Evaluation of Conversational Agents for Aerospace Domain

Ying-Hsang LIU*
yingliu@sdu.dk
University of Southern Denmark
Kolding, Denmark

Alexandre ARNOLD†
Gérard DUPONT*
Catherine KOBUS*
François LANCELOT*
<name>.<surname>@airbus.com
AIRBUS AI Research, Toulouse France

ABSTRACT

The use of conversational agents within the aerospace industry offers quick and concise answers to complex situations. The aerospace domain is characterized by products and systems that are built over decades of engineering to reach high levels of performance within complex environments. Current development in conversational agents can leverage the latest retrieval and language model to refine the system's question-answering capabilities. However, evaluating the added-value of such a system in the context of industrial applications such as pilots in a cockpit is complex. This paper describes how a conversational agent is implemented and evaluated, with particular references to how state-of-the-art technologies can be adapted to the domain specificity. Preliminary findings of a controlled user experiment suggest that user perception of the usefulness of the system in completing the search task and the system's responses to the relevance of the topic are good predictors of user search performance. User satisfaction with the system's responses may not be a good predictor of user search performance.

CCS CONCEPTS

• **Human-centered computing** → **Laboratory experiments**; *Natural language interfaces*; • **Information systems** → *Search interfaces.*

KEYWORDS

Enterprise search; Conversational search; Aerospace industry; Conversational agent; Question answering; Evaluation protocol

1 INTRODUCTION

The aerospace industry relies on massive collections of documents covering system descriptions, manuals or procedures. Most of these are subjected to dedicated regulation and/or have to be used in the context of safety of life scenarios such as cockpit procedures for pilots. A user looking for specific information in response to a given situation in this large corpus is often seen spending a significant amount of precious time navigating through the documents. Even experienced pilots who are familiar with the structure of the documents can sometimes have difficulties in finding known items in a constrained time.

The dedicated structure of the information helps to quickly target a specific piece of information. The search system helps in any other case. However, they come with their limitations. Most of the time, it is the user's responsibility to adapt their search needs using specific keywords and/or syntax, known as the difficulty of articulating information needs [30, 51]. For simple queries that have a ready-made answer in the document, this is not always a difficult problem. However, for the understanding of complex procedures or troubleshooting system errors, it can lead to multiple queries and thus a cumbersome experience for the user.

Various types of systems are associated with conversational agents. A recent survey of different types of dialogue systems has identified three main types: task-oriented dialogue system, conversational agents, and interactive question-answering [14]. From the perspectives of human-computer interaction (HCI), user experience (UX), and information retrieval (IR), issues associated with the voice-based user interface, such as recognition error, user experience, and voice queries have gained traction recently [20, 32, 33].

In this study, our "Smart Librarian" (SL) mixed a task-oriented dialogue system with a conversational agent and an interactive question-answering component (with/without a voice-based interface). Specifically, the assistant is envisioned as a task-oriented system in a restricted domain, with mixed system/user initiatives and a multimodal interface to support situation awareness in a cockpit. Therefore, the evaluation objective is to assess the benefit of smart search and conversational search for cockpit documentation.

One of the primary objectives of conversational search systems is to enable the provision of information services in interactive styles, similar to human-human interactions in information-seeking conversations. User interfaces for conversational search systems ideally are similar to natural dialogue interactions [21] in which user's questions can be clarified during conversations. This thread of research has received much attention from the research communities of natural language processing, information retrieval, and human-computer interaction, just to name a few [2, 31, 34, 55].

This paper describes how a conversational agent is implemented and evaluated, with particular references to how state-of-the-art technologies can be adapted to the domain specificity in aerospace. We propose a user-centered approach to the design and evaluation of conversational search user interfaces (SUIs), termed Smart Librarian, to support the pilot in cockpits. The skills of assistants are intended to translate into the requirements of conversational search systems to support the tasks performed by the pilot. Our preliminary findings suggest that there were significant interaction effects between the task difficulty and the types of system; the Smart Librarian system performed well for difficult search tasks.

*Also with The Australian National University.
†Authors in alphabetical order.

Future directions for research and development of conversational agents are suggested.

2 RELATED WORK

2.1 Aviation Cockpits and Controls

In the aviation cockpits and controls environment, research has focused on the consideration of cognitive strategies and cognitive processing in a stressful environment for the design of automated support systems. For example, the role of cognitive processes inherent in the tasks and specific considerations of cognitive strategies adopted by the pilots for the design of automated support systems, i.e., automated cockpit have been emphasized [10]. In a study of how procedures such as Quick Reference Handbook (QRH) are used in a cockpit for emergencies, the results showed that "pilots employed strategies that interleaved a range of resources, often consulting fragments of the QRH checklists rather than following them from start to finish" [10, p. 147].

From the perspectives of human-computer interaction (HCI), an observational and interview study of cockpit activities for tangible design noted that "Pilots mention the value of having tools separated from the aircraft systems. Speaking about the physical QRH (Quick Reference Handbook), that you can hold in your hands, a pilot valued it in case of degraded contexts, when there is no longer control available." [27, p. 663]. And the usability of an information visualization system in flight to improve aviation safety has been evaluated in a flight simulator setting [3].

Overall, these studies suggest that systems designed for aviation cockpits and controls need to consider the issues regarding the role of cognitive processes inherent in tasks and cognitive strategies employed by pilots. The role of context in designing user interfaces and usability is also emphasized.

2.2 Information Seeking Conversation

Informed by theories of human-human communication and linguistics, IIR (interactive information retrieval) research has attempted to identify the purposes and communicative functions of elicitations (i.e., questions to request information) in information seeking conversations. User's elicitation behavior was found to be affected by individual differences, such as status, age, and experience, and interacted with situational variables, such as interaction time and the number of utterances [52]. Further studies have developed the concept of elicitation styles, characterized by linguistic forms, utterance purposes, and communicative functions, with particular references to user satisfaction [53, 54]. However, these findings have not been directly applied to the design of conversational search systems.

Recent studies have focused on developing system design guidelines from user studies. For example, in a study that observes people's interactions in a laboratory setting, researchers have compared human-human interactions and models of well-established search models to inform the design of spoken conversational search system [44]. And the system requirements for intelligent conversational assistants for improving user experience have been explored [50].

Following the paradigm of computers as social actors, a taxonomy of social cues of conversational agents based on interpersonal communication theories was built [18]. Using ethnomethodology and conversational analysis, the role of conversational interfaces in everyday life revealed the voice user interface design implications for the request and response design in embedded social interactions [35]. Together with IIR research, this thread of research contributes to our current discussions regarding the theories that can be borrowed from other disciplines and/or re-conceptualized to design conversational search systems.

2.3 Conversational Search System

System requirements for conversational search systems were defined as "a system for retrieving information that permits a mixed-initiative back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user" [37, p. 160]. An evaluation framework for conversational agents in the aerospace domain was proposed [4]. The research was conducted to identify the conversational styles for building computational models at scale for speech-based conversational agents [43]. These studies suggest the existing methods used to explore the conversational search systems from the perspectives of system design.

From technical perspectives, research on conversational search systems has focused on identifying user intent in information seeking conversations, designing user interfaces for different modes of interaction, and provision of clarification questions. For example, structural features (i.e., the position of an utterance in a dialogue) contribute the most to the identification of user intents, using neural classifiers [36]. The generation of clarification questions from community question-answering websites formulated the tasks as noun phrase ranking problems [9]. Neural models were used to generate clarification questions by considering sequences of purposes of interaction [1]. A formal model of information seeking dialogues that consists of the query, request, feedback, and answer for identifying the frequency of sequence patterns was proposed [48].

Overall, research and practice in conversational search systems have received lots of attention recently, but the usefulness of these systems has not been rigorously evaluated in the system design process from user perspectives.

2.4 Evaluation of Conversational Search System

A recent approach to the evaluation of conversational search systems, such as chatbots has intended to enhance user experience and thus select user satisfaction as the main evaluation criterion for success. For example, the Alexa Prize Socialbot Grand Challenge was designed as research competitions to advance our understanding of human interactions with socialbots, with the support of large amounts of user data from Amazon.com. This evaluation approach was derived from computer science research and AI perspective.

Within the NLP community, one of the key distinctions of evaluation approaches is the intrinsic and extrinsic evaluation of machine outputs [40]. The intrinsic evaluation focuses on the internal outputs from the system, whereas the extrinsic evaluation is concerned with how the use of the system contributes to external outputs,

https://developer.amazon.com/alexaprize

such as task completion. In the IR community, the evaluation efforts have focused on the creation of test collection to compare system performance, using appropriate evaluation metrics for different types of question-answering tasks. In this study, we take a holistic approach to understand user experience and user performance to bridge the gap between system-centric evaluation (i.e., automatic metrics) and human evaluation, using crowdsourcing platforms.

3 USER EXPERIMENT

3.1 Evaluation Objective

The alignment between system design requirements and evaluation objectives is important for a user-centered approach to system design and evaluation. Our evaluation objective is to *determine the relationship between the search tasks in the typical flight operation scenarios and the perceived usefulness of the system for task completion.*

3.2 Research Hypothesis

Research on user information seeking suggests that people's levels of domain expertise and experience, work roles, tasks, and procedures affect their information-seeking strategies and perceived usefulness of information resources [17, 19, 22, 28]. In the context of a safety-critical environment, information behavior research reveals that "Overly conditioned information behaviors, which would correspondingly limit methodical information behaviors, can lead crews to miss crucial steps in the process of projecting the future state of the aircraft and suitably planning ahead" [49, p. 1567]. Therefore, our proposed research hypotheses are as follows:

H1. Types of search systems and user perceptions will affect user search performance.
H2. Perceived search task difficulty and user perceptions will affect user search performance.

3.3 Research Design

In this study, since we focus on the design and evaluation of conversational search systems from user perspectives, we are concerned with user interactions with a prototype system in a laboratory setting. This approach has been adopted because we can 1) determine the relationship among the variables in a laboratory environment and 2) transfer the findings into specific system design decisions. This approach is scientifically rigorous when the experiment is conducted properly. However, it is very resource-intensive and time-consuming and requires different sets of expertise. And the results may be affected by the variability of individuals considerably [e.g. 41, 42, 51]. Specific examples include a flight simulator experiment with pilots co-designing system [3] and a turbulent touch design experiment with students [12].

The experiment protocol has been approved by the Toulouse University research ethics committee. The participant was presented with an informed consent form to sign-off before the experiment started.

3.4 Experiment Setting

The experiment was conducted in the environment of a flight simulator (ENAC BIGONE A320/A330 cockpit simulator) within the ACHIL platform. The setting was intended to create an environment that can elicit the information needs of participants, as suggested in simulated work task situations [8].

The subjects were given access to a tablet - similar to the ones used by the pilot in flight - to access the Flight Crew Operating Manual (FCOM) through one of the two systems: Smart Librarian (SL) and electronic flight bag (FB). This source document incorporates aircraft manufacturer guidance on how to use the systems onboard the aircraft for enhanced operational safety, as well as for increased efficiency. Overall it can be seen of several PDF documents counting several thousand pages.

3.5 Search Task

In designing search tasks we have considered the complexity of tasks from the perspectives of search as learning by classifying the search scenarios as easy and complex [46]. User perceived search task complexity after using the system [28] was assessed by a questionnaire.

Specifically, the easy task involves fact-finding while the hard task requires a higher level of understanding of the problems and/or some cognitive reasoning for answering the questions. In easy search tasks, the problem description contains relevant words that can be used to craft the "best question" pointing to a unique procedure (or document unit) that contains the solution. By contrast, in hard search tasks, the problem description does not contain any words matching the "best question" and the subject will need to rephrase the problem. Moreover, the user needs to explore at least two document units to find the answer. There is a need to reformulate the problem with new words/question and at least two document units are necessary to find the solution to the problem. Several successive questions are needed to identify the solution (See Table 1).

For each task, the ground truth has been defined by a set of domain experts by pointing the exact expected answer(s) and the exact procedure(s) in which these can be found in the FCOM document. The very narrow of the aeronautical domain and the particular form of documents, allowed us to ensure that the answers are unique for each task and that their location in the documentation is unique.

3.6 Arrangement of Experimental Conditions

Tasks were presented to subjects following a traditional Graeco-Latin square design [24, 25]. This study is a 2 × 2 design with two types of search systems (SL and FB) and two types of search tasks (easy and hard), to minimize the effect of presentation order of treatments [25].

3.7 Metrics

Search performance: The tasks defined are pure goal-oriented search task: the user is asked to find the exact answer and locates the procedure used. Classic precision and recall metrics used in information retrieval do not apply in this context and the score used can be seen as a Boolean success metric based on the expert ground truth (one could note it is similar to the precision@1 - but measured based on user's response).

Thus performance was evaluated through [0;1] scores for each step in the tasks (finding the right procedure, finding the right

Label	Level	Title	Initial Trigger/Message	Flight Condition
Tutorial	Easy	Captain's duty	N/A	cruise
Task A	Easy	Cockpit windshield cracked	Bird strike/window crack	cruise FL370
Task B	Easy	Bomb on board	N/A	cruise
Task C	Hard	ALL ENGINE FAILURE over the sea	ALL ENGINE FAILURE	cruise flying FL350 over the ocean, >70NM from coast
Task D	Hard	Air too hot in the cockpit	Air too hot	cruise FL370
Bonus	Hard	Engine fire over mountain	ENG 1 FIRE	climbing over the Alps in FL350

Table 1: Difficulty level of search tasks, with description and key aspects (FL means 'Flight Level')

Notes: The first task familiarizes the participant with the experiment setup, whereas the final task introduces the participant to the setup of a flight simulator.

answer to the situation in the procedure, finding the next procedure, etc.). Since hard tasks had more steps, the score was averaged in a single task score in [0;1] for each task (1 being the maximum score). This scoring strategy relies primarily on the task that can be understood from the user's point of view as a fact-finding problem. The classic precision and recall measures of the system are only taken into account through the lens of the user's selection in this interactive experimentation.

This can be seen as a quantitative metric with 1 data point per user per search task.

User's perception of the problem and system: Other metrics were collected through post-search questionnaires after each task and a final exit questionnaire. They were designed using 5 points Likert scale to collect the subject's perception of task difficulty and familiarity as well as system relevance and usefulness. These questionnaires were submitted by each user after each task and each system usage.

3.8 Prototype System

In our user experiment, we have developed a prototype Smart Librarian system to address the evaluation objective of determining the relationship between the types of search tasks and the perceived usefulness of search. The system was built around three main components:

- A dialog engine (based on RASA platform [7]) handling the conversation and identifying user's intents;
- A search engine (based on Solr [45]) where the documents collection is indexed following the BM25F relevance framework [39];
- A QA engine, based on a BERT large model [15], fine-tuned using the FARM framework . A multi-task setup was used for the fine-tuning: one task is the classical QA task (detecting the span of text) on SQUAD 2.0 dataset [38]; the other is a classification task (i.e. whether the answer to the question is contained or not in the document extract).

On top of these, additional capabilities to process speech inputs and produce speech outputs are available as an alternative to the traditional textual input. Figure 1 offers an overview of the whole architecture.

The whole system is made available through a reactive web interface enabling conversation and document exploration (See Figure 2). It was deployed in a cloud environment and made available to users through a tablet.

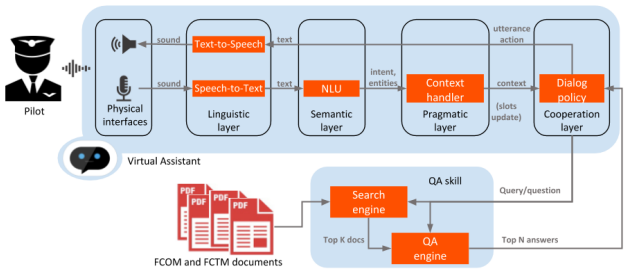

Figure 1: Overview of the prototype architecture.

The reference system, electronic flight bag (FB) was the Navblue software used by many pilots in commercial flight. It is distributed on tablets and customized for each aircraft. Only the library features (for access to documents and procedures) were used in the experiment.

3.9 Participants

We recruited students from an aviation school, currently in their early years of training for becoming commercial aircraft pilots. These students have a good understanding of aircraft technologies and flying physics. At the time of the experiments, they would not have followed a particular course on a specific aircraft, such as A320.

3.10 Evaluation Protocols

We have developed realistic scenarios for our user experiment by engaging with engineers and consulting with an ergonomic expert with specialties in designing systems for pilots (See Table 1). The simulated work task situations toolkit has been followed for triggering user information needs and the evaluation of user search behavior and system performance [8]. In designing search tasks we have considered the complexity of tasks from the perspectives of search as learning [46]. Several questionnaires, including a demographic questionnaire, post-search, and exit questionnaires, have been developed to assess the user perceptions during the search process as well as overall perceptions about the whole interaction

https://github.com/deepset-ai/FARM

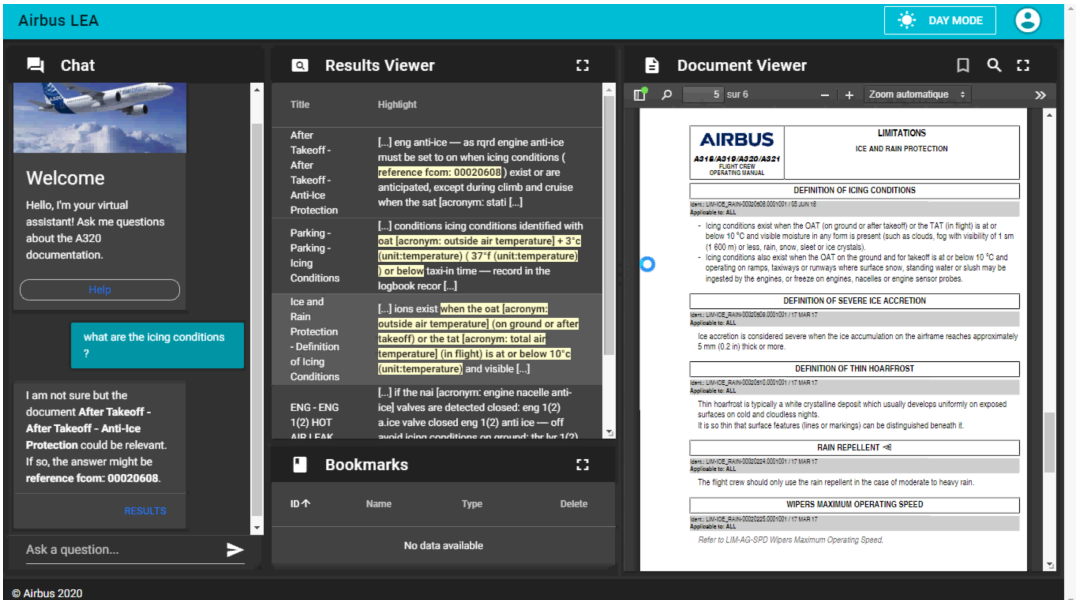

Figure 2: Screenshot of the prototype showing conversation on the left, search result panel in the center and document view on the right.

process. Finally, to ensure that the training for each participant is consistent across all sessions, experimental guidelines have been developed and used in our experiment.

4 DATA ANALYSIS

We construct mixed-effects models for determining the effects of system and user perceptions on search performance. Mixed-effects distinguish between fixed effects that are due to experimental conditions and random effects that are due to individual differences in a sample. We are concerned with both fixed effects of system and user perception and random effects of individual differences.

We choose the mixed-effects models because they are useful for the examination of the random effects of subjects and search tasks [5]. Examples of mixed-effects models in information retrieval research have included modeling of search topics effect [11], analysis of eye gaze behavior [23], and user characteristics [29]. We primarily use the lme4 package in R statistical computing software for model fitting [6].

We find that there was no significant relationship between the task order and the time spent ($R = 0.012, p = 0.92$). In addition to considering the fixed effects of system and user perception, the random effects of search task and user were considered in our full model construction and data fitting. To fit the data, we performed an automatic backward model selection of fixed and random parts of the linear mixed model [26]. Since the random intercepts for task and user were significant for time spent, with $p < .001$ and $p < .01$ respectively, we chose a mixed-effects model with search task and user-controlled as random effects. Model assessments based on diagnostic checks for non-normality of residuals and outliers, distribution of random effects, and heteroscedasticity were conducted. The random intercepts for the task were significant for task score, with $p < .001$, we chose a mixed-effects model with search task as random effects.

5 RESULTS

5.1 Participant Characteristics

A total of 16 pilot school students participated in the study. Most students were between 18 and 25 years old. Almost all students had flying experiences as an amateur or student pilot (less than 70 flight hours on average), and three students had general aviation experiences for more than 5 years. None of them had commercial flying experiences. Most participants used search engines every day or several times a day or more, whereas they had limited experiences using virtual assistants. Overall, the participants are homogeneous by experiences and age.

5.2 Search Performance by Search Task Difficulty

The overall results suggest that there was no significant difference in search performance by task score and time spent. However, there were very significant differences in search performance by search task. Figures 3 and 4 indicate that the proposed SL (Smart Librarian) system enhanced task score for hard search tasks, but there was no significant difference by time spent. As expected, task difficulty had significant effects on search performance. The SL system performed particularly well for hard search tasks.

5.3 Effect of System and Perception on Task Score

Table 2 presents the results of model selection for the mixed-effects of system and user perception on task scores. It shows that the

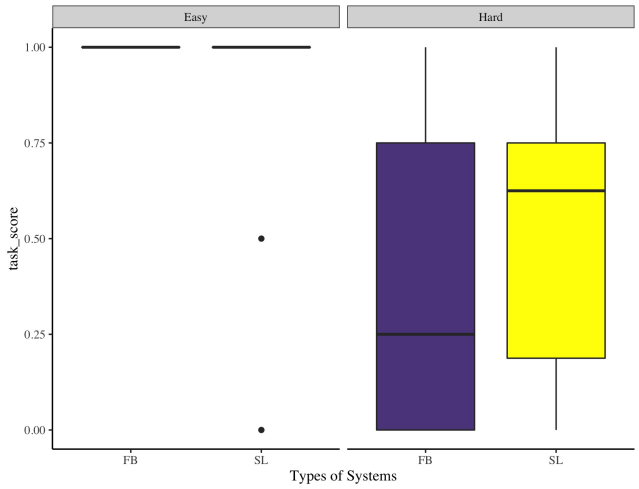

Figure 3: Boxplot of the types of systems and task score by search task difficulty

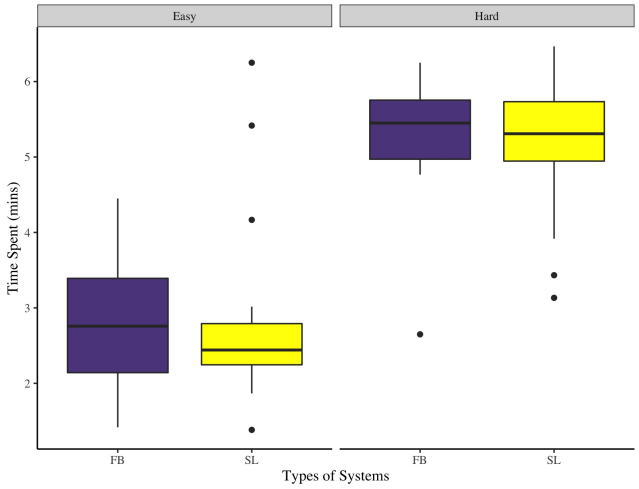

Figure 4: Boxplot of the types of systems and time spent by search task difficulty

system alone did not significantly affect the task score. The random effect of the user was present in the perceived usefulness of the system, whereas the random effect of the task appeared in the relevance of the system's responses to the topic and user satisfaction with the search process. The random effect of both user and task was present in the usefulness of the system's responses to find answers and user satisfaction with system responses.

Table 3 reveals that all the user perception measures made significant differences in the task score. The best model based on the Akaike information criterion (AIC) was Model 1 in which the user-perceived usefulness of the system accounted for 65% of the variances, followed by the relevance of the system's responses to the topic in Model 2, with 41% of the variances. Interestingly, the

Table 2: Model selection of fixed and random effects for user perception measures.

	Fixed and Random Effects Model
Model 1	sys_usefulness + (1 | user)
Model 2	topic_relevance + (1 | task)
Model 3	utility + (1 | task) + (1 | user)
Model 4	sys_satisfaction + (1 | task) + (1 | user)
Model 5	process_satisfaction + (1 | task)

Notes: sys_usefulness refers to how useful the system was in completing the task; topic_relevance is how relevant to the topic the system's responses were; utility refers to how useful the system's responses were to find answers; sys_satisfaction is how satisfied with the system's responses; process_satisfaction refers to how satisfied with the search process; random intercepts for task and user are specified with (1|task) and (1|user) respectively.

user's satisfaction with the system's responses represented the effect size of 21% in Model 4. Therefore, the results suggest that the system design should focus on the user-perceived usefulness of the design features (related to usability issues) and the relevance of the system's responses to the topic (related to effectiveness issues). User satisfaction may not be the best predictor of user search performance.

Therefore, our research hypothesis H1: Types of search systems and user perceptions will affect user search performance is partially supported. Specifically, the search system is not correlated with user search performance; user perception of the usefulness of the system in completing the search task and the system's responses to the relevance of the topic are good predictors of user search performance.

5.4 Effect of Perceived Difficulty and Perception on Task Score

Table 4 shows that the best model was perceived search task difficulty and its interactional effect with the relevance of the system's responses to the topic, which accounts for 52% of the variances. In other words, when a search task was considered difficult, participants had more problems judging the relevance of the system's responses to the topic. The user perceptions about the system utility and satisfaction about the system had significant effects on the task score, together with significant interactional effects. It is worth noting that both Tables 3 and Table 4 suggest that user perception of how useful the system was in completing the task was the best predictor of task score and there was no correlation between the user-perceived search task difficulty and the task score. Importantly, our constructed mixed-effects models have relatively large effect sizes, suggesting that participants in the study are very good at judging their performance.

Therefore, our research hypothesis H2: Perceived search task difficulty and user perceptions will affect the user search performance is supported. Specifically, perceived search task difficulty and its interactional effect with the relevance of the system's responses to the topic make a significant difference in user search performance.

Table 3: Effect of system and user perception on task score

	Task_Score				
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
sys_usefulness	0.25***				
	(0.02)				
topic_relevance		0.19***			
		(0.03)			
utility			0.13***		
			(0.03)		
sys_satisfaction				0.12***	
				(0.03)	
process_satisfaction					0.08***
					(0.03)
Constant	−0.20**	−0.01	0.27*	0.28*	0.44**
	(0.09)	(0.15)	(0.15)	(0.15)	(0.18)
N	64	64	64	64	64
Log Likelihood	−1.90	−2.20	−5.79	−6.92	−11.99
AIC (Akaike information criterion)	11.79	12.39	21.57	23.85	31.99
ICC (Intraclass correlation)	0.21	0.33	0.56	0.55	0.55
R^2 (fixed)	0.65	0.41	0.22	0.21	0.06
R^2 (total)	0.72	0.60	0.66	0.65	0.58

***p < .01; **p < .05; *p < .1

Table 4: Effect of perceived search task difficulty and user perception on task score

	Task Score		
	Model 1	**Model 2**	**Model 3**
perceived_difficulty	−0.33***	−0.33***	−0.35***
	(0.10)	(0.08)	(0.09)
topic_relevance	−0.11		
	(0.10)		
perceived_difficulty:topic_relevance	0.06***		
	(0.02)		
utility		−0.16**	
		(0.08)	
perceived_difficulty:utility		0.06***	
		(0.02)	
sys_satisfaction			−0.18**
			(0.09)
perceived_difficulty:sys_satisfaction			0.06***
			(0.02)
Constant	1.46***	1.66***	1.80***
	(0.46)	(0.37)	(0.41)
N	64	64	64
Log Likelihood	−2.16	−5.50	−6.71
AIC (Akaike information criterion)	16.31	23.01	25.42
ICC (Intraclass correlation)	0.30	0.36	0.34
R^2 (fixed)	0.52	0.41	0.41
R^2 (total)	0.66	0.63	0.61

***p < .01; **p < .05; *p < .1

6 DISCUSSION

This study is concerned with the design and evaluation of conversational search systems to support the pilot in cockpits, with particular references to the system evaluation issues from the user-centered perspectives. Our findings suggest that the system alone cannot predict search performance and search efficiency; participants in the study are very good at judging their performance. Specifically, their perceptions about the usefulness of the system in completing the task and the relevance of the system's responses to the topic are good predictors of search performance.

Our findings reveal that user satisfaction with the system's responses may not a good predictor of user search performance. Since the Alexa Prize Socialbot Grand Challenge is designed as research competitions to advance our understanding of human interactions with socialbots, to enhance the user experience, specifically user satisfaction when interacting with Alexa, it is not surprising that user satisfaction has been selected as the main evaluation criterion for success. The evaluation criteria which consist of automatic metrics from the system and human evaluation with Amazon's Mechanical Turk. Since the objective is to judge the system performance based on approximations of user satisfaction, it is found that there was a discrepancy between automatic metrics and human evaluation results [16]. Our findings suggest that if the goal is to enhance user search performance, the perceived usefulness of the system and the relevance of the system's responses to the topic are better predictors than user satisfaction with the system.

Our holistic approach to understanding user experience and user performance is intended to bridge the gap between system-centric evaluation (i.e., automatic metrics) and human evaluation. This approach is in line with the extrinsic evaluation that is concerned with how the use of system contributes to external outputs, such as task completion [40]. The user's judgment of usefulness has been proposed and used as an evaluation criterion for IIR (interactive information retrieval) studies [13, 47]. Our finding that user perceptions about the usefulness of the system in completing the task and the relevance of the system's responses to the topic are good predictors of search performance suggest user-perceived usefulness and relevance of the system's responses to the topic can be used for the evaluation of current conversational search systems. It demonstrates the applicability of the holistic approach adopted in previous information-seeking conversations [52, 54] to the design and evaluation of conversational search systems in a specific domain.

Given these findings, future research and development work needs to focus on the design of system support features for the relevance judgment, such as the snippets in search engine results page and system feedback. This work involves both the usability and effectiveness issues in system development. Future research on the correlations between the system-centric metrics and the user task score is suggested. Since the participants are homogeneous by age and experience in a specific domain, the generalizability of the research findings to other settings may be limited. Larger sample size would also enhance the validity of the results.

7 CONCLUSION

In this paper, we demonstrate a user-centered approach to the design and evaluation of conversational search user interfaces through a collaborative research project between academia and industry. It presents an approach for developing conversational search systems from the user perspectives by considering the user search behavior as well as individual differences when interacting with the proposed conversational search system. The controlled user experiment suggests that user perception of the usefulness of the system in completing the search task and the system's responses to the relevance of the topic are good predictors of user search performance. User satisfaction with the system's responses may not a good predictor of user search performance.

8 ACKNOWLEDGMENTS

This study was funded by Airbus Central Research & Technology with the support from the Aeronautical Computer Interaction Lab (ACHIL), from the Ecole Nationale de l'Aviation Civile (ENAC) and in particular the following researchers assistance for the experimentation: Alexandre DUCHEVET, Géraud GRANGER, Jean-Paul IMBERT, Nadine MATTON and Yves ROUILLARD.

REFERENCES

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the SIGIR '19*. ACM, New York, 475–484. https://doi.org/10.1145/3331184.3331265

[2] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational Search (Dagstuhl Seminar 19461). *Dagstuhl Reports* 9, 11 (2020), 34–83. https://doi.org/10.4230/DagRep.9.11.34

[3] Cecilia R Aragon and Marti A Hearst. 2005. Improving Aviation Safety with information Visualization : A Flight Simulation Study. In *Proceedings of the CHI '05*. ACM, New York, 441–450. https://doi.org/10.1145/1054972.1055033

[4] Alexandre Arnold, Gérard Dupont, Catherine Kobus, and François Lancelot. 2019. Conversational agent for aerospace question answering: A position paper. In *Proceedings of the 1st Workshop on Conversational Interaction Systems (WCIS at SIGIR)*. Paris.

[5] R. H. Baayen, D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 4 (2008), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

[6] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1 (2015), 51. https://doi.org/10.18637/jss.v067.i01

[7] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. (2017). arXiv:1712.05181

[8] Pia Borlund. 2016. A study of the use of simulated work task situations in interactive information retrieval evaluations: A meta-evaluation. *J. Doc.* 72, 3 (2016), 394–413. https://doi.org/10.1108/JD-06-2015-0068

[9] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What do you mean exactly? Analyzing clarification questions in CQA. In *Proceedings of the CHIIR '17* (Oslo, Norway). ACM, New York, 345—348. https://doi.org/10.1145/3020165.3022149

[10] Guido C Carim, Tarcisio A Saurin, Jop Havinga, Andrew Rae, Sidney W.A. Dekker, and Éder Henriqson. 2016. Using a procedure doesn't mean following it: A cognitive systems approach to how a cockpit manages emergencies. *Saf. Sci.* 89 (2016), 147–157. https://doi.org/10.1016/j.ssci.2016.06.008

[11] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. 2011. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the CIKM '11* (Glasgow, Scotland, UK). ACM, New York, 611–620. https://doi.org/10.1145/2063576.2063668

[12] Andy Cockburn, Carl Gutwin, Philippe Palanque, Yannick Deleris, Catherine Trask, Ashley Coveney, Marcus Yung, and Karon MacLean. 2017. Turbulent touch: Touchscreen input for cockpit flight displays. In *Proceedings of the CHI '17* (Denver, Colorado, USA). ACM, New York, 6742–6753. https://doi.org/10.1145/3025453.3025584

[13] Michael Cole, Jingjing Liu, Nicholas Belkin, Ralf Bierig, Jacek Gwizdka, C Liu, Jin Zhang, and X Zhang. 2009. Usefulness as the criterion for evaluation of

interactive information retrieval. In *Proceedings of the HCIR '09*. 1–4. http://cuaslis.org/hcir2009/HCIR2009.pdf

[14] Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2019. Survey on evaluation methods for dialogue systems. (2019). arXiv:1905.04071

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT 2019*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[16] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020. *The Second Conversational Intelligence Challenge (ConvAI2)*. Springer International Publishing, Cham, 187–208.

[17] Ralph H Earle, Mark A Rosso, and Kathryn E Alexander. 2015. User preferences of software documentation genres. In *Proceedings of the Annual International Conference on the Design of Communication (SIGDOC '15)*. ACM, New York, 46:1–46:10. https://doi.org/10.1145/2775441.2775457

[18] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *Int. J. Hum. Comput. Stud.* 132 (2019), 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

[19] Luanne Freund. 2013. A cross-domain analysis of task and genre effects on perceptions of usefulness. *Inf. Process. Manage.* 49, 5 (2013), 1108–1121. https://doi.org/10.1016/j.ipm.2012.08.007

[20] Ido Guy. 2018. The characteristics of voice search: Comparing spoken with typed-in mobile web search queries. *ACM Trans. Inf. Syst.* 36, 3 (2018), 30:1–30:28. https://doi.org/10.1145/3182163

[21] Marti A. Hearst. 2011. 'Natural' search user interfaces. *Commun. ACM* 54, 11 (2011), 60–67. https://doi.org/10.1145/2018396.2018414

[22] Morten Hertzum and Jesper Simonsen. 2019. How is professionals' information seeking shaped by workplace procedures? A study of healthcare clinicians. *Inf. Process. Manage.* 56, 3 (2019), 624–636. https://doi.org/10.1016/J.IPM.2019.01.001

[23] Kajta Hofmann, Bhaskar Mitra, Filip Radlinski, and Milad Shokouhi. 2014. An eye-tracking study of user interactions with query auto completion. In *Proceedings of the CIKM '14* (Shanghai, China). ACM, New York, 549—558. https://doi.org/10.1145/2661829.2661922

[24] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.* 3, 1–2 (2009), 1–224.

[25] Roger E. Kirk. 2013. *Experimental Design: Procedures for the Behavioral Sciences* (4th ed.). Brooks/Cole, Pacific Grove, CA.

[26] Alexandra Kuznetsova, Per Brockhoff, and Rune Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *J. Stat. Software, Artic.* 82, 13 (2017), 1–26. https://doi.org/10.18637/jss.v082.i13

[27] Catherine Letondal, Jean-Luc Vinot, Sylvain Pauchet, Caroline Boussiron, Stéphanie Rey, Valentin Becquet, and Claire Lavenir. 2018. Being in the sky: Framing tangible and embodied interaction for future airliner cockpits. In *Proceedings of the TEI '18*. ACM, New York, 656–666. https://doi.org/10.1145/3173225.3173229

[28] Yuelin Li and Nicholas J Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manage.* 44, 6 (2008), 1822–1837. https://doi.org/10.1016/j.ipm.2008.07.005

[29] Chang Liu, Ying-Hsang Liu, Tom Gedeon, Yu Zhao, Yiming Wei, Fan Yang, and Fan Zhangs. 2019. The effects of perceived chronic pressure and time constraint on information search behaviors and experience. *Inf. Process. Manage.* 56, 5 (2019), 1667–1679. https://doi.org/10.1016/j.ipm.2019.04.004

[30] Ying-Hsang Liu and Nicholas J Belkin. 2008. Query reformulation, search performance, and term suggestion devices in question-answering tasks. In *Proceedings of the IIiX '08*. ACM, New York, NY, 21–26. https://doi.org/10.1145/1414694.1414702

[31] Varvara Logacheva, Valentin Malykh, Aleksey Litinsky, and Mikhail Burtsev. 2020. ConvAI2 dataset of non-goal-oriented human-to-bot dialogues. In *The NeurIPS '18 Competition. The Springer Series on Challenges in Machine Learning*, Sergio Escalera and Ralf Herbrich (Eds.). Springer International Publishing, Cham, 277–294.

[32] Robert J Moore and Raphael Arar. 2019. *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework*. ACM, New York.

[33] Christine Murad and Cosmin Munteanu. 2019. "I Don'T Know What You'Re Talking About, HALexa": The case for voice user interface guidelines. In *Proceedings of the CUI '19*. ACM, New York, 9:1–9:3. https://doi.org/10.1145/3342775.3342795

[34] Chelsea M Myers. 2019. Adaptive suggestions to increase learnability for voice user interfaces. In *Proceedings of the IUI '19 Companion*. ACM, New York, 159–160. https://doi.org/10.1145/3308557.3308727

[35] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *Proceedings of the CHI '18* (Montreal QC, Canada). ACM, New York, 1—12. https://doi.org/10.1145/3173574.3174214

[36] Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. 2019. User intent prediction in information-seeking conversations. In *Proceedings of the CHIIR '19* (Glasgow, Scotland UK). ACM, New York, 25—33. https://doi.org/10.1145/3295750.3298924

[37] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the CHIIR '17* (Oslo, Norway). ACM, New York, 117–126. https://doi.org/10.1145/3020165.3020183

[38] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the Annual Meeting of the ACL*. Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

[39] Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.* 3, 4 (2009), 333–389. https://doi.org/10.1561/1500000019

[40] Anne Schneider, Ielka Van Der Sluis, and Saturnino Luz. 2010. Comparing intrinsic and extrinsic evaluation of MT output in a dialogue system.. In *Proceedings of the 7th International Workshop on Spoken Language Translation*. 329–336.

[41] Ben Steichen, Cristina Conati, and Giuseppe Carenini. 2014. Inferring visualization task properties, user performance, and user cognitive abilities from eye gaze data. *ACM Trans. Interact. Intell. Syst.* 4, 2 (2014), 1–29. https://doi.org/10.1145/2633043

[42] Muh-Chyun Tang, Ying-Hsang Liu, and Wan-Ching Wu. 2013. A study of the influence of task familiarity on user behaviors and performance with a MeSH term suggestion interface for PubMed bibliographic search. *Int. J. Med. Informatics* 82, 9 (2013), 832–843. https://doi.org/10.1016/j.ijmedinf.2013.04.005

[43] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the CHIIR '18* (New Brunswick, NJ, USA). ACM, New York, 42—51. https://doi.org/10.1145/3176349.3176388

[44] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the design of spoken conversational search. In *Proceedings of the CHIIR '18*. ACM, New York, 32–41. https://doi.org/10.1145/3176349.3176387

[45] Doug Turnbull and John Berryman. 2016. *Relevant Search: With Applications for Solr and Elasticsearch*. Manning Publications, Shelter Island, NY.

[46] Kelsey Urgo, Jaime Arguello, and Robert Capra. 2019. Anderson and Krathwohl's two-dimensional taxonomy applied to task creation and learning assessment. In *Proceedings of the ICTIR '19* (Santa Clara, CA, USA). ACM, New York, 117–124. https://doi.org/10.1145/3341981.3344226

[47] Pertti Vakkari, Michael Völske, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Modeling the usefulness of search results as measured by information use. *Inf. Process. Manage.* 56, 3 (2019), 879–894. https://doi.org/10.1016/j.ipm.2019.02.001

[48] Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A data-driven model of information-seeking dialogues. In *Advances in Information Retrieval. ECIR 2019. Lecture Notes in Computer Science, vol 11437*, Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra (Eds.). Springer International Publishing, Cham, 541–557.

[49] Terry L von Thaden. 2008. Distributed information behavior: A study of dynamic practice in a safety critical environment. *J. Am. Soc. Inf. Sci. Technol.* 59, 10 (2008), 1555–1569. https://doi.org/10.1002/asi.20842

[50] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L A Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the CHI EA '17*. ACM, New York, 2187–2193. https://doi.org/10.1145/3027063.3053175

[51] Peter Wittek, Ying-Hsang Liu, Sándor Darányi, Tom Gedeon, and Ik Soo Lim. 2016. Risk and ambiguity in information seeking: Eye gaze patterns reveal contextual behavior in dealing with uncertainty. *Front. Psychol.* 7 (2016), 1790. https://doi.org/10.3389/fpsyg.2016.01790

[52] Mei-Mei Wu. 2005. Understanding patrons' micro-level information seeking (MLIS) in information retrieval situations. *Inf. Process. Manage.* 41, 4 (2005), 929–947. https://doi.org/10.1016/j.ipm.2004.08.007

[53] Mei-Mei Wu and Ying-Hsang Liu. 2003. Intermediary's information seeking, inquiring minds, and elicitation styles. *J. Am. Soc. Inf. Sci. Technol.* 54, 12 (2003), 1117–1133. https://doi.org/10.1002/asi.10323

[54] Mei-Mei Wu and Ying-Hsang Liu. 2011. On Intermediaries' Inquiring Minds, Elicitation Styles, and User Satisfaction. *J. Am. Soc. Inf. Sci. Technol.* 62, 12 (2011), 2396–2403. https://doi.org/10.1002/asi.21644

[55] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. ACM, New York, 418–428. https://doi.org/10.1145/3366423.3380126