for this, we used linear mixed-effects regression models in our analyses rather than repeated measures ANOVAs. Mixed-effects models are well-suited for imbalanced, repeated measures data [13]. Also, by using mixed-effect models, we were able to account for random effects due to variations at the participant-pair level (n = 27) and at the participant level (n = 54). We tested the significance of our mixed-effects overall models by computing the \( \chi^2 \) statistic using a likelihood-ratio test against a null model (i.e., one without the searchbot condition as a co-variate).

4 RESULTS
Before presenting results for our three research questions, we report on the overall engagement with the searchbot. There were 54 search sessions (27 × 2) in which participants could trigger the searchbot in either the bot_q or bot_auto condition. Participants triggered the searchbot in 49 out of 54 sessions (90%). Of these, there were 37 sessions (70%) where at least one participant clicked on a searchbot result. These 37 sessions were almost equally divided between the bot_q and bot_auto conditions (20 and 17, respectively).

From this preliminary analysis, we can conclude that engagement with the searchbot was fairly high. Additionally, at least in terms of this binary measure of engagement (“interacted” vs. “did not interact”), engagement with the searchbot was roughly equal in the bot_q and bot_auto conditions. In Section 4.2, we revisit the differences between participants’ engagement with the searchbot in the bot_q and bot_auto conditions (RQ2).

4.1 RQ1: Participants’ collaborative experience
To address our first research question, we analyzed participants’ responses to the post-task questionnaire section about their collaborative experience, and we also analyzed several measures related to their collaborative effort.

4.1.1 Post-Task perceptions about the collaborative experience.
First, we analyze participants’ responses to the post-task questions about their collaborative experience. As described in Table 1, these questions focused on three main themes: (1) awareness of each other’s activities, (2) effort, and (3) enjoyment. Figure 2 shows the mean of participants’ responses for each question across all three searchbot conditions. To analyze the effect of the searchbot condition on participants’ responses, we used a linear mixed-effects model (LMM) with nested random effects. The participant id was nested within the participant-pair id. We ran analyses using both the no_bot and bot_q conditions as the baseline to test for differences between all pairs of searchbot conditions.

**Awareness:** Of the three post-task questions about awareness, we found a **marginally** significant effect of searchbot condition on aware_browse (\( \chi^2(2) = 5.53, p = 0.06 \)). There were significant differences between the bot_auto and no_bot conditions (\( \beta = 0.56, S.E. = 0.26, p < 0.05 \)) and between the bot_auto and bot_q conditions (\( \beta = 0.59, S.E. = 0.28, p < 0.05 \)), with participants reporting greater awareness of their partner’s browsing activities in the bot_auto condition. Searchbot condition was not a significant predictor for the other two awareness measures (aware_myprefs and aware_pprefs). That said, there was about a 0.5 point difference between the bot_auto and no_bot conditions for aware_pprefs, with participants reporting greater confidence that they were considering alternatives that would satisfy their partner’s preferences in the bot_auto condition.

**Effort:** Of the four post-task questions about effort, we found significant effects of searchbot condition on ease_share (\( \chi^2(2) = 12.00, p < 0.01 \)), ease_comm (\( \chi^2(2) = 6.03, p < 0.05 \)), and ease_cons (\( \chi^2(2) = 10.45, p < 0.01 \)). In terms of ease_share, participants reported a greater ease in sharing information in the bot_q condition.
In terms of ease_comm, participants reported a greater ease in communicating their preferences with their partner in the bot_auto condition (β = 0.42, S.E. = 0.17, p < 0.05) as compared to the no_bot condition. Similarly, in terms of ease_cons, participants reported a greater ease in reaching consensus with their partner in the bot_auto condition (β = 0.63, S.E. = 0.19, p < 0.01) as compared to the no_bot condition.

Enjoyment: Searchbot condition was a significant predictor for both questions about enjoyment: enjoy.me (χ²(2) = 12.80, p < 0.01) and enjoy.part (χ²(2) = 11.95, p < 0.01). In terms of enjoy.me, participants reported greater levels of enjoyment during the task in the bot_q condition (β = 0.54, S.E. = 0.17, p < 0.01) and the bot_auto condition (β = 0.60, S.E. = 0.18, p < 0.01) as compared to the no_bot condition. In terms of enjoy.part, participants reported that they perceived their partner to have enjoyed the task more in the bot_q condition (β = 0.44, S.E. = 0.15, p < 0.01) and the bot_auto condition (β = 0.47, S.E. = 0.15, p < 0.01) as compared to the no_bot condition.

4.1.2 Measures of Collaborative Effort. In addition to analyzing participants’ perceptions about their collaboration, we also computed several measures associated with the level of collaborative effort expended during the task. We focused our analysis on four measures: (1) task completion time (in seconds), (2) number of messages exchanged, (3) average message length (in words), and (4) number of URLs exchanged between participants. Figures 3a-3d show the mean value of these measures across all three searchbot conditions. To analyze the effect of the searchbot condition on these measures, we used a linear mixed-effects model (LMM) with participant-pair id as a random effects variable. We ran analyses using both the no_bot and bot_q conditions as the baseline to test for differences between all pairs of searchbot conditions.

Searchbot condition was a marginally significant predictor of the number of URLs exchanged between participants (χ²(2) = 5.65, p = 0.06). Participants exchanged a greater number of URLs in the no_bot condition as compared to the bot_q condition (β = −0.49, S.E. = 0.21, p < 0.05). This trend was also present for the bot_auto condition, but did not reach significance. Searchbot condition was not a significant predictor for the other three measures. As one might expect, participants in the no_bot condition were forced to search independently “out of channel” and had to share their findings by copy/pasting URLs via Slack.

4.2 RQ2: Searchbot perceptions & engagement

To address our second research question, we analyzed participants’ responses to the post-task questionnaire section about the searchbot, and we also analyzed two measures related to their level of engagement with the searchbot.

4.2.1 Post-Task perceptions about the searchbot. First, we analyze participants’ responses to the post-task questions about the searchbot, which were given to participants during the 49 sessions in which the searchbot was triggered. As described in Table 2, these post-task questions focused on three main themes: (1) perceptions about the searchbot’s intervention, (2) confidence in the usefulness of the searchbot’s results, and (3) gains obtained from the searchbot. Figure 4 shows the mean of participants’ responses for each question in the bot_q and bot_auto conditions. To analyze the effect of the searchbot’s intervention type on participants’ responses, we used a linear mixed-effects model (LMM) with nested random effects. The participant id was nested within the participant-pair id. The bot_q condition was used as the baseline.

Searchbot condition was not a significant predictor of participants’ responses for any of the post-task questions about the searchbot. That said, a few trends are worth noting. First, participants found the searchbot’s intervention to be slightly more distracting in the bot_auto versus the bot_q condition. We believe that this is because the top-three search results displayed in the bot_auto condition (see Figure 1b) took a much larger portion of the Slack screen than the first question asked by the searchbot in the bot_q condition (see Figure 1a). Second, participants were slightly more confident in the searchbot’s results in the bot_auto versus the bot_q condition. We believe that this is because participants were able to immediately see how the searchbot’s results were relevant to both of their personal preferences. Finally, participants reported slightly higher gains from the searchbot in the bot_auto versus the bot_q condition. Participants’ responses were slightly higher for saved_time and discover_info. Participants reported gaining fewer ideas about things to search for on their own in the bot_auto versus
the bot_q condition. However, this may simply be because participants were less inclined to search on their own in the bot_auto versus the bot_q condition.

4.2.2 Measures of Searchbot Engagement. In addition to analyzing participants’ perceptions about the searchbot, we also computed two measures associated with participants’ level of engagement with the searchbot: (1) the number of clicks (from either participant) on the searchbot’s results and (2) the number of items selected by participants (out of three) that came from the searchbot’s results.

In terms of both measures, engagement with the searchbot was roughly equal. The number of clicks on the searchbot’s results were 2.80 ± 0.52 in the bot_q condition and 2.92 ± 0.52 in the bot_auto condition. Similarly, the number of searchbot results selected by participants in their final solution were 2.12 ± 0.18 in the bot_q condition and 2.00 ± 0.17 in the bot_auto condition. To analyze the effect of the searchbot’s intervention type on these measures, we used a linear mixed-effects model (LMM) with participant-pair id as a random effects variable and used the bot_q condition as the baseline. The searchbot’s intervention type was not a significant predictor for either measure.

4.3 RQ3: Impressions of the Searchbot

In our third research question (RQ3), we investigate participants’ impressions about the searchbot. To address this question, we analyzed participants’ responses to the two open-ended questions that were included in the second part of the post-task questionnaire. Recall that this part was given to participants in the bot_q and bot_auto conditions who actually triggered the searchbot (n = 98 out of 108). The first open-ended question asked: “If the searchbot helped you during the task, briefly explain how.” The second open-ended question asked: “If the searchbot did not help you during the task, briefly explain why not.”

To analyze participants’ responses, two of the authors conducted two rounds of qualitative coding. During the first round, both authors independently coded participants’ responses using open coding and then resolved their codes to derive a closed set of codes. During the second round, both authors independently re-coded participants’ responses using the closed set of codes. Ultimately, the closed set of codes included 11 different codes for the first question and 6 different codes for the second. A code was assigned to a participant’s response only if both authors agreed on the presence of the code during the second round of coding. The Cohen’s Kappa (κc) agreement during the second round of coding was at the level of “almost perfect” (κc > .80) for 15 codes and “substantial” (0.60 < κc ≤ .80) for 2 codes [21]. Participants’ responses to the first question were grouped into two categories: (1) motivations for engaging with the searchbot and (2) gains obtained from the searchbot.

Motivations for engaging with the searchbot: Participants reported six motivations for engaging with the searchbot: (1) the task was difficult and I had little prior knowledge (difficult task); (2) the searchbot intervened at an appropriate time, for example, as I was about to start searching (appropriate intervention); (3) the searchbot provided results that matched both of our preferences (relevant results); (4) the searchbot asked us questions that were relevant to the task (relevant questions); (5) the searchbot gave us immediate results without asking any questions (immediate results); and (6) the searchbot provided a limited number of results (limited results). Figure 5 shows the number of responses associated with each code for both bot_q and bot_auto conditions.

Our results show five interesting trends. First, participants reported that the task difficulty and their level of prior knowledge...
in the task domain was an important factor in the searchbot’s usefulness during the task. Second, participants were more inclined to engage with the searchbot when the intervention came at an appropriate time. Interestingly, this code was more common in the bot_auto versus bot_q condition (7 vs. 2), suggesting that the intervention was more salient in the bot_auto condition. Third, the most common code about why participants engaged with the searchbot was that the search results matched the preferences of both participants (n = 27). Interestingly, this code was also more common in the bot_auto versus bot_q condition (18 vs. 9), suggesting that relevant results were more impressive when the searchbot did not ask questions. Fourth, participants were motivated to engage with the searchbot in both conditions, but for different reasons. In the bot_q condition, six participants reported that the searchbot asked relevant questions, while in the bot_auto condition, seven participants expressed a positive reaction to the searchbot not asking any questions (immediate results). Finally, a few participants reacted positively to the searchbot producing only a few relevant results.

**Gains obtained from the searchbot:** Participants reported five gains obtained from engaging with the searchbot: (1) the searchbot provided ideas on how to get started (help started); (2) the searchbot provided auxiliary information such as ratings on items and an interactive map on the landing page (auxiliary info); (3) the searchbot saved us time (saved time); (4) it was useful for my partner and I to be able to see the same search results (shared context); and (5) the searchbot provided ideas about things to search for on our own (provided ideas). Figure 6 shows the number of responses associated with each code for both bot_q and bot_auto conditions.

Our results show four interesting trends. First, participants reported that the searchbot was useful even if it was not the only resource used to complete the task. For example, participants reported that it helped in getting started with the task and provided ideas about things to search for. Second, participants reacted positively to the auxiliary information included in the searchbot’s results, such as ratings on items and the interactive map. Third, participants (n = 21) reported that the searchbot saved the collaborators time, which re-enforces previous results that it enhanced the collaborative experience. Finally, participants (n = 8) also reported that the shared context helped them collaborate more effectively. Participant responses included statements such as: “We were both able to see the museums on the same screen and better agree on the places we wanted to go.” and “We could both see the information and did not have to guide each other through different websites.”.

**Reasons for not gaining from the searchbot:** Participants reported six reasons for not obtaining any gains from the searchbot: (1) the searchbot intervened at an inappropriate time (inappropriate intervention); (2) I/we had already started searching on our own (already started); (3) I/we had already solved the task (already solved); (4) the searchbot’s results were not relevant (non-relevant results); (5) the searchbot asked questions about things we had already discussed (redundant questions); and (6) we both replied to the searchbot’s questions, but it used the most general (rather than specific) reply. Figure 7 shows the number of responses associated with each code for both bot_q and bot_auto conditions.

Our results show six interesting trends. First, five participants reported that the searchbot’s intervention was not at an appropriate time. One participant reported that it intervened “too early”, possibly when the participants were still making sense of the task, and another participant reported that it intervened “in the middle of typing”. Interestingly, this code was more common in the bot_auto condition, which suggests that the intervention was more salient (and therefore more disruptive) in the bot_auto condition. Second, one participant reported that the searchbot was not useful because he/she had already started searching. This result is consistent with prior research which shows that users tend to avoid help systems when it is difficult to cognitively disengage with the current task [8]. Third, as one might expect, the searchbot was not useful when the participants had already solved the task. Fourth, 11 participants reported not gaining from the searchbot because the results were not relevant. Fifth, one participant found it annoying to answer questions about things already mentioned in the conversation. Finally, there was one case where both participants replied to the same question from the searchbot, and the searchbot apparently used the wrong input (i.e., the most general response). This result suggests that searchbots that ask questions may need to accommodate (or gracefully ignore) multiple replies from users.

5 DISCUSSION

**Research Question RQ1:** In terms of our first research question, our results suggest that the searchbot improved participants’ collaborative experience. In the bot_q and bot_auto conditions, participants reported: (1) greater awareness of each other’s activities; (2) greater ease in sharing, communicating, and reaching consensus; and (3) greater levels of enjoyment. Additionally, in the no_bot condition, participants exchanged a greater number of URLs through Slack. This result suggests that the absence of the searchbot forced participants to search independently and share their findings through Slack.

In terms of the searchbot’s intervention type, participants’ perceptions of their collaborative experience were slightly better in the bot_auto versus bot_q condition. For one of our post-task measures (aware_browser), there was a significant difference between the bot_auto and bot_q conditions. For two of our post-task measures (ease_comm, and ease_con), there were significant differences
between the bot_auto and no_bot conditions, but no significant differences between the bot_q and no_bot conditions.

Our RQ1 results suggest that integrating search tools (e.g., searchbots) into communication channels such as Slack provides some of the same benefits provided by dedicated collaborative search systems [4, 12, 26, 28, 32, 37]. Prior studies involving dedicated systems have found that tools to support chat-based communication are heavily used during collaborative search [37]. In this respect, searchbots have the advantage that they are directly integrated into the communication channel. Additionally, dedicated systems typically include features that raise collaborators’ awareness of each other’s activities. Prior research has found that these features improve users’ experience [26, 28] and reduce communication and coordination efforts [33]. Our results suggest that the shared context provided by a searchbot (i.e., allowing collaborators to see the same information directly in the communication channel) can improve the collaborative experience and reduce the need to search independently and coordinate by copy/pasting URLs.

Research Question RQ2: In terms of our second research question, we did not observe strong effects from the searchbot’s type of intervention on participants’ perceptions about the searchbot and level of engagement with the searchbot. It is important to note that in this particular study, we simulated the best-case scenario for both intervention types. In other words, in the bot_q condition, the searchbot elicited information that was relevant to the task, while in the bot_auto condition, that searchbot “inferred” the needed information in order to provide contextually relevant results. Participants’ perceptions about the searchbot were equally positive in both of these best-case scenarios.

Our RQ2 results have two important implications for future work. First, our results suggest that both intervention types (elicitation versus inference) are equally good if they are done well. In the bot_q condition, participants were not annoyed by having to respond to the searchbot’s questions, while in the bot_auto condition, participants did not strongly mistrust the searchbot’s ability to provide relevant results without eliciting information (perhaps because the results were visibly relevant). This result is consistent with prior research on interruptions, which shows that interruptions that are relevant to the current task tend to be less disruptive [7, 17].

The second implication is that future work is needed to understand the differences between these two intervention types under non-ideal conditions. What is the cost of eliciting information that is not relevant to the task? What is the cost of embedding non-relevant results into the communication channel? What is the cost of eliciting information, if it means the difference between relevant versus non-relevant search results? These are questions for future work. Prior research on different interruption protocols [23] may generate ideas about how a searchbot should intervene given its level of confidence that it has inferred the users’ needs.

Research Question RQ3: Our qualitative analysis of participants’ responses reveal interesting opportunities and challenges for dynamic searchbots.

In terms of opportunities, participants reported different gains obtained from the searchbot. Participants reported that the searchbot provided relevant results, saved the participants time, helped them get started with the task, generated ideas about things to search for, provided a limited set of relevant results, and provided auxiliary tools that were useful for the task (e.g., an interactive map). Most importantly, the searchbot provided a shared context that made the collaboration easier. In a large scale user study, Xi and Cool [36] found that individual searchers encounter difficulty with seven general processes: (1) getting started, (2) identifying relevant sources, (3) navigating a source, (4) constructing queries, (5) constraining the results, (6) recognizing relevant content, and (7) monitoring the task process. The gains reported by our participants touch upon several of these processes.

In terms of challenges, participants’ responses suggest several factors that may influence their decision to engage with a searchbot. Several participants reported engaging with the searchbot because the task was difficult and they had little prior knowledge in the task domain. This result is consistent with prior work which found that users are more likely to engage with search assistance tools during complex tasks [3]. Furthermore, participants’ responses clearly indicate that the point of intervention is key. Participants reported avoiding the searchbot because the intervention was too soon (i.e., before they fully understood the task), too late (i.e., after they had completed the task), during the middle of some activity (e.g., while typing), and after participants had already engaged with their own approach to the task (e.g., already searching on their own). Similarly, participants reported engaging with the searchbot because it intervened right as they were about to start searching. These participants’ responses echo previous findings from two lines of prior work. Prior work on dynamic help systems has found that people avoid help systems when they do not understand how the system can help (e.g., when the intervention happens too soon), or when it is costly to cognitively disengage from the current activity (e.g., after they have already started searching). Similarly, prior work on interruptions has consistently shown that interruptions are less disruptive during periods of low mental workload (e.g., during sub-task transitions) [1, 16, 17].

Practical Concerns: The searchbots that we explored in this paper rely on the ability to monitor a conversation between collaborators and to use the conversational history in order to decide when and how to intervene. There are many privacy, security, and ethical issues that would need to be addressed in such a system. Some of these issues are common to existing systems that monitor an environment in order to provide services (e.g., online email services, home monitoring systems, and sensors in Internet of Things devices). However, the collaborative nature of the searchbot presents additional issues to be considered. For example, it might be advisable to “activate” a searchbot only if all members of the chat channel have enabled it as a feature (i.e., a global setting). In addition, the searchbot could clearly announce itself at the beginning of a conversation and should be “visible” like any other member of the chat channel.

6 CONCLUSION
We reported on a Wizard of Oz user study that investigated the use of a dynamic searchbot during collaborative information-seeking tasks coordinated using the Slack messaging system. The searchbot intervened in two different ways: (1) by eliciting information and (2) by “inferring” the needed information from the conversation and directly providing search results.
In terms of our first research question (RQ1), our results show that the searchbot improved our participants’ collaborative experience and reduced the need to search independently. Moreover, participants’ perceptions about their collaborative experience were slightly better in the condition where the searchbot intervened by directly providing contextually relevant results (without eliciting).

In terms of our second research question (RQ2), the searchbot’s type of intervention did not greatly affect participants’ perceptions about the searchbot. However, in this study, we simulated the best-case scenarios for both intervention types: the searchbot asked follow-up questions that were relevant to the task and always produced relevant results. Future research is needed to understand the trade-offs between asking potentially non-relevant questions and returning potentially non-relevant results. The manner of intervention may have a greater effect in non-ideal conditions.

In terms of our third research question (RQ3), participants reported different motivations for engaging with the searchbot, different gains obtained from the searchbot, and different reasons for avoiding the searchbot. Our results suggest that the point of intervention is key. Participants reported avoiding the searchbot when the intervention was too soon (before understanding the task), too late (after solving the task), or during periods when they were deeply engaged with other tasks. This finding is consistent with prior work on dynamic help systems and interruptions.

The work presented in this paper is an instantiation of a greater area for future research: embedding dynamic search tools into communication channels that are widely used for collaboration. From an IR perspective, many open questions remain: (1) inferring collaborators’ needs from their communication, (2) predicting when to intervene, and (3) deciding when and how to elicit information in order to provide relevant results. Outside of IR, the CSCW community has also identified collaborative agents (e.g., intelligent personal assistants, chatbots, and embodied robots) as an important area for research. Recent workshops and panel discussions at CSCW ’16 and CSCW ’17 have identified important challenges and opportunities for emerging collaborative agents. These include issues about: (1) how agents can integrate into the collaborative process [31], (2) how people coordinate and co-manage the use of an agent [31], (3) the social dynamics involved when people engage with agents in collaborative settings [10, 31], (4) questions about whether people engage with agents as they do with their human collaborators [10], and (5) questions about whether an agent’s behavior might influence the way human collaborators interact with each other [19].

Acknowledgements: This work was supported in part by NSF grants IIS-1451668 and IIS-1552587. Any opinions, findings, conclusions, and recommendations expressed in this paper are the authors’ and do not necessarily reflect those of the sponsor.

REFERENCES


