# How to Guide Task-oriented Chatbot Users, and When: A Mixed-methods Study of Combinations of Chatbot Guidance Types and Timings

Su-Fang, Yeh*
sfy.iem07g@nctu.edu.tw
National Yang Ming Chiao Tung University
Hsinchu, Taiwan, Taiwan

Meng-Hsin, Wu*
menghsin.wu@mail.utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Tze-Yu, Chen
Yen-Chun, Lin
alexchen.ms07@nctu.edu.tw
barney0817.cs10@nycu.edu.tw
National Yang Ming Chiao Tung University
Hsinchu, Taiwan, Taiwan

Xi-Jing, Chang
siliconcrystal.c@nycu.edu.tw
National Yang Ming Chiao Tung University
Hsinchu, Taiwan, Taiwan

You-Hsuan, Chiang
youxuanjiang19960901@gmail.con
National Yang Ming Chiao Tung University
Hsinchu, Taiwan, Taiwan

Yung-Ju, Chang†
armuro@cs.nctu.edu.tw
National Yang Ming Chiao Tung University
Hsinchu, Taiwan, Taiwan

## ABSTRACT

The popularity of task-oriented chatbots is constantly growing, but smooth conversational progress with them remains profoundly challenging. In recent years, researchers have argued that chatbot systems should include guidance for users on how to converse with them. Nevertheless, empirical evidence about what to place in such guidance, and when to deliver it, has been lacking. Using a mixed-methods approach that integrates results from a between-subjects experiment and a reflection session, this paper compares the effectiveness of eight combinations of two guidance types (example-based and rule-based) at four guidance timings (service-onboarding, task-intro, after-failure, and upon-request), as measured by users' task performance, improvement on subsequent tasks, and subjective experience. It establishes that each guidance type and timing has particular strengths and weaknesses, thus that each type/timing combination has a unique impact on performance metrics, learning outcomes, and user experience. On that basis, it presents guidance-design recommendations for future task-oriented chatbots.

## CCS CONCEPTS

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

## KEYWORDS

chatbot; lab study; non-progress; guidance

*Both authors contributed equally to this research.
†Corresponding author.


*CHI '22, April 29-May 5, 2022, New Orleans, LA, USA*

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501941

**ACM Reference Format:**
Su-Fang, Yeh, Meng-Hsin, Wu, Tze-Yu, Chen, Yen-Chun, Lin, Xi-Jing, Chang, You-Hsuan, Chiang, and Yung-Ju, Chang. 2022. How to Guide Task-oriented Chatbot Users, and When: A Mixed-methods Study of Combinations of Chatbot Guidance Types and Timings. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3491102.3501941

## 1 INTRODUCTION

In recent years, there has been tremendous growth in the use of chatbot applications on text-based messaging platforms. In 2018, more than 300,000 chatbots were live on Facebook Messenger, and were commonly used to transact business [6] in multiple domains [28]. Task-oriented chatbots, in particular, are aimed at helping their users perform domain-specific tasks, and can be important tools for saving time on repetitive tasks [71]. Perhaps for these reasons, "efficiency" has been identified as the most important motivation for chatbot use [7–9, 23]. Nevertheless, making efficient progress in conversations with chatbots remains a fundamental challenge for their users: obstacles to conversation commonly arise from the chatbot expressing uncertainty about users' intentions (conversation breakdown; true or false negative) or misunderstanding users' messages (false positive), and providing unexpected messages [4, 24, 44, 48].

While some obstacles arise from the fact that chatbots' technical capability to understand natural language is still evolving, research has suggested that it sometimes results from people's overestimation or other misunderstanding of their chatbot's capability [37, 38, 44, 48, 55]. To reduce such gaps in understanding, various scholars have recommended providing specific guidance on how to interact with a system [33, 38, 61]. Yet, regarding when to provide it, there has been inconsistent suggestions. For instance, whereas some researchers have highlighted that preventing users from experiencing such breakdowns in the first place is as important [46, 59], such as presenting them at the initial stage of chatbot interaction [3, 38], others have suggested that it be provided when requested by

users [38, 40, 72]. To date, there appear to have been no empirical investigations of the impacts of different timings of conversational guidance for chatbot users. Thus, it is unclear how the diverse rival timings recommended so far would actually be most helpful to users using task-oriented chatbots.

There has also been scant attention to what kind of guidance should be offered at these various guidance-delivery time-points. Example-based guidance, for example, has frequently been recommended or adopted in various kinds of intelligent user interface [10, 38, 40, 72]. While examples can explain complex concepts easily, however, they do not readily indicate the limits of bots' capability. Yet, rules can explicitly state limits of a system's capability, but meanwhile have been found to require a steeper learning curve [20, 65]. Given these different characteristics, the question of which type of guidance is more suitable to provide at particular points in time is a potentially complex issue.

The present study aims to fill these research gaps. Specifically, our objective is to shed light on which of eight combinations of two guidance types (Example-based and Rule-based) and four timings (Service-onboarding, Task-intro, Upon-request, and After-failure) yield better user performance and subjective experience. Accordingly, we developed chatbots equipped with one of the eight possible combinations of guidance types and timings, in the expectation that they would allow us to answer the following research questions:

- RQ1. Which combination of guidance type and timing enables users to a) complete their tasks more efficiently, b) make better conversational progress, and c) improve their performance during subsequent chatbot use?
- RQ2. What are users' subjective experiences of each of these combinations?

In addition, to facilitate the design of future task-oriented chatbots, we asked a third research question:

- RQ3. What are users' desired characteristics for the combination of a chatbot-conversation guidance type and its timing?

To answer these questions, we used a two-phase mixed-methods approach, which was a mixture of an experiment and reflection sessions that allowed us to obtain quantitative results to answer RQ1, and qualitative results to answer RQ2 and RQ3. This paper makes four main contributions to the literature as follows:

- It identifies overall patterns of task performance and improvement by chatbot-conversation guidance type, including that example-based guidance yields better initial task performance but poor task-on-task improvement, and that rule-based guidance yields weak initial task performance but more improvement.
- It reveals overall patterns of task performance among guidance timings, including that providing guidance as soon as a task is introduced generally yields good performance in overall tasks, and that providing it upon request generally yields poor performance in the case of initial tasks.
- It ascertains that specific guidance type/timing combinations that lead to particularly high and low task performance and improvement, e.g., showing rules after failure is associated with excellent performance; showing examples upon request, with strong improvement; and showing examples at service onboarding, with weak improvement.
- It provides qualitative insights into the performance and improvement of the eight guidance type/timing combinations.

Based on these findings, the paper provides design recommendations for the guidance-provision features of future task-oriented chatbots.

## 2 RELATED WORK

### 2.1 Guidance Timing

Helping users recognize, diagnose, and recover from errors has long been viewed as an important heuristic for building good usability [3, 46, 59]. Investigations of how to repair conversational breakdowns has been studied in recent years. [4, 24, 77]. Ashktorab et al. [4] looked at eight repair strategies, and reported that people preferred a chatbot to provide its guesses as to what they had meant. However, conversation breakdowns, or non-progresses, referred to as users not making conversation progress[48], which include misunderstandings [24, 44, 48], may already harm users' conversational experience [39, 70] and may lead to conversation abandonment [38, 48].

Numerous researchers have recommended managing users' expectations by progressively issuing them clear guidance about AI capability and functionality [3, 15, 21, 33, 35, 38, 48, 52, 56, 58]. Some have focused on the timing of such guidance, and others, on the format for expressing it.

In terms of timing, some researchers have recommended that guidance be provided at the start of the interaction. For example, Amershi et al. [3], in their guidelines for human-AI interaction, suggested that an AI system should state its capabilities and functionalities at the initial stage of the interaction. Similarly, Jain et al. [38] interviewed 16 first-time users of several chatbots about areas that stood in need of improvement. As well as after a failure occurred, their participants said that they wanted to access information on the chatbots' capability and functionality either explicitly at the start of the interaction, or on demand at any time, in the form of examples. Unfortunately, neither Amershi et al. nor Jain et al. clearly explained when they deemed interactions to have started, or when the "initial" part of an interaction ended.

Nielsen [59], on the other hand, recommended that an interface guidance should always be accessible to users; and Langevin et al. [46] likewise suggested that chatbot systems should guide users by clarifying their capabilities throughout an interaction. Yet, proactive guidance may be unexpected, annoying and/or distracting [12, 14, 62], so providing users with guidance upon request might be more in line with their expectations [22].

However, while the research reviewed above has usefully revealed perceptual aspects of users' guidance-timing preferences, the question of whether catering to such preferences would yield empirically better communicative outcomes remains unanswered, despite a high proportion of chatbots use being outcome-focused [7, 9, 23].

One study highly relevant to the present one focused on VUIs [40], where a Wizard of Oz experiment was conducted to compare task performance and user satisfaction across three conditions: proactive guidance, reactive guidance, and a no-guidance baseline.

Their participants were asked to complete a food-ordering task three times, once for each experimental condition. The researchers found no significant difference between these two guidance timings in terms of task performance. In their interviews, however, the participants commented that when using a VUI for the first time, proactive guidance would help them, but that over the long term, they would prefer on-demand guidance. Complementing this prior work, our study examines three timings for proactive guidance, and it suggests that each combination of a guidance type and a guidance timing has specific effects on users' progress in chatbot communication, which have not previously been identified or discussed in the chatbot literature.

## 2.2 Guidance Types

Prior research has indicated that people tend to converse with chatbots using their existing mental models of communication [37, 38, 44, 48, 55]. This means that they may regard a conversational user interface as anything from a human-like agent that can understand natural languages, to a system only capable of dealing with a series of predetermined commands and syntax [2, 19]. Accordingly, we conducted a broad review of the literature not only on example-based and rule-based instructions, but also in the spheres of syntax and language learning, including work on computer programming and query formulations for information-retrieval (IR) systems, even where it was not explicitly related to the use of chatbots.

Taking example-based instructions first, a considerable number of studies have shown that providing explicit problem-solution examples generates worked-example effects [66, 67], which help people acquire skills by decreasing their cognitive load. However, the positive impact of example-based learning depends on how well learners generate their own explanations of why the example works, and generalize such understanding to other problems they later face [68]. The quality of this generalization process has been found to be rooted in individuals' pre-existing problem-solving ability [16, 63]. Therefore, various researchers have argued that providing specific principles-based guidance (similar to rule-based guidance) is as important as providing examples [5, 64, 65, 69].

In the specific case of English writing acquisition, Kyun et al. [45] demonstrated that students who were exposed to worked examples learned significantly more than those who were exposed to no guidance. Similarly, Ellis [20] found that examples helped second-language (L2) learners achieve competence faster than rules did. Yet, the same study reported that explicitly stating rules facilitated these learners' broad understanding of the language and thus, their grammar performance; and recommended that both examples and rules be used in L2 learning contexts. Similar to these results, our study also showed that rule-based guidance generally led to more improvement in task performance.

Examples and rules have both been highlighted as effective scaffolding for learning IR (e.g., [76]). Halttunen [30, 31] used both examples and query-formulation cues for this purpose, and found that providing feedback alongside such guidance could help students learn and query more effectively than a traditional IR learning environment, in which the tutor was the only source of instruction.

In programming-related research, *programming by example* is a powerful paradigm that alleviates the complexity of learning by demonstrating actions concretely [13, 34, 57]. Researchers have also found that, when learning an application programming interface (API), programmers adopting either a concept-oriented or a code-oriented learning strategy would like to read information about parameters [49]; and documentation serves as a good reference both for why programming issues happen and how to address them [53]. Example-based and rule-based guidance are both prevalent in such documentation currently.

These two broad types of guidance have also been explored in research on human interaction with intelligent systems. For example, Waa et al. [74] compared rule-based and example-based explanations of an Explainable AI system that generated explanations for decision support systems (DSS), and found that while the example-based ones were better at encouraging people's compliance with the system's advice, rule-based explanations helped the participants more accurately identify decisive issues in DSS's feedback messages. Similarly, Stumpf et al. [73] showed that regarding feedback of an E-mail spam filter system, rule-based feedback was more understandable than, and preferred by users over, similarity-based feedback, which is being comparable to example-based guidance) .

However, despite the large body of evidence that rule-based guidance is as valuable as example-based guidance, various kinds of intelligent user interfaces continue to recommend or incorporate example-based guidance only (e.g., [10, 38, 40, 72]).

Research investigating the effects of different combinations of guidance type and timing to help users of task-oriented chatbot has hitherto been rare to nonexistent. This mixed-methods study fills that gap by investigating the effectiveness of eight combinations of guidance timing and type in assisting the users of task-oriented chatbots to communicate with them smoothly, as measured by task-execution performance, the promotion of learning, and users' subjective experience.

# 3 METHODS

## 3.1 Chatbot and Tasks

Our target tasks were based on two dimensions: task context and complexity. We designed two contexts, one related to arranging travel and the other to movie booking, both of which have been extensively used in prior chatbot research [4, 38, 47, 50]. For each context, we developed chatbots that handled tasks at three levels of complexity, based on Campbell's proposition [11] that the more requirements the task involves, the higher its complexity. Specifically, the low-, medium-, and high-complexity tasks required four, six, and eight pieces of information to accomplish, respectively. The detailed task requirements can be seen in supplementary materials.

The chatbots in this study were built on IBM Watson platform [1]. We developed versions of the movie and travel chatbots specific to each of nine guidance conditions, i.e., eight combinations of guidance type and guidance timing, plus a control group that received no guidance. Our conversation-design process was based on a synthesis of the guidance offered by multiple chatbot platforms [26, 29, 36, 54]. We gathered candidate travel and movie tasks from

[1] https://www.ibm.com/watson

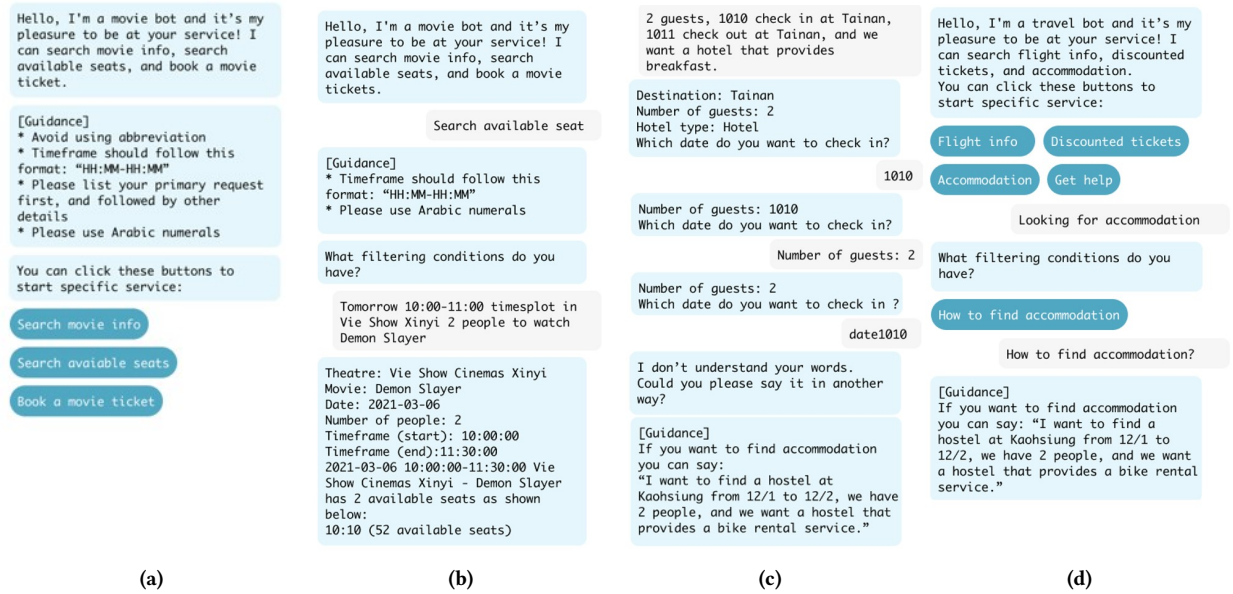



**Figure 1: English translation of example dialogue snippets in medium-complexity tasks with chatbot guidance: (a) the movie chatbot with rule-based guidance and service-onboarding timing; (b) the movie chatbot with rule-based guidance and task-intro timing; (c) the travel chatbot with example-based guidance and after-failure timing; (d) the travel chatbot with example-based guidance and upon-request timing.**

websites that offer these services and identified the information that completing such tasks would require. Then, we generated sample dialogues for each task, diagrammed the conversation flows, and iterated the conversation-flow design and content with several pilot participants. In building our language-understanding model, we brainstormed 20 training phrases that had varying formats and contained varying numbers of entities. In addition, the chatbot was designed to allow the user to enter multiple pieces of information in one utterance, in their own desired order. It was also designed to recognize users' intent to break off communication, and to provide options for them to exit the conversation flow when that happened.

## 3.2 Study Design and Chatbot Guidance

This study's four selected guidance timings and two guidance types (see Figure 1) were based on the foregoing review of the relevant literature. Each is defined below.

*Guidance Timings*

- Service-onboarding (ONB): The user receives the guidance along with a brief introduction to the chatbot at the start of the service.
- Task-intro (TASK): The user receives the guidance at the start of any task, irrespective of whether s/he has used the same chatbot system before.
- After-failure (FAIL): The user receives the guidance after the chatbot indicates that it does not recognize his/her intent.
- Upon-request (REQ): The user receives the guidance after requesting it, either by clicking a "Help" button or via their own typed utterance.

*Guidance Types*

- Example-based (EXMP): The user receives an example of an utterance that can be understood by the chatbot, with an average length of between 39 and 68 Chinese characters [2].
- Rule-based (RL): The user receives between one and four rules of syntax and format, depending on the task, in a bullet-pointed format and a random order. All rules were decided based on the natural language processing limitations of IBM Watson, as carefully checked by the research team. These rules consisted of reminders to avoid using abbreviations, specific formats for dates and times, and the preferred order of requested information (e.g., departure city first, in the case of airline-ticket requests).

The above guidance types and timings resulted in eight possible guidance type/timing combinations (i.e., 2 types x 4 timings), each of which is abbreviated in the remainder of this paper using a "timing-type" format (i.e., ONB-EXMP, ONB-RL, TASK-EXMP, TASK-RL, FAIL-EXMP, FAIL-RL, REQ-EXMP, and REQ-RL). Moreover, we examined one condition in which participants received no guidance, resulting in nine experimental conditions in total.

## 3.3 Participants

A total of 126 people, 63 male and 63 female, aged between 20 and 45 ($M$=26), participated in the study. They were recruited via the main social-media platforms in Taiwan, including PTT, Dcard, and Facebook groups. All participants provided their demographic information and familiarity with chatbots[4] when answering our

[2]Each task had its own unique example. However, in the service-onboarding timing condition, all three examples of the tasks they might perform using that chatbot were simultaneously presented, rather than (as in the other three timing conditions) just the example of the one task they were trying to execute at that moment.

online sign-up questionnaire. Taking together self-reported prior chatbot usage experience, i.e. 0 or 1, and familiarity with chatbots, i.e. 1 to 5 on a five-point Likert scale, half (n=63) reported high familiarity with chatbots, and the remaining reported low familiarity. Participants' backgrounds were balanced when they were assigned to the nine conditions, via a semi-randomization approach [27]. Each participant received NT$300 (approximately US$11) as compensation for their participation.

## 3.4 Study Procedure

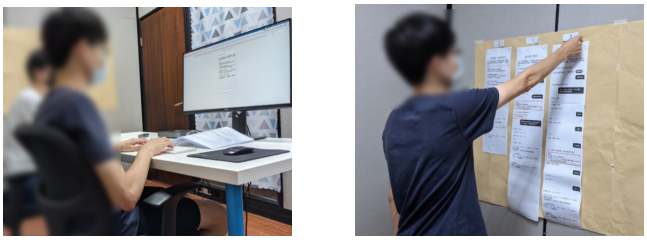
The study took place face-to-face in a laboratory environment, with each participant attending alone (Figure 2).

*3.4.1 Phase 1: Between-subjects Experiment.* First, they were informed about the study process and signed a consent form, which included permission for audio and screen recording during the entire study. They were next asked to perform a warm-up task to familiarize themselves with the keyboard they would be using and the experimental environment. In the warm-up task, they were given a task script (similar to those for the actual experimental tasks), which asked them to reserve a table in a restaurant, and then to converse with a conversation agent in a blank document. As well as keyboard familiarization, this allowed us to observe how each participant talked to a conversational user interface.

Next, the Phase 1 between-subjects experiment was executed on a desktop computer. The host first informed the participants that they should treat the chatbot as they would do in any casual setting, and that therefore, they could abandon a task when/if they decided they did not want to use a chatbot to complete it. Then, the participants took part in two trials. In the first trial, the participants performed a total of six closed-ended tasks, three (of three different complexity levels) on one movie chatbot, and the other three (also of three different complexity levels) on one travel chatbot. The order of the context and of the complexity were partially counterbalanced, following Gravetter et al.[27]. During a given person's conversations, the chatbot provided just one of the eight guidance combinations to assist them in accomplishing their tasks (or no guidance, in the case of the control group).

After the first trial was finished, the participants took a five-minute break. Then, in the second trial, they performed the same number and kinds of tasks, with slight variations in the task descriptions and requested information, with the same guidance type/timing combination of chatbot guidance as before. The purpose of this was to examine whether they could successfully accomplish similar tasks based on what they had learned in the first trial.

After finishing both trials, the participants filled out an online questionnaire, in which they rated their satisfaction with the guidance [33] on five-point Likert scales. The phase 1 took 30–40 minutes. Then, they moved on to Phase 2, the reflection study.



(a) (b)

**Figure 2: The laboratory set up: (a) a participant interacting with the chatbot via a desktop computer; (b) a participant looking at printed screenshots of the conversational context of each guidance combination, and arranging them based on his preferences.**

*3.4.2 Phase 2: Reflection Study.* The purpose of the reflection study was to capture the participants' perceptions, attitudes, preferences, and concerns about each guidance combination. It started with 10- to 15-minute semi-structured individual interviews, in which the participants were asked to review their conversation histories with the chatbot and walk us through their thoughts. Next, we showed them the eight guidance modalities that they had not been exposed to in Phase 1, and had them rank all nine in order of personal preference. Then, we carefully probed their rationales for these rankings. Lastly, the participants were asked to reflect on the combinations and modify their rankings if they felt it necessary. Throughout, we also asked them about their desired/ideal characteristics for chatbot guidance. This ranking process took 20–30 minutes.

## 3.5 Measures

*3.5.1 Performance Metrics.* In Phase 1, we recorded three performance metrics: task success, task-completion time, and number of non-progress events (following Li et al. [48]). Additionally, given that learnability is an important aspect of usability of an interactive system [60] and one that has previously been used for assessing whether a conversational user interface allows users to achieve successful interaction [52], we also measured learnability of the chatbot. Specifically, for each metric, we measured individual participants' improvement between the two sets of trials, as a proxy for system learnability [1].

A participant was defined as having succeeded in a task only if all three of the following criteria were met: 1) the task was not abandoned, 2) all task requests were fulfilled, and 3) the number of non-progress events was less than four. We set three as the upper limit of non-progress events based on prior conversation analysis by Li et al. [48], which established that this was the typical maximum number of such events a user could tolerate without them leading to service abandonment. Efficiency was measured based on task-execution time. Non-progress events were counted following Li et al.'s [48] definition, i.e., when the participants became aware that the chatbot could not recognize or had mis-recognized their intent. However, it should be noted that efficiency and non-progress events were only measured for tasks that were successfully completed.

*3.5.2 Post-Study Scales.* The participants rated their satisfaction with the guidance they had received on five-point Likert scales translated into Mandarin from the Explanation Satisfaction Scale [33]. Due to our research purpose, this research focused on understandability, feelings of satisfaction, and the perceived usefulness of the guidance/timing combinations.

**Table 1: Regression table for fixed effects (df=727). The group with the lowest number of successes (in this case, the Control group) was coded as the reference group. *p<.05, **p<.01, ***p<.001**

| | Trial1 | | | | Trial2 | | | |
|---|---|---|---|---|---|---|---|---|
| Guidance | Estimates | SD | Z | p | Estimates | SD | Z | p |
| (Intercept) | 1.386 | 0.486 | 3.971 | 0.004 ** | 2.354 | 1.095 | 2.149 | 0.032 * |
| REQ-EXMP | 0.455 | 0.609 | 0.747 | 0.455 | 0.845 | 0.863 | 0.979 | 0.328 |
| Upon-request Rule | 0.697 | 0.627 | 1.111 | 0.267 | 2.152 | 1.205 | 1.787 | 0.074 |
| FAIL-EXMP | 0.855 | 0.651 | 1.314 | 0.189 | 0.348 | 0.780 | 0.446 | 0.656 |
| FAIL-RL | 1.098 | 0.686 | 1.600 | 0.110 | 0.149 | 0.772 | 0.193 | 0.847 |
| ONB-EXMP | 0.861 | 0.648 | 1.328 | 0.184 | 2.119 | 1.206 | 1.756 | 0.079 |
| ONB-EXMP | 0.208 | 0.579 | 0.359 | 0.720 | 1.116 | 0.897 | 1.244 | 0.214 |
| TASK-EXMP | 1.795 | 0.849 | 2.114 | 0.034 * | 1.063 | 0.881 | 1.207 | 0.228 |
| TASK-RL | 0.656 | 0.624 | 1.053 | 0.293 | 0.716 | 0.828 | 0.864 | 0.388 |
| Task order | 0.164 | 0.086 | 1.912 | 0.056 | 0.000 | 0.101 | 0.005 | 0.993 |

### 3.6 Data Cleaning and Analysis

We obtained the aforementioned performance metrics via data coding of screen-recordings and chat logs. Specifically, three coders coded a sample comprising 2% of the full dataset and discussed and revised the coding protocol until consensus was reached. Then, another 2% of the full dataset was used as a pilot test of the coding schema's reliability, and the same iterative process of discussing and resolving disagreements completed again. After that, the coders tested their revised codes with a sample comprising 10% of the remaining 96% of the data. Lastly, the coders divided the final 108 participants' data into three equal groups of 36 and coded them independently. The code book covered both categorical and continuous attributes. Krippendorff's Alpha [32, 43, 51] was used to guarantee the reliability of the categorical attributes, which included 1) whether the participant gave up, 2) whether s/he completed all task requirements, and 3) whether his/her data should be removed, such as because the study instructions were not followed, or because there was no chatbot output due to Internet instability. The Krippendorff's Alpha value for "gave up" was .71, for the "requirements complete" was .91, and for "data removed" was .71, all of which were comfortably above the reliability threshold of 0.67 [42].

We ensured the reliability of the continuous attributes, task time and number of non-progress events, by calculating intraclass correlation coefficients (ICCs) [41], using a single-rater, consistency, two-way mixed model with three raters across the test dataset. The ICC value we obtained for task time was .93, and for the number of non-progress events was .94, suggesting that reliability of these continuous attribute codes was excellent.

Because they met the criteria for our code "data removed", 35 tasks were eliminated from the complete set of 1,512. A further 95 were then removed because they failed to meet the criteria for our code "task success", leaving a pool of 1,382 tasks for efficiency analysis.

Since each participant performed 12 tasks in the experiment, we used mixed-regression models to analyze the performance metrics to take random effects into account for individual differences. We also took account of the impact of task order. To examine the experimental manipulations' influence on task time and the number of non-progress events, we used linear regression. As task-completion time is not normally distributed, we took the logarithm of raw task-completion time and transformed it into a log-normal distribution to enable linear regression [18]. And for task success, which was a binary metric, logistic regression was used. As for the post-study questionnaire, since each participant only contributed one sample point, we used one-way analysis of variance (ANOVA) to analyze the satisfaction scores. Finally, for the guidance type/timing combination rankings, we conducted Friedman test [25] and Wilcoxon signed-rank tests [75] as post hoc tests.

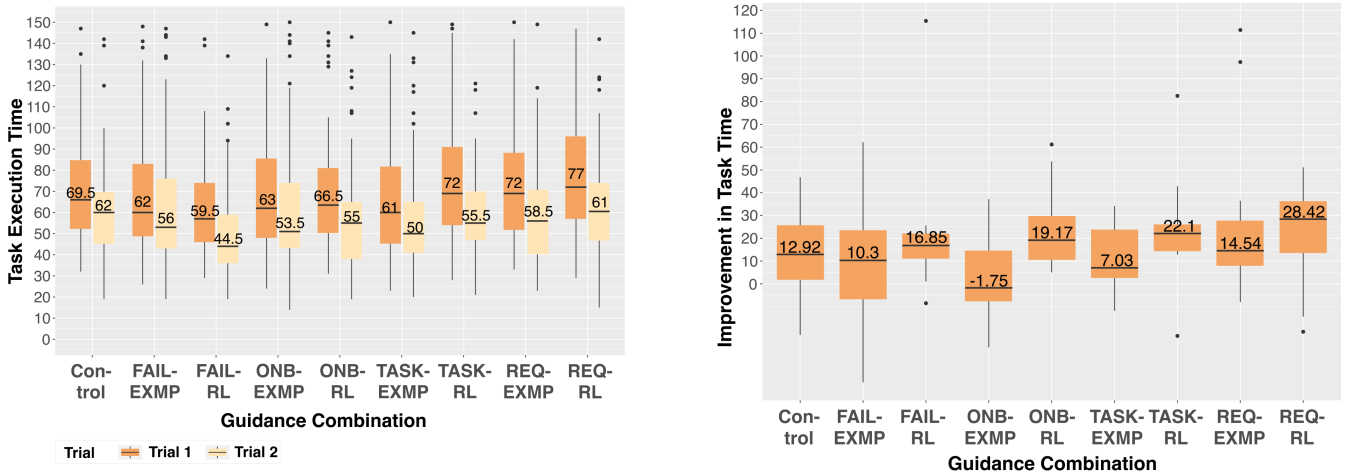
For qualitative analysis, we adopted the affinity diagramming [17]. This process yielded high-level themes relating to how guidance types and guidance timing affected the participants' behaviors and perceptions, and to the impact of use context.

## 4 QUANTITATIVE RESULTS

Our study design made it likely that, in some tasks, the participant would not see any guidance: i.e., because s/he did not request any guidance in the upon-request timing condition, or did not encounter failure that triggered a guidance in the after-failure timing condition. Sometimes, however, participants failed to notice guidance that appeared, or saw it but chose not to read it. By group, the percentages of the participants who self-reported seeing guidance that appeared were 100% for Task-intro/Example and Service-onboarding/Example; 92.9% for Task-intro/Rule; 85.7% for Service-onboarding/Rule; 71.4% for Upon-request/Example and Upon-request/Rule; 57.1% for After-failure/Example; and just 35.7% for After-failure/Rule. These distributions should be borne in mind when interpreting some of the task-performance results.

### 4.1 Task Success

The TASK-EXMP group successfully completed most tasks on average ($M$=5.79, $SD$=0.43), and the control group, the least ($M$=5.00, $SD$=1.47). We ran logistic regression on task success in Trail 1, and most guidance combinations did not differ significantly from each other, the sole exception being the TASK-EXMP group and the control group ($Z$=2.11, $p$=0.034). In Trial 2, none of the pairwise comparisons revealed significant differences (Table 1).




(a) (b)

**Figure 3: (a) Boxplot of the task-completion times for each task in Trial 1 (orange) and Trial 2 (peach). Numbers represent the medians for each guidance/timing combination; (b) Boxplot of the participants' improvement in task-completion times. Numbers represent the median for each guidance/timing combination.**

## 4.2 Task-completion Time

*4.2.1 Task-completion Time in Trial 1.* Task execution time was the average time a participant spent on each successfully completed task. On average in Trial 1, this was 79.8 seconds (*SD*=48.8) overall, 57.2 seconds (*SD*=38) for the low-complexity tasks, and 120.2 seconds (*SD*=63.1) for the complex ones. Interestingly, as shown in Figure 3a, most tasks with example-based guidance took the participants less time to complete in Trial 1 than their rule-based counterparts did. The only exception was the FAIL timing, in which those who received rules rather than examples completed the task an average of 7.6 seconds faster. We speculate that the generally higher efficiency of the example-based groups was due to many participants in those groups copied-and-pasted the example provided in their own tasks. The FAIL-RL group, as well as being markedly faster than its FAIL-EXMP counterpart, was the fastest group overall, with its participants completed task significantly faster than the REQ-RL ($t(112.06)=3.10$, $p=0.002$), REQ-EXMP ($t(118.03)=2.43$, $p=0.022$) and TASK-RL groups ($t(112.40)=2.31$, $p=0.022$) completed theirs. In part, this was because in 146 of their 154 successful tasks, the FAIL-RL participants did not encounter any conversation error that triggered guidance. The average task-completion time of these successful tasks was 58.7 seconds (*SD*=31.6), whereas for those in which at least one conversational error happened, it was 140.1 seconds (*SD*=99.9), i.e., more than twice as long. Moreover, 64.3% of the participants in the FAIL-RL group never saw guidance during any of their 12 tasks. The fact that these participants did not spend any time reading guidance could explain why their group had the highest overall efficiency.

Among the example-based groups, participants in TASK-EXMP and ONB-EXMP spent significantly less time completing their tasks than those in the REQ-RL group (TASK-EXMP vs. REQ-RL: $t(113.42)=2.45$, $p=0.016$; ONB-EXMP vs. REQ-RL: $t(113.08)=2.18$, $p=0.031$). This suggests that showing participants an example before they interacted with the chatbot enabled them to complete tasks more efficiently, probably because they could apply the provided example earlier in the task process.

*4.2.2 Task-completion Time in Trial 2.* As shown in Figure 3b, participants in all guidance type/timing groups improved their task-completion times between Trial 1 and Trial 2. The average improvement was 17.3 seconds (*SD*=8). On average, in Trail 2, the participants spent 63.6 seconds (*SD*=35.1) completing each task, ranging from 47.3 seconds for the least complex ones (*SD*=19.7) to 90.3 seconds for the most complex ones (*SD*=41).

As well as this overall improvement, the participants' performance in Trial 2 showed a dramatic change in group-specific trends, as compared to Trial 1. Specifically, we found that the participants

**Table 2: Mean and Standard Deviation of the Task Completion Time per task.**

| | Trial1 | | Trial2 | | Overall | |
|---|---|---|---|---|---|---|
| Guidance | Mean | SD | Mean | SD | Mean | SD |
| Control | 84.77 | 55.72 | 68.18 | 40.25 | 76.24 | 48.93 |
| FAIL-EXMP | 77.62 | 51.91 | 68.50 | 38.43 | 73.09 | 45.79 |
| FAIL-RL | 68.11 | 35.61 | 51.54 | 25.70 | 60.15 | 32.23 |
| ONB-EXMP | 75.87 | 45.97 | 71.06 | 42.83 | 73.45 | 44.34 |
| ONB-RL | 80.03 | 42.62 | 56.70 | 23.35 | 67.82 | 36.84 |
| TASK-EXMP | 69.93 | 35.96 | 58.67 | 26.74 | 64.30 | 32.09 |
| TASK-RL | 84.70 | 50.13 | 60.46 | 23.50 | 72.50 | 40.80 |
| REQ-EXMP | 88.43 | 58.89 | 66.49 | 37.39 | 77.15 | 50.08 |
| REQ-RL | 91.12 | 56.05 | 70.02 | 43.02 | 80.18 | 50.67 |

**Table 3: Regression table for fixed effects. The group that took the longest to complete the task (here, the Upon-request/Rule group) was coded as the reference group. *p<.05, **p<.01, ***p<.001**

| | Trial1 | | | | | Trial2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Guidance | Estimates | SD | df | Z | p | Estimates | SD | df | Z | p |
| (Intercept) | 1.973 | 0.031 | 206.193 | 64.274 | 0.000 *** | 1.917 | 0.049 | 588.100 | 39.363 | 0.000 *** |
| REQ-EXMP | -0.023 | 0.038 | 120.323 | -0.594 | 0.554 | -0.018 | 0.039 | 110.590 | -0.471 | 0.639 |
| Control | -0.035 | 0.038 | 118.931 | -0.927 | 0.356 | -0.117 | 0.039 | 115.416 | -2.998 | 0.919 |
| FAIL-EXMP | -0.075 | 0.037 | 114.986 | -2.022 | 0.046 * | -0.008 | 0.039 | 113.999 | -0.205 | 0.838 |
| FAIL-RL | -0.114 | 0.037 | 112.059 | -3.095 | 0.002 ** | -0.004 | 0.039 | 110.725 | -0.102 | 0.003 ** |
| ONB-EXMP | -0.08 | 0.037 | 113.078 | -2.179 | 0.031 * | -0.003 | 0.038 | 110.055 | -0.073 | 0.942 |
| ONB-RL | -0.045 | 0.038 | 119.747 | -1.196 | 0.234 | -0.078 | 0.038 | 110.903 | -2.041 | 0.044 * |
| TASK-EXMP | -0.102 | 0.037 | 111.114 | -2.777 | 0.006** | -0.057 | 0.038 | 109.684 | -1.498 | 0.137 |
| TASK-RL | -0.029 | 0.037 | 114.725 | -0.783 | 0.435 | -0.034 | 0.038 | 111.785 | -0.881 | 0.380 |
| Task order | -0.02 | 0.005 | 572.924 | -4.366 | 0.000 *** | -0.014 | 0.004 | 586.162 | -3.188 | 0.002 ** |

in example-based guidance groups did not improve their task-completion times as their counterparts in the rule-based groups did, with the REQ groups being the only exception. This implies that example-based guidance allowed users to perform tasks efficiently from the start, but did not substantially help them learn how to use the chatbot. In contrast, particularly large improvements were made by the members of the ONB-RL, TASK-RL, and REQ-RL groups; and in comparisons against the ONB-EXMP group in particular, these differences were statistically significant (ONB-RL vs. ONB-EXMP: $t(117)$=-2.57, $p$=0.012; TASK-RL vs. ONB-EXMP: $t(117)$=-2.751, $p$=0.007; REQ-RL vs. ONB-EXMP: $t(117)$=-2.932, $p$=0.004).

It should be noted here that our statistical analysis of improvement was conducted on a participant level (i.e. whether a given individual's task completion time improved in Trial 2 as compared to Trial 1) rather than on a task level. This was because there was no one-to-one correspondence between tasks: the order of the tasks was randomly assigned, and some participants successfully accomplished fewer tasks in Trial 1 than in Trial 2. Probably because this level change shrank the sample size, we did not see as many improvement results that reached the .05 significance level. Yet, we regard some improvements as noteworthy, even where no statistical significance was found. For example, participants had markedly more improvement in REQ-EXMP than in other example-based groups. This implies that showing examples when requested might yield better learning outcomes than showing examples at other times. Likewise, the least improvement was among members of the ONB-EXMP group. As seen in Figure 3b, more than half the participants in that group spent more time on Trial 2 than on Trial 1. ONB-EXMP was also the only group for which average improvement was negative.

## 4.3 Number of Encountered Non-progress

*4.3.1 Number of Non-progress Events, Trial 1.* The overall trend in non-progress was similar to that of task-completion time. Participants on average encountered 0.25 non-progress events per task ($SD$=0.61), ranging from 0.17 on the least complex tasks ($SD$=0.51) to 0.38 for the most complex ones ($SD$=0.72). As Figure 4a shows, most of the eight guidance groups encountered significantly fewer non-progress events than the control group in Trial 1, the exceptions being the REQ groups, which, as mentioned in the previous section, were also two of the worst-performing groups in terms of task-completion time. Both TASK groups and the ONB-EXMP group encountered the fewest non-progress events, and significantly fewer than the control group (TASK-EXMP: $t(113.83)$=-3.68, p<0.001; TASK-RL: $t(117.38)$=-3.25, $p$=0.002; ONB-EXMP: $t(115.77)$=-3.19, $p$=0.002). Participants in both REQ-EXMP and REQ-RL also encountered relatively more non-progress events, with REQ-RL group encountered significantly fewer than TASK-EXMP($t(112.89)$=-2.139, $p$=0.011).

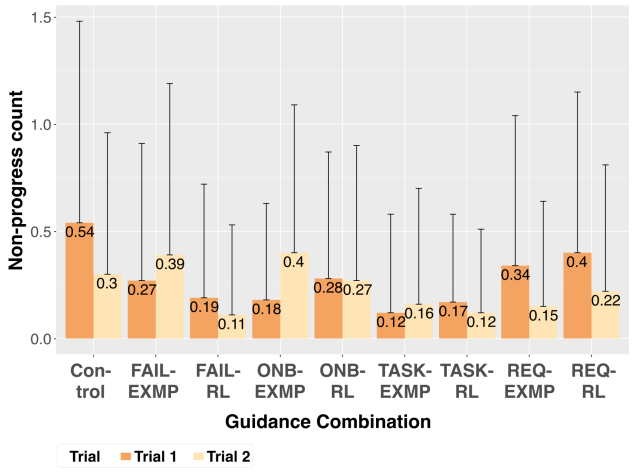
*4.3.2 Number of Non-progress Events, Trial 2.* Despite the participants having had relatively few conversational failures in either trial, we still observed change in the number of non-progress events between trials. The overall trend remained similar to that of task-completion time. Whereas participants in all rule-based guidance groups encountered fewer non-progress events in Trial 2 than in Trial 1 (ONB-RL: $M$=-0.03, $SD$=0.40; TASK-RL: $M$=-0.06, $SD$=0.23; REQ-RL: $M$=-0.21, $SD$=0.40; FAIL-RL: $M$=-0.07, $SD$=0.21), participants in three of the four example-based guidance groups did

**Table 4: Mean and Standard Deviation of Number of Encountered Non-progress per task.**

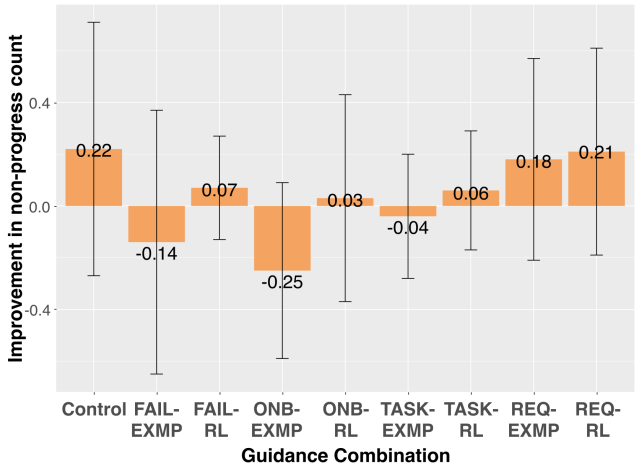
| | Trial1 | | Trial2 | | Overall | |
|---|---|---|---|---|---|---|
| Guidance | Mean | SD | Mean | SD | Mean | SD |
| Control | 0.54 | 0.94 | 0.3 | 0.66 | 0.42 | 0.82 |
| FAIL-EXMP | 0.27 | 0.64 | 0.39 | 0.80 | 0.33 | 0.73 |
| FAIL-RL | 0.19 | 0.53 | 0.11 | 0.41 | 0.15 | 0.48 |
| ONB-EXMP | 0.18 | 0.45 | 0.40 | 0.69 | 0.29 | 0.59 |
| ONB-RL | 0.28 | 0.59 | 0.27 | 0.63 | 0.27 | 0.61 |
| TASK-EXMP | 0.12 | 0.46 | 0.16 | 0.54 | 0.14 | 0.50 |
| TASK-RL | 0.17 | 0.41 | 0.12 | 0.39 | 0.14 | 0.40 |
| REQ-EXMP | 0.34 | 0.70 | 0.15 | 0.49 | 0.25 | 0.62 |
| REQ-RL | 0.40 | 0.75 | 0.22 | 0.59 | 0.31 | 0.67 |

**Table 5: Regression table for fixed effects. The group that encountered the most non-progress events in a single task (i.e., the Control group in the case of Trial 1, and the After-failure/Example group for Trial 2) was coded as the reference group. $^{*}$p<.05, $^{**}$p<.01, $^{***}$p<.001.**

| | Trial1 | | | | | Trial2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Guidance | Estimates | SD | df | Z | p | Estimates | SD | df | Z | p |
| (Intercept) | 0.508 | 0.096 | 212.5 | 5.272 | 0.000 *** | 0.530 | 0.144 | 603.203 | 0.677 | 0.000 *** |
| Control | - | - | - | - | - | -0.102 | 0.113 | 120.022 | -0.896 | 0.372 |
| FAIL-EXMP | -0.267 | 0.115 | 117.623 | -2.33 | 0.026 * | - | - | - | - | - |
| FAIL-RL | -0.351 | 0.114 | 114.791 | -3.079 | 0.003 ** | -0.291 | 0.114 | 114.872 | -2.557 | 0.012 * |
| ONB-EXMP | -0.363 | 0.114 | 115.768 | -3.184 | 0.002 ** | 0.007 | 0.112 | 114.365 | 0.061 | 0.951 |
| ONB-RL | -0.26 | 0.116 | 112.231 | -2.237 | 0.027 * | -0.134 | 0.112 | 115.238 | -1.198 | 0.233 |
| TASK-EXMP | -0.418 | 0.114 | 113.83 | -3.678 | 0.000 *** | -0.236 | 0.112 | 116.192 | -2.116 | 0.036 * |
| TASK-RL | -0.372 | 0.115 | 117.379 | -3.238 | 0.002 ** | -0.286 | 0.112 | 114.093 | -2.541 | 0.012 * |
| REQ-EXMP | -0.196 | 0.117 | 112.781 | -1.675 | 0.097 | -0.250 | 0.114 | 114.718 | -2.192 | 0.030 * |
| REQ-RL | -0.131 | 0.115 | 117.039 | -1.144 | 0.255 | -0.181 | 0.111 | 112.348 | -1.630 | 0.106 |
| Task order | 0.009 | 0.014 | 571.239 | 0.644 | 0.520 | -0.014 | 0.013 | 0.000 | -1.086 | 0.278 |




(a)




(b)

**Figure 4: (a) Bar chart of the non-progress events encountered in each task, by trial. The numbers represent the means for each combination. A bar chart instead of a boxplot was used here because in the majority of tasks, the number of non-progress events was zero; (b) Bar chart of improvement in non-progress events. Numbers represent the mean for each guidance/timing combination.**

not show as much improvement, or even experienced more non-progress in Trial 2 than in Trial 1 (ONB-EXMP: $M$=+0.22, $SD$=0.29 , TASK-EXMP: $M$=+0.04, $SD$=0.24; REQ-EXMP: $M$=-0.30, $SD$=0.40; FAIL-EXMP: $M$=+0.14, $SD$=0.51). In particular, the ONB-EXMP group encountered the largest increase in non-progress events in Trial 2. This performance was significantly worse than that of the REQ-RL group ($t(117)$=3.24, $p$=0.002), the REQ-EXMP group ($t(117)$=3.28, $p$=0.001), the FAIL-RL group($t(117)$=2.41, $p$=0.018) and the control group ($t(117)$=3.34, $p$=0.001). The increase in non-progress encountered by the FAIL-EXMP group was also significantly larger than that in the REQ-RL group ($t(117)$=2.45, $p$=0.016), the REQ-EXMP group ($t(117)$=2.48, $p$=0.014) and the control group ($t(117)$=2.54, $p$=0.012).

These results suggest that presenting example-based guidance after failure and at service onboarding may not help the participants learn how to use the chatbot effectively.

In contrast, both REQ groups showed a distinct trend of improvement, as they also had in terms of task-completion time; that is, they encountered significantly fewer non-progress events in Trial 2 than in Trial 1. This again suggests that showing guidance after the user requested it improved their subsequent performance. In particular, showing examples upon request was associated with the

**Table 6: Guidance/timing combinations' preference ranking scores and relative frequencies of being ranked in the top-three and bottom-three preferred combinations. Higher rank scores represent greater preference**

| Scenario | Median | Mean | SD | Relative frequency of top-three placement | Relative frequency of bottom-three placement |
|---|---|---|---|---|---|
| TASK-EXMP | 8 | 7.33 | 1.81 | 71.43% | 5.56% |
| REQ-EXMP | 6 | 6.10 | 1.99 | 43.65% | 10.32% |
| TASK-RL | 6 | 5.94 | 2.22 | 44.44% | 19.05% |
| ONB-EXMP | 5 | 5.33 | 2.44 | 41.27% | 27.78% |
| REQ-EL | 5 | 5.25 | 2.04 | 31.75% | 23.81% |
| ONB-RL | 5 | 4.94 | 2.17 | 23.02% | 25.40% |
| FAIL-EXMP | 3 | 4.32 | 2.32 | 22.22% | 51.59% |
| FAIL-RL | 3 | 3.93 | 2.12 | 15.08% | 51.59% |
| Control | 1 | 1.87 | 2.08 | 7.14% | 84.92% |

largest improvement among the example-based guidance groups, and this improvement was significantly larger than that of both the FAIL-EXMP group ($t(117)$=-2.54, $p$=0.012) and the ONB-EXMP group ($t(117)$=-3.33, $p$=0.001). REQ-RL participants also achieved substantial improvement, which was also significantly larger than that of the FAIL-EXMP group ($t(117)$=-2.45, $p$=0.016) and the ONB-EXMP group ($t(117)$=-3.24, $p$=0.002). It should also be noted that the REQ groups' improvement did not equate to substantially better performance than the EXMP groups in Trial 2, because the former encountered relatively more non-progress events in Trial 1. In other words, their more marked improvement merely led them to converge to a similar level of non-progress as the EXMP groups by the end of the experiment. The control-group members also made large improvements, but likewise, this did not lead to notably high performance in Trial 2, due to their initially high level of non-progress.

Overall, the TASK-EXMP, TASK-RL, and FAIL-RL remained the top-performing groups as measured by fewest non-progress.

## 4.4 Rankings

The participants' rankings of their ideal combination of guidance type and timing were quite diverse. Each combination, and the control group, had its advocates, and all of the nine were ranked in the top three and in the bottom three by at least one participant. However, there was still a relatively clear pattern in their preferences.

We computed the preference score for each guidance type/timing combination from all participants by assigning nine points to each person's favorite, and one point to his/her least favorite. The average scores of all combinations are shown in Table 6. The table also shows the frequency with which each combination was placed in participants' top-three and bottom-three positions. A Friedman test [25] indicated that differences among the rankings were statistically significant ($\chi^2(8)$=320.188, p<.001). Post hoc pairwise comparison was conducted using Wilcoxon signed-rank tests [75] with Bonferroni correction

The Wilcoxon tests showed that TASK-EXMP was unarguably the most-favored guidance type/timing combination. As well as being the highest overall, its score was significantly higher than those of all the other combinations (TASK-EXMP vs. TASK-RL: $Z$=4.91, $p$<.001; vs. REQ-EXAMP: $Z$=5.53, $p$<.001; vs. REQ-RL: $Z$= 6.132, $p$<.001; vs. ONB-EXAMP: $Z$=6.905, $p$<.001; vs. ONB-RL: $Z$=7.02, $p$<.001; vs. FAIL-EXAMP: $Z$=7.76, p<.001; vs. FAIL-RL: $Z$=7.71, $p$<.001; vs. Control: $Z$= 9.37, $p$<.001). The frequency with which TASK-EXMP was placed in the top three was also overwhelming: nearly 28% greater than the next highest contender, REQ-EXAMP.

The least preferred combinations were ONB-RL ($M$=4.94, $SD$=2.17), FAIL-EXMP ($M$=4.32, $SD$=2.32), and FAIL-RL ($M$=3.93, $SD$=2.12), all of which scored only better than the control group (Control vs. ONB-RL: $Z$= - 7.755, $p$<.001; vs. FAIL-EXMP: $Z$= -7.265, $p$<.001; vs. FAIL-RL: $Z$= -7.01, $p$<.001). The two FAIL groups were both ranked in the bottom three by nearly half of the participants. FAIL-RL – which had enabled the participants to complete tasks more efficiently than any other guidance/timing combination – was only ranked in the top three 15% of the time, and received the lowest ranking overall.

Another pattern was that the example-based guidance combinations received significantly higher preference rankings than their rule-based counterparts only at the TASK and REQ timings (TASK: $Z$=4.905, $p$<.001; REQ: $Z$=3.33, $p$<.001). Example-based guidance at the ONB and FAIL timings, on the other hand, received preference rankings similar to their rule-based counterparts (ONB: $Z$=1.281, $p$=.20; FAIL: $Z$=1.745, $p$=.081). The fact that ONB-EXMP and FAIL-EXMP were not more favored than their rule-based counterparts resonated with their poorer performance in helping participants learn the chatbot, as described in the previous sections. Finally, within a given guidance type, the participants generally liked receiving guidance at the TASK timing the most, and at FAIL the least.

## 4.5 Satisfaction

Only those participants who had seen guidance completed the Explanation Satisfaction Scale. ANOVA showed significant main effects of guidance type/timing combination on understandability ($F(7)$=3.21, $p$=.005) and usefulness ($F(7)$=2.39, $p$=.029). Tukey post-hoc testing revealed that TASK-EXMP was rated significantly higher than ONB-RL for understandability (4.86±0.4 vs. 3.67±1.1, $p$=.03<.05). For usefulness, REQ-EXMP was rated significantly more useful than the FAIL-RL (4.9±0.3 vs. 3.4±1.3, $p$=.048<.05). We did

not observe any statistically significant differences in satisfaction ratings.

# 5 QUALITATIVE FINDINGS

Below, we present the qualitative data related to task efficiency first, followed by those on learning how to use the chatbot; the specifics of guidance/timing combinations; and other desired characteristics of guidance.

## 5.1 Task Efficiency

Most participants expressed concern about efficiency. Some said that they would rather use other alternatives if a task-oriented chatbot did not allow them to efficiently complete the task. As P44 put it, "*If I feel like using it does not make me more efficient, then I rather book the tickets via a website.*" Participants tended to conceive of efficiency in three aspects: time, physical effort, and cognitive effort. The first refers to spending less time on completing the tasks; the second, to typing as few words or performing as few actions as possible; and the third, to using fewer mental resources to process or think about how to complete the tasks. The participants wanted to complete the tasks in as few turns as possible to save time, such as by "*providing all relevant information in one message*" (P17), so that they could "*use the minimum number of conversations [i.e., conversational turns] in the shortest time to get the information*" they wanted (P23). As P79 noted, "*The [example] guidance just showed me how to put all information like date, location, and my task in one sentence, so that I didn't need to figure this out for myself.*" To these participants, "*If one message would suffice, I'd just use one*" (P25).

In terms of physical efficiency, 28.6% (16/56) participants who received the example-based guidance tended to copy-and-paste the entire example, and then change only the key details, so that they would have to type fewer words and thus avoid feeling "*tired*" (P46). Some participants mentioned that this copy-and-pasting strategy also reduced their cognitive effort in processing how to compose a message, resulting in better cognitive efficiency. P15, for instance, said: "*I just needed to change 12/1 to 10/10, which did not cause me any burden because I didn't have to think; it's like editing a template*". The common use of this strategy underlay the participants' dislike of examples presented at service-onboarding, because that timing meant they needed to "*keep scrolling up*" (P37) to access the guidance. P36 specified that service-onboarding examples were inferior to those that appeared upon request, saying of the latter: "*I can immediately see them when I open them, and I just copy them.*"

Several participants also mentioned that examples were easier to understand and absorb because, as compared to rules, they were more "*colloquial*" (P20, P22, P34). In contrast, rules were regarded as demanding comprehension and transformation into one's own words, which was "*more troublesome*" (P116).

Despite the efficiency gains generally associated with the copy-and-paste method, several participants said that it made them inattentive, occasionally resulting in more time being spent on repairing their conversations, because they had "*missed some information and had to start over*" (P116). On the other hand, some participants preferred processing less text, and thus preferred rule-based guidance. As P92 noted, "*I still think examples look 'blah-blah'. I don't want to read them unless I can't type anything myself.*" In contrast, rule-based guidance was deemed cleaner and more organized, and made some participants feel "*less impatient and irritated*" (P97).

## 5.2 Improvement in Performance

Learning how to use the chatbot was also something that many participants said they cared about. Example-based and rule-based guidance each played a distinctive role in helping them with such learning.

Regarding example-based guidance, participants mentioned that they were not sure how to interact with the chatbot at first. Therefore, example-based guidance served as a good point of reference for message framing, flow, and elements. This helped them understand "*what a message should be like*" (P94), "*what to put in the message first*" (P14), and "what information is needed" (P65). As P52 said, examples "*let me understand how [the chatbot] works, so that I have a clearer and concrete picture. After knowing this, I just need to follow the rules*". P89 also explained that concise rules might not be sufficiently specific about the wording, terms, and format possibilities that users would wonder about: "*Like searching for seats in a certain period, it only mentions the format for times, but does not mention other time-related keywords that you should use in the message.*" P23 likewise said he hesitated about whether to use commas or spaces to separate statements, because this was not mentioned in the rules. In short, despite their specificity, examples demonstrated how a message should be built and what elements are essential to it.

On the other hands, the participants perceived the rule-based guidance as allowing them to learn how to "*quickly modify their expressions*" (P67) and save time that they might otherwise have spent dithering "*among the many formats*" that could be used. As P92 said, "*having this rule helps [. . .] I will not be wondering 'How I should put this to allow the chatbot to recognize it?'*" Similarly, P9 explained, "*It's easy to mix up date and number, or some special rules which you may not know in advance. Indicating where I might make a mistake saves me a lot of trouble.*"

More importantly, many participants mentioned having to process and make sense of rules, and then think about how to compose a message on their own, which they did not need to do when guidance was example-based. As P15 commented, "*I need to comprehend the rules first and then convert them into sentences. It's not like using my own way to say it, but making sense of the rules and then producing what [the chatbot] needs.*" Similarly, P95 said: "*Examples are more like scenarios, which are easier to understand. But rules are something you have to digest*". Despite requiring more time and effort initially, such processes over time enabled participants to communicate with chatbots more smoothly than their counterparts who tended to copy example-based guidance.

Participants often mentioned that rules-based guidance is more helpful in the long run, because rules would be more applicable than examples across disparate tasks and even different chatbots, assuming that they shared similar, if not entirely identical, underlying mechanisms for understanding language. For example, P79 explained, "*Rules are more generalizable and universal. It works for looking up times, movie tickets, and so on.*" P23 also commented: "

*[Rules-based guidance] lets me customize my messages, and is therefore much better than giving me a specific example.*"

Finally, when guidance was provided upon request rather than upfront, many participants were inclined to explore the chatbot's capabilities by composing messages on their own first, instead of requesting examples or rules right away. P61 provided an explanation for this exploratory approach.: "*It's like you buy a new vacuum and you don't want to check the manual but try it first. [. . .] It's probably because I was confident in myself using [this chatbot].*" Possibly due to such preferences for and/or confidence in a trial-and-error approach, the participants took longer time to complete tasks with upon-request guidance. However, possibly also due to this exploration process, participants had great improvements in task performance.

## 5.3 Different Feelings and Perceptions about Guidance at Different Timings

Finally, echoing our quantitative results about how specific guidance/timing combinations had differential effects on task performance and improvement, our qualitative results show that the participants had feelings and perceptions that were unique to each such combination.

*5.3.1 Rules-based Guidance Seen as Reminders, vs. a Manual, vs. Maxims.* Several participants noted that rules-based guidance presented after conversational failures served more as a *reminder* or an *error message*, indicating where the message might have gone wrong, as opposed to guidance presented upfront, which served more as a manual for how to use the system (P124). P15 commented, "*it's nice when the chatbot tells me [its] rule after I make a mistake, as I feel like I'm being advised by the chatbot so I know how to move on.*" However, among those participants who perceived after-failure rules as error messages, some reported negative feelings such as frustration and worry. P52 said, "*It made me think of the red warning text box [. . .], and made me feel kind of frustrated and caught off-guard*". Similarly, it made P70 "*feel quite alarmed,*" as well as that the chatbot was "*somewhat pushy*".

On the other hand, rules-based guidance presented at service onboarding made some participants feel that the rules were *maxims* (P120) that they must obey when using the service. When they perceived the rules as communication maxims or orders to obey, participants tended to regard them as inflexible, as well as somewhat "*impolite*" (P124). As a service is supposed to encourage more usage, but rules at onboarding were perceived as having the opposite effect: i.e., discouraging use unless the rules were followed (P18), or the user passed a test (P91). As P18 put it, "*I've not even started using it yet! It made me feel that the chatbot might have strict limits, and would give me errors if I typed something.*" P91 powerfully critiqued the service-onboarding timing as not like interaction, "*but more like taking a test, and you need to pass a lot of them to get its help*", which was "*very troublesome.*"

*5.3.2 Example-based Guidance Seen as Template, vs. Sign of a Different Mentality, vs. Visual Clutter.* Examples were generally seen as templates that could be adapted to users' own tasks, or reference materials about how sentences can be composed to be understood by the chatbot. As P96 noted, "*Examples should be provided at the very beginning, so people can have a template, or like an easy guidebook to follow, especially for the kids or elderly*". Nonetheless, participants reacted to examples differently when they perceived it as arriving late, i.e., after failures. Specifically, many participants did not feel that seeing examples after failure helped them to recognize their conversational mistakes, because they were "*not specific enough*" (P25). Some even felt that they were being tricked or mocked: as P39 put it, "*I feel like I am being played by the chatbot; [...] I'm taking lots of turns to interact with it, and then it tells me that I can say everything in one sentence. Why didn't it tell me earlier, then?*"

Seeing examples only after they had strived to construct sentences in their own ways, but failed to make the chatbot understand them, made some participants realize that they were talking to "someone" with a different mentality. As P23 explained, "*When I found out that it failed to understand my message then told me how a message should look, I was thinking 'Your message was actually not as good as mine' [. . . ] I'd be a little annoyed when it showed me something that should have failed but worked. [. . .] Then you realize that you're not its audience; the way you talk to it is not the way it is intended to be used.*" P104 also complained, "*Must I type exactly the same as yours [i.e., the example] to make you get my meaning?*"). Interestingly, when examples were presented upfront, the participants did not challenge the logic in the same examples, or regard them with distaste or suspicion, but instead, tended to adopt them wholesale. Therefore, their negative feelings were probably provoked by a sense that after-failure examples negated or devalued messages they had expended considerable effort in crafting.

Despite generally preferring that examples be presented earlier, the participants thought it important that they not be presented so early as to become irrelevant. When seeing example-based guidance presented at service-onboarding, many participants perceived the examples as visual clutter and simply skipped over them, characterizing them as excessively long and wordy (P125), not pertinent (P116), trivial (92), and/or like "*spam text*" (P39). As P26 commented, "*Whenever I see a lot of text while onboarding I would rather skip it. It applies to all these examples at onboarding. They take up a lot of space, and even occupy half the screen if you use your phone*". As a result, most participants said they wished examples would be offered when they "*really need them when performing the task*" (P17).

# 6 DISCUSSION

Our mixed-methods study has yielded rich results regarding the empirical effects of eight guidance type/timing combinations, and also how users reacted subjectively to the characteristics of specific combinations. Below, we discuss these findings and their implications.

## 6.1 A Mismatch between Task Performance and User Experience

Our results show that presenting rules after failures, and presenting examples at the introduction of a task, allowed participants to accomplish their tasks the quickest. However, our participants' perceptions of these guidance type/timing combinations differed considerably. Specifically, the main advantages of presenting examples at task-intro were 1) allowing the participants to quickly start

their task by adapting the provided examples to their own needs, and 2) informing them in a timely manner about how they could integrate all the requests into a single message. Seeing examples at this point in the communication process aided their understanding of the structure of messages that the chatbot could deal with, and applying the examples saved them the time they would otherwise have spent guessing about this; and thus, they spent less time on their tasks. Probably due to these benefits, the TASK-EXMP combination received the highest satisfaction scores.

In contrast, rule-based guidance provided after a conversational failure allowed participants to perform their tasks quickest, but they nevertheless placed it last in their preference rankings, suggesting that the participants did not perceive themselves as having been helped by this guidance, because they would not see this guidance until they encountered a conversational breakdown. Some participants also mentioned that seeing rules after failures was like seeing error messages, perhaps another reason for this guidance type/timing combination being disfavored. These contrasts between TASK-EXMP and FAIL-RL indicate more generally that guidance features that lead to great performance do not necessarily lead to pleasant user experiences. Future chatbot designers may wish to consider including both these combinations, especially since helping users recognize, diagnose, and recover from errors is a widely acknowledged heuristic to help them accomplish their tasks [46, 59].

## 6.2 Examples Warranted a Good Start, whereas Rules Promoted Understanding

Our results show that, in Trial 1, when example-based guidance was provided early – i.e., either at task-intro or service-onboarding – the participants performed better than when rule-based guidance was provided at those time-points. As discussed above, this was because examples were more concrete to be adapted to the tasks. Examples were also considered more likely to contain specific details needed when composing a message, many of which were not included in the rules. Nevertheless, participants' performance, whether in terms of task-completion time or number of non-progress events, did not improve between trials in most example-based guidance groups. We observed that this was because many participants just lightly modified the examples, without thinking about the chatbot's underlying mechanisms, or how such mechanisms impacted how users' messages ought to be written, therefore they did not benefit from the worked-example effect [68].

In contrast, rule-based guidance led participants to take more time to execute tasks initially. However, probably due to repeated rehearsals of this higher level of processing, members of most rule-based guidance groups displayed significant improvement in their tasks in Trial 2. Indeed, some of the rule-based groups, despite having had inferior performance to the example-based groups with the same timings in Trial 1, performed better than them in Trial 2.

While multiple prior studies of intelligent user interfaces have recommended adopting examples as guidance content [10, 38, 40, 72], our results suggest that showing examples upfront may prompt chatbot users to simply adapt the examples to their immediate needs, rather than truly processing them as guidance. While this approach can boost their time efficiency, at least in the short term, it does not necessarily help them to develop a mental model of the chatbot and its capabilities. The distinct advantages of these two types of guidance have been mentioned in other fields, such as language learning [20], and our results show that these distinctions also hold when it comes to guiding people's communication with task-oriented chatbots.

## 6.3 The Timing of Providing Examples Matters

Our results also confirm the importance of guidance timing. While Jain [38] recommended that examples should be provided during every phase of chatbot interaction, our quantitative results indicate that example-based guidance delivered at different timings led to different task-efficiency and varying degrees of task-to-task improvement. In particular, we found that examples received during service onboarding were widely misperceived as not pertinent, and thus were not read carefully. Probably for this reason, the ONB-EXMP group improved the least between trials of any group, and its number of non-progress events actually increased from one trial to the next. Moreover, when attempting to apply examples to tasks, participants in the ONB condition found it inconvenient to repeatedly scroll up to the guidance. Therefore, while some prior work [3, 38] concluded that guidance should be placed at the initial stage of interaction, the present study specifically clarifies that the example-based guidance would be better provided at the introduction of the task and not at service-onboarding.

Among the example-guidance groups, in contrast, the most between-trial improvement was seen when the guidance was provided upon request. This was because the participants could best absorb this type of guidance when they actually needed assistance. Some also seemed to pride themselves on exploring the chatbot's capabilities without any guidance for as long as possible – an approach that naturally reduced the incidence of the copy-and-paste strategy discussed above. Such strategy enabled the participants to make sense of the mechanism of the chatbot, via the worked-example effect [66–68]. This seems to contradict the findings of Kirschthaler et al. [40], who compared the effectiveness of proactive and reactive guidance on VUIs and reported no significant difference between the two. However, they did not distinguish example-based from rule-based guidance, or have their participants undertake a series of different tasks.

Finally, presenting example-based guidance after failure resulted in worse performance and less improvement. This result is understandable, since one main advantage of examples is serving as templates for users' own messages to the chatbot, and not showing examples until after failure negates this advantage. It also led some participants to feel "played" by the chatbot, which could have fed FAIL-EXMP into receiving one of the worst ranking outcomes.

All in all, these results suggest that the timing of the provision of example-based guidance is crucial not only to users' performance but also to their subjective experience of chatbot use.

## 6.4 Implications for Chatbot Guidance Design

Our results have three key high-level implications. First, there was no clear "winner" in either the guidance type or timing. Instead, it is clear that choices of both type and timing should depend on the purpose of the guidance: chiefly, facilitating task execution vs.

promoting learning. Second, despite there being no "best" guidance type or timing, there were specific combinations that we would recommend against practitioners adopting: notably, showing example-based guidance at service-onboarding and after-failure. However, there may be specific cases in which even these combinations would be highly suitable, but which were simply not covered by our study. Third, in line with our study's original purpose, it is indeed beneficial to leverage the strengths of both example- and rule-based guidance at specific timings, tailored to chatbot designers' various needs.

One example of such tailoring is as follows. A chatbot provider can decide whether to make example-based guidance visible upon request, or visible at all times. Although the former requires an extra step by the user, it will encourage more exploration of the chatbot's capability, and thus promote learning among those users who want to explore. Importantly, for users who may have a need to accomplish a particular task as soon as possible, making the guidance button available at all times is important. It should be noted, however, that users may tend to neglect such buttons; thus, we recommend that the text introducing each task explicitly indicate where the user can access examples. More advanced systems could detect the device in use and the user's recent transportation activity[3] to determine whether the user would be efficiency-oriented or exploration-oriented at any given time-point. To accelerate task execution, such chatbots could even proactively shift to example-based guidance when they detect that the user is on the go.

It might also be desirable for example-based and rule-based guidance to be integrated together: e.g., each with its own button, located side by side, or in successive paragraphs, such as by listing rules first and then showing an example of how to compose a message using them. If the chatbot is complex and involves a number of rules, these should be presented progressively – e.g., with more detailed/advanced rules provided only upon further request – to avoid overwhelming the user.

Even in such a setup, however, we would recommend that a concise subset of key rules, perhaps in the form of a list of "common mistakes", be provided at task introduction in place of a long list of rules. It might even be worthwhile to list the most crucial rules as early as service-onboarding, but only if they are generalizable to multiple tasks; otherwise, users are likely to perceive them as visual clutter, and only skim or even ignore them. Whenever the conversation fails, the chatbot should provide rule-based guidance, framed as a reminder of where messages are most likely to go wrong. If the purpose is to help users efficiently accomplish tasks, we suggest the chatbot also provide likely options for users to proceed, following Ashktorab et al.'s [4] recommendation.

To sum up, this set of recommendations is based on an assumption that guidance should serve not only to facilitate the immediate task at hand, but also to help build mental models of the chatbot to support the user's completion of subsequent tasks. Different kinds of task-oriented chatbots will inevitably be used in different contexts, with some being more likely to be used on-the-go and on the phone, and others in static, desktop settings. Thus, it is important that practitioners take context of use into account when determining whether guidance should be more efficiency-directed or learning-directed. However, we would argue that learning-directed guidance is more scalable and sustainable, given that service providers may, at any time, expand their task sets or develop additional chatbots for delivering specific services. Empowering users to better understand the mechanisms underlying their chatbots will also make the former less reliant on examples when performing such new tasks, and working with such new chatbots. In the long run, as more conversational user interfaces become available, the better users understand them, the less common non-progress events will be. Thus, a good combination of guidance types and timings is likely to make a positive contribution to all chatbot services, not just the one that provides it.

[3] e.g., via the physical-activity sensors available from both Android and iOS

## 7 LIMITATIONS AND FUTURE WORK

The current paper is subject to the following limitations. First, it was a lab-based study, in which the participants were using the chatbots under conditions moderated by the researchers. Thus, they might not have not been using these systems as naturally as they would in real life. Second, in the reflection phase, participants generated their rankings after seeing example conversations of the other combinations, rather than actually using them. It is likely that their rankings would have differed at least somewhat if they had used them in the experiment. Third, despite our high-complexity tasks involving the entry of eight pieces of information, most participants managed to accomplish most tasks without encountering many conversational failures. Thus, the high efficiency we report for the after-failure timing might be an overestimate. Future research could usefully seek to replicate this study with more difficult and complex tasks. Fourth, the design tasks were limited to scenarios in two domains: movies and flights. As such, our findings may not be generalizable to other application areas (e.g., business, insurance, or support). Fifth, each participant only undertook two trials in total. Therefore, we cannot make claims about the participants' longer-term learning outcomes. Sixth, we did not explore the effectiveness of different content for the two types of guidance, and recommend that future research do so. Seventh, our participant pool was young-skewing and mostly urban, and thus the results may not be generalizable to dissimilar populations of chatbot users and potential chatbot users; and the sample size for each condition was relatively small (i.e., 14), so future researchers should consider a larger sample size. Last but not least, a few of our results were not supported by statistical significance, and thus our conclusions about them can only be tentative. Because this lack of significance might have been because of sample size, we mentioned all trends we deemed to be noteworthy to avoid type II (false-negative) error. Nevertheless, future research will be needed to validate any such conclusions.

## 8 CONCLUSION

Task-oriented chatbots have been increasingly popular in recent years, especially as complements to existing channels such as mobile apps and websites. As the chronic problem of human-chatbot conversational failures is unlikely to be resolved in the near future, or easily, it is vital to understand how this emerging medium can better support users task accomplishment. Our mixed-methods investigation of the impact of guidance types and timings on user communication with task-oriented chatbots was not only able to

identify patterns in the effectiveness of guidance type/timing pairings, but why these patterns were present. This, in turn, enabled us to generate a set of design recommendations for task-oriented chatbots. While we anticipate that those recommendations can benefit chatbot practitioners, we concede that – as the first study of its kind – it is merely a starting point; and we encourage future researchers to validate the effectiveness of the proposed designs in real-life settings.

## ACKNOWLEDGMENTS

We sincerely thank our study participants and Zi-Yi Li and Yu-Chi Hsiao for helping the study. This research was supported in part by the Ministry of Science and Technology, R.O.C (MOST 109-2218-E-009 -016), and by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan.

## REFERENCES

[1] Bill Albert and Tom Tullis. 2013. *Measuring the user experience: collecting, analyzing, and presenting usability metrics.* Newnes, Burlington, MA.

[2] James F Allen, Donna K Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2001. Toward conversational human-computer interaction. *AI magazine* 22, 4 (2001), 27–27.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, 1–13.

[4] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–12.

[5] Robert K Atkinson, Alexander Renkl, and Mary Margaret Merrill. 2003. Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of educational psychology* 95, 4 (2003), 774.

[6] Marion Boiteux. 2018. *Messenger at F8 2018.* Meta. Retrieved January 16, 2021 from https://blog.messengerdevelopers.com/messenger-at-f8-2018-44010dc9d2ea

[7] Petter Bae Brandtzaeg and Asbjørn Følstad. 2017. Why people use chatbots. In *International conference on internet science*. Springer, pringer, Cham, New York, NY, 377–392.

[8] Petter Bae Brandtzaeg and Asbjørn Følstad. 2018. Chatbots: changing user needs and motivations. *Interactions* 25, 5 (2018), 38–43.

[9] Raluca Budiu. 2018. *The user experience of chatbots.* Nielsen Norman Group. Retrieved August 16, 2021 from https://www.nngroup.com/articles/chatbots/

[10] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, 258–262.

[11] Donald J Campbell. 1988. Task complexity: A review and analysis. *Academy of management review* 13, 1 (1988), 40–52.

[12] John M Carroll and Caroline Carrithers. 1984. Training wheels in a user interface. *Commun. ACM* 27, 8 (1984), 800–806.

[13] Michael E Caspersen and Jens Bennedsen. 2007. Instructional design of a programming course: a learning theoretic approach. In *Proceedings of the third international workshop on Computing education research*. Association for Computing Machinery, New York, NY, 111–122.

[14] Richard Catrambone and John M Carroll. 1986. Learning a word processing system with training wheels and guided exploration. *ACM SIGCHI Bulletin* 18, 4 (1986), 169–174.

[15] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758.

[16] Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science* 13, 2 (1989), 145–182.

[17] Rikke Friis Dam and Teo Yu Siang. 2020. *Affinity Diagrams – Learn How to Cluster and Bundle Ideas and Facts.* Interaction Design Foundation. Retrieved September 2, 2021 from https://www.interaction-design.org/literature/article/affinity-diagrams-learn-how-to-cluster-and-bundle-ideas-and-facts

[18] Alan Dix. 2020. Statistics for HCI: Making Sense of Quantitative Data. *Synthesis Lectures on Human-Centered Informatics* 13, 2 (2020), 1–181.

[19] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech communication* 50, 8-9 (2008), 630–645.

[20] Nick Ellis. 1993. Rules and instances in foreign language learning: Interactions of explicit and implicit knowledge. *European Journal of Cognitive Psychology* 5, 3 (1993), 289–318.

[21] Dario Fiore, Matthias Baldauf, and Christian Thiel. 2019. " Forgot your password again?" acceptance and user experience of a chatbot for in-company IT support. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*. Association for Computing Machinery, New York, NY, 1–11.

[22] Asbjørn Følstad and Ragnhild Halvorsrud. 2020. Communicating Service Offers in a Conversational User Interface: An Exploratory Study of User Preferences in Chatbot Interaction. In *32nd Australian Conference on Human-Computer Interaction*. Association for Computing Machinery, New York, NY, 671–676.

[23] Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st international conference on conversational user interfaces*. Association for Computing Machinery, New York, NY, 1–9.

[24] Asbjørn Følstad and Cameron Taylor. 2019. Conversational repair in chatbots for customer service: the effect of expressing uncertainty and suggesting alternatives. In *International Workshop on Chatbot Research and Design*. Springer, Springer, Cham, New York, NY, 201–214.

[25] Milton Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics* 11, 1 (1940), 86–92.

[26] Google. 2021. *Conversation design.* Google. Retrieved August 16, 2021 from https://developers.google.com/assistant/conversation-design/welcome

[27] Frederick J Gravetter and Lori-Ann B Forzano. 2018. *Research methods for the behavioral sciences.* Cengage Learning, Boston,MA.

[28] Jonathan Grudin and Richard Jacques. 2019. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–11.

[29] Lakisha Hall. 2018. *6 steps to successful conversational design.* IBM. Retrieved August 16, 2021 from https://www.ibm.com/blogs/watson/2018/09/6-steps-to-successful-conversational-design/

[30] Kai Halttunen. 2003. Scaffolding performance in IR instruction: Exploring learning experiences and performance in two learning environments. *Journal of Information Science* 29, 5 (2003), 375–390.

[31] Kai Halttunen. 2011. Pedagogical design and evaluation of interactive information retrieval learning environment. In *Teaching and learning in information retrieval*. Springer, New York, NY, 61–73.

[32] Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures* 1, 1 (2007), 77–89.

[33] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects.

[34] Reid Holmes and Gail C Murphy. 2005. Using structural context to recommend source code examples. In *Proceedings of the 27th international conference on Software engineering*. Association for Computing Machinery, New York, NY, 117–125.

[35] Kristina Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with computers* 12, 4 (2000), 409–426.

[36] IBM. 2021. *IBM Cloud Docs/Watson Assistant.* IBM. Retrieved November 16, 2021 from https://cloud.ibm.com/docs/assistant?topic=assistant-dialog-slots

[37] Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. 2018. FarmChat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–22.

[38] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*. Association for Computing Machinery, New York, NY, 895–906.

[39] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. Association for Computing Machinery, New York, NY, 143–152.

[40] Philipp Kirschthaler, Martin Porcheron, and Joel E Fischer. 2020. What Can I Say? Effects of Discoverability in VUIs on Task Performance and User Experience. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. Association for Computing Machinery, New York, NY, 1–9.

[41] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* 15, 2 (2016), 155–163.

[42] Klaus Krippendorff. 2004. *Content analysis: An introduction to its methodology.* Sage publications, Washington, D.C.

[43] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

[44] Knut Kvale, Olav Alexander Sell, Stig Hodnebrog, and Asbjørn Følstad. 2019. Improving Conversations: Lessons Learnt from Manual Analysis of Chatbot Dialogues. In *International Workshop on Chatbot Research and Design*. Springer, Springer, Cham, New York, NY, 187–200.

[45] Suna Kyun, Slava Kalyuga, and John Sweller. 2013. The effect of worked examples when learning to write essays in English literature. *The Journal of Experimental Education* 81, 3 (2013), 385–408.

[46] Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic Evaluation of Conversational Agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–15.

[47] SeoYoung Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103 (2017), 95–105.

[48] Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–12.

[49] Jing Li, Aixin Sun, and Zhenchang Xing. 2018. Learning to answer programming questions with software documentation through social context embedding. *Information Sciences* 448 (2018), 36–52.

[50] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems.

[51] Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2010. Practical resources for assessing and reporting intercoder reliability in content analysis research projects.

[52] Ewa Luger and Abigail Sellen. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, 5286–5297.

[53] Michael Meng, Stephanie Steinhardt, and Andreas Schubert. 2018. Application programming interface documentation: what do software developers want? *Journal of Technical Writing and Communication* 48, 3 (2018), 295–330.

[54] Microsoft. 2021. *Best practices for building a language understanding (LUIS) app*. Microsoft. Retrieved August 16, 2021 from https://docs.microsoft.com/en-us/azure/cognitive-services/luis/luis-concept-best-practices

[55] Mohammed Slim Ben Mimoun, Ingrid Poncin, and Marion Garnier. 2012. Case study—Embodied virtual agents: An analysis on reasons for failure. *Journal of Retailing and Consumer services* 19, 6 (2012), 605–612.

[56] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R Cowan. 2018. Design guidelines for hands-free speech interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. Association for Computing Machinery, New York, NY, 269–276.

[57] Brad A Myers. 1986. Visual programming, programming by example, and program visualization: a taxonomy. *ACM sigchi bulletin* 17, 4 (1986), 59–66.

[58] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–7.

[59] Jakob Nielsen. 1994. *How to Conduct a Heuristic Evaluation*. Nielsen Norman Group. Retrieved June 10, 2021 from https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/

[60] Jakob Nielsen. 1994. *Usability engineering*. Morgan Kaufmann, San Francisco, CA.

[61] Jakob Nielsen. 2010. *Mental Models*. Nielsen Norman Group. Retrieved August 16, 2021 from https://www.nngroup.com/articles/mental-models/

[62] Jakob Nielsen. 2011. *Workflow Expectations: Presenting Steps at the Right Time*. Nielsen Norman Group. Retrieved September 2, 2021 from https://www.nngroup.com/articles/workflow-expectations/

[63] Alexander Renkl. 1997. Learning from worked-out examples: A study on individual differences. *Cognitive science* 21, 1 (1997), 1–29.

[64] Alexander Renkl. 2002. Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and instruction* 12, 5 (2002), 529–556.

[65] Alexander Renkl. 2014. Toward an instructionally oriented theory of example-based learning. *Cognitive science* 38, 1 (2014), 1–37.

[66] Alexander Renkl, Robert K Atkinson, and Cornelia S Große. 2004. How fading worked solution steps works–a cognitive load perspective. *Instructional science* 32, 1 (2004), 59–82.

[67] Alexander Renkl, Tatjana Hilbert, and Silke Schworm. 2009. Example-based learning in heuristic domains: A cognitive load theory account. *Educational Psychology Review* 21, 1 (2009), 67–78.

[68] Alexander Renkl, Robin Stark, Hans Gruber, and Heinz Mandl. 1998. Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary educational psychology* 23, 1 (1998), 90–108.

[69] Julian Roelle, Sara Hiller, Kirsten Berthold, and Stefan Rumann. 2017. Example-based learning: The benefits of prompting organization before providing examples. *Learning and Instruction* 49 (2017), 1–12.

[70] Ruhi Sarikaya. 2017. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine* 34, 1 (2017), 67–81.

[71] Agnieszka Sienkiewicz. 2021. *11 Chatbot Statistics and Trends You Need to Know in 2021*. Tidio. Retrieved January 16, 2021 from https://www.tidio.com/blog/chatbot-statistics/

[72] Arjun Srinivasan, Mira Dontcheva, Eytan Adar, and Seth Walker. 2019. Discovering natural language commands in multimodal interfaces. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, 661–672.

[73] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International journal of human-computer studies* 67, 8 (2009), 639–662.

[74] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.

[75] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*. Springer, New York, NY, 196–202.

[76] Koos Winnips and Catherine McLoughlin. 2001. *Six WWW based learner supports you can build*. Association for the Advancement of Computing in Education (AACE), Waynesville, NC.

[77] Svetlana Yarosh, Stryker Thompson, Kathleen Watson, Alice Chase, Ashwin Senthilkumar, Ye Yuan, and AJ Bernheim Brush. 2018. Children asking questions: speech interface reformulations and personification preferences. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*. Association for Computing Machinery, New York, NY, 300–312.