



HAL
open science

User Evaluation of Conversational Agents for Aerospace Domain

Ying-Hsang Liu, Alexandre Arnold, Gérard Dupont, Catherine Kobus, François Lancelot, Géraud Granger, Yves Rouillard, Alexandre Duchevet, Jean-Paul Imbert, Nadine Matton

► **To cite this version:**

Ying-Hsang Liu, Alexandre Arnold, Gérard Dupont, Catherine Kobus, François Lancelot, et al.. User Evaluation of Conversational Agents for Aerospace Domain. International Journal of Human-Computer Interaction, 2023, pp.1-20. 10.1080/10447318.2023.2239544 . hal-04575779

HAL Id: hal-04575779

<https://enac.hal.science/hal-04575779v1>

Submitted on 20 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



User Evaluation of Conversational Agents for Aerospace Domain

Ying-Hsang Liu, Alexandre Arnold, Gérard Dupont, Catherine Kobus, François Lancelot, Géraud Granger, Yves Rouillard, Alexandre Duchevet, Jean-Paul Imbert & Nadine Matton

To cite this article: Ying-Hsang Liu, Alexandre Arnold, Gérard Dupont, Catherine Kobus, François Lancelot, Géraud Granger, Yves Rouillard, Alexandre Duchevet, Jean-Paul Imbert & Nadine Matton (2024) User Evaluation of Conversational Agents for Aerospace Domain, International Journal of Human-Computer Interaction, 40:19, 5549-5568, DOI: [10.1080/10447318.2023.2239544](https://doi.org/10.1080/10447318.2023.2239544)

To link to this article: <https://doi.org/10.1080/10447318.2023.2239544>



© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC



Published online: 02 Aug 2023.



Submit your article to this journal [↗](#)



Article views: 1449



View related articles [↗](#)













View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)

User Evaluation of Conversational Agents for Aerospace Domain

Ying-Hsang Liu^{a,b} , Alexandre Arnold^c , Gérard Dupont^d , Catherine Kobus^c , François Lancelot^c ,
Géraud Granger^e , Yves Rouillard^e , Alexandre Duchevet^e , Jean-Paul Imbert^e , and Nadine Matton^{e,f} 

^aDepartment of ALM (Archival, Library, Information & Museum Studies), Uppsala University, Uppsala, Sweden; ^bPredictive Analytics, Chemnitz University of Technology, Chemnitz, Germany; ^cAIRBUS AI Research, Toulouse, France; ^dMavenoid, Stockholm, Sweden; ^eENAC, Ecole Nationale de l'Aviation Civile, Toulouse, France; ^fCLLE, University of Toulouse, Toulouse, France

ABSTRACT

The aerospace industry can benefit from conversational agents that provide efficient solutions for safety-of-life scenarios. This industry is characterized by products and systems that require years of engineering to achieve optimal performance within complex environments. With recent advances in retrieval and language models, conversational agents can be developed to enhance the system's question-answering capabilities. However, evaluating the added-value of such systems in the context of industrial applications, such as pilots in a cockpit, can be challenging. This article presents the design, implementation, and user evaluation of a conversational agent called Smart Librarian, which is tailored to the aerospace domain's specific requirements to support pilots in their tasks. Our results based on a controlled user experiment with flight school students indicate that the user's perception of the usefulness of the system in completing the search task is a good predictor of both task score and time spent. The system's responses to the relevance of the topic is also a good predictor of task score. The perceived difficulty of the search task and its interaction with the relevance of the system's responses to the topic also play a key role in search performance. The mixed-effects models constructed in this study had large effect sizes, demonstrating participants' ability to assess their performance accurately. Nevertheless, user satisfaction with the system's responses may not be a reliable predictor of user search performance. Implications for the design of conversational agents based on the domain's specific requirements are discussed.

KEYWORDS



Enterprise search;
conversational search
system; aerospace industry;
question answering; user
evaluation

1. Introduction

There has been a notable rise in the use of conversational search systems, both in personal and professional contexts in recent years. Popular examples of these systems include Apple Siri, Microsoft Cortana, Google Assistant, Amazon Alexa, and ChatGPT, which have all become a regular presence in our homes and on our mobile devices, thereby shaping our user experiences (UXs), such as the user's perceptions and responses about the mistakes made by chatbots (de Sá Siqueira et al., 2023) and the factors affecting the trust building of users (Rheu et al., 2021). Moreover, conversational search systems have been deployed in e-commerce websites and call centers to provide customer service support and enhance overall user satisfaction. Recently, there has been a shift toward evaluating conversational search systems, including chatbots, with a focus on improving UX and selecting user satisfaction as the primary criterion for success. For instance, research has shown that interaction signals with Cortana can be utilized to predict user satisfaction with search dialogues with a high degree of accuracy when interacting with an intelligent assistant (Kiseleva et al., 2016). Additionally, the Alexa

Prize Socialbot Grand Challenge¹ was created as a research competition aimed at advancing our understanding of human interactions with socialbots, with the support of large amounts of user data from Amazon.com. This evaluation approach has been developed from a system design perspective.

A recent survey of dialogue systems has highlighted various kinds of conversational systems associated with conversational agents, including task-oriented dialogue systems, conversational agents, and interactive question-answering systems (Deriu et al., 2021). Issues related to the voice-based user interface, such as recognition errors, UX, and voice queries, have become more prominent in the fields of human-computer interaction (HCI), UX, and information retrieval (IR). While systematic reviews of empirical user studies of conversational agents have revealed the importance of agent performance quality, such as knowledge level and task completion, for building trust in these artificial intelligence devices (Rheu et al., 2021), the efficacy and health outcomes of using these systems in the healthcare domain have rarely been evaluated (Laranjo et al., 2018). From a user or human-centered AI perspective, the usefulness of these

CONTACT Ying-Hsang Liu  ying-hsang.liu@abm.uu.se  Department of ALM (Archival, Library, Information & Museum Studies), Uppsala University, Uppsala, Sweden; Predictive Analytics, Chemnitz University of Technology, Chemnitz, Germany

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

systems in supporting user tasks has not been well-established (C. Liu et al., 2021; Shneiderman, 2022).

One of the main goals of conversational search systems is to provide information services in a natural and interactive style, akin to human–human interactions in information-seeking conversations. Ideally, user interfaces for conversational search systems should allow for clarifying questions and a smooth flow of conversation (Hearst, 2011). This research area has garnered significant attention from various fields, such as Natural Language Processing (NLP), IR, and HCI (Anand et al., 2020; Logacheva et al., 2020; Myers, 2019; Zamani et al., 2020).

The aerospace industry heavily depends on extensive collections of documents that include system descriptions, manuals, and procedures. Many of these documents are subject to specific regulations and must be utilized in the context of safety-of-life scenarios, such as aircraft procedures in the cockpit. However, pilots searching for particular information in this vast corpus often spend a considerable amount of time navigating through the documents. Even experienced pilots who are familiar with the structure of the documents can struggle to find known items within a limited timeframe.

Having a structured information organization enables quick and accurate retrieval of specific information. In cases where the structured information is not sufficient, search systems can be employed; however, these systems have their limitations. Typically, it is up to the user to adapt their search needs by using specific keywords and/or syntax, which is known as the difficulty of articulating information needs (Y.-H. Liu & Belkin, 2008; Wittek et al., 2016). While simple queries with readily available answers may not pose a significant challenge, understanding complex procedures or troubleshooting system errors can result in multiple queries and a cumbersome UX.

Our study focuses on the development of a “Smart Librarian” system that combines a task-oriented dialogue system with a conversational agent, incorporating both voice-based and non-voice-based interfaces, as well as an interactive question-answering component. The assistant is designed as a task-oriented system for a specific domain, featuring mixed system and user initiatives and a multi-modal interface to enhance situation awareness in a cockpit. Our objective is to evaluate the effectiveness of smart and conversational search for cockpit documentation.

This article outlines the implementation and evaluation of a conversational agent, with a specific focus on adapting state-of-the-art technologies for aerospace domain specificity. The design and evaluation of the conversational search user interfaces (SUIs), called Smart Librarian (SL), are centered around a user-centered approach to support pilots in cockpits. The assistant’s abilities are intended to align with the requirements of conversational search systems to aid pilots in their tasks. Our results indicate significant interaction effects between task difficulty and system types, with the SL system performing better for more challenging search tasks. We suggest potential research and development directions for conversational agents in the future.

This article represents an extension of our previous study (Y. H. Liu et al., 2020), where we have broadened our

examination of relevant literature to include information-seeking behavior and the evaluation of conversational search systems. To further augment our research, we have formulated and examined an additional research hypothesis, which asserts a correlation between search task difficulty and user perception measures, using logarithmic cross-ratio analysis. Our extended findings include the relationship between search task difficulty and user perception measures, the impact of system and user perception on time spent, as well as the influence of perceived difficulty and user perception on time spent. To accommodate these supplementary outcomes, our discussion has been expanded accordingly.

2. Related work

2.1. Information-seeking behavior

The investigation of human information behavior concentrates on users’ requirements for information, the process of seeking information, and the utilization of information across diverse situations (Case & Given, 2016). It constitutes a significant part of information studies and highlights the approaches that people utilize to search for information and its application in tasks such as decision-making, problem-solving, and sense-making. To comprehend the UX in work environments, it is essential to have an accurate perception of the information atmosphere and the techniques individuals utilize to find information to accomplish their tasks.

Research on information behavior has shed light on how individuals in safety-critical environments make sense of information. One notable contribution in this area is the distributed information behavior system introduced by von Thaden (2008), which identifies three dimensions of information behavior: information need, information seeking, and information use, spanning the exploration–exploitation continuum. In a flight simulation study with 19 pilot training students, video-recorded transcripts were analyzed to understand the social information practice between high- and low-performance crews. One significant finding was 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” (von Thaden, 2008, p. 1567). The finding emphasizes the significance of procedures in information-seeking behavior within a safety-critical environment like that of pilots in cockpits. Hertzum and Simonsen (2019) demonstrate that procedures in the workplace have an impact on information-seeking behavior. More specifically, the triage and timeout procedures implemented in emergency departments for normal, abnormal, and emergency situations trigger different tasks. The study also uncovers differences in the information-seeking strategies used by experts and novices.

The correlation between tasks and information resources has been studied, and it has been found that the type of task, such as learning, fact-finding, doing, decision-making, and problem-solving, and the genre of document, such as best practices, frequently asked questions (FAQs), product documentation, and whitepapers, as well as their

interactions, influence the perceived usefulness of documents in the workplace (Freund, 2013). Earle et al. (2015) have demonstrated that work roles and experience also impact the use of software documentation. These findings suggest that individuals' levels of domain expertise and experience, work roles, tasks, and procedures all play a crucial role in determining their information-seeking strategies and the perceived usefulness of information resources.

2.2. Aviation cockpits and controls

Research into the aviation cockpit and controls environment has centered on cognitive strategies and processing in high-stress situations for the purpose of designing automated support systems. The role of cognitive processes inherent in tasks and the specific cognitive strategies employed by pilots in designing automated support systems, such as automated cockpits, has been a key area of focus (Carim et al., 2016). A study of the use of procedures, such as the Quick Reference Handbook (QRH), in emergency situations in the cockpit found that pilots used a range of resources, frequently consulting fragments of the QRH checklists rather than following them in sequence (Carim et al., 2016).

Endsley's model of situation awareness has been widely tested and applied to the design of systems in critical safety environments (Endsley, 2018). Situation awareness, which is defined as "a current understanding of the state of the world and the systems being operated, is a critical foundation for expertise in many complex domains, including driving, aviation, military operations, and medical practice" (Endsley, 2018, p. 714). There are three levels of situation awareness, namely perception, comprehension, and projection, with research indicating that approximately 20% of pilot errors occur at the comprehension level. Recent research on professional pilots' situation awareness reveals that individual differences predict situation awareness (Cak et al., 2020). Specifically, the lack of correlation between offline and online measures of situation awareness suggests that both measures should be used simultaneously to obtain a comprehensive evaluation of situation awareness. In the context of predicting offline situation awareness scores, working memory and level of expertise exhibited the strongest predictive capacity. In contrast, expertise, divided attention, and inhibition emerged as the most influential predictors for online situation awareness scores (Cak et al., 2020). Furthermore, a systematic review and meta-analysis of direct objective measures of situation awareness found that the Situation Awareness Global Assessment Technique (SAGAT) is a reliable measure of situation awareness that can be applied across many domains (Endsley, 2021).

An observational and interview study of cockpit activities for tangible design from the perspective of HCI found that pilots expressed the importance of having tools separated from the aircraft systems. The physical QRH, which can be held in the hands, was especially valued by pilots in degraded contexts where control is no longer available (Letondal et al., 2018). In addition, an evaluation of the usability of an information visualization system in flight to

enhance aviation safety was conducted in a flight simulator setting (Aragon & Hearst, 2005).

The use of the electronic flight bag (e-flight bag) system has replaced traditional paper-based documentation in aviation. In order to improve current flight bags, SUIs for within-document searches have been developed. Research has focused on designing user interfaces that support access to segments of full-text documents, specifically in the context of book selection (Wacholder, 2008), within-document retrieval (Harper et al., 2004), and focused retrieval (Arvola et al., 2012). One proposed interface, ProfileSkim, uses an interactive bar graph to retrieve relevant segments of a document. A user study with manual indexing tasks found that ProfileSkim was more efficient than other interfaces with a "Find" command in web browsers, and was at least as effective as the "Find" command tool (Harper et al., 2004). Another proposed interface, the Focus + Context user interface, an extension of TitleBars for within-document search and navigation (Schwartz et al., 2010), showed that users did not have usability issues using the term-distribution visualization system. Overall, these studies suggest that interfaces with visualizations of term distributions in long documents can be efficient in supporting user access to portions of the document, which is particularly important in aviation for efficient and timely information access.

From the perspectives of HCI, interfaces displaying spatially stable overviews of the whole document have been identified as efficient for navigating documents (Gutwin et al., 2017). Field experiments have shown that the overview is useful for both pattern matching and revisiting pages, particularly for locating previously visited pages. To develop a spatial memory of long documents, interfaces with augmented scrollbars that use visual items as landmarks for revisiting long documents have been proposed (Mollashahi et al., 2018). The findings suggest that double-icons, which are two-level augmented scrollbars that use icons, result in improved search performance and user preferences.

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 for specific tasks, within-document retrieval, and usability issues are emphasized.

2.3. Information seeking conversation

Interactive information retrieval (IIR) study has made an effort to define the goals and communicative functions of elicitation (i.e., questions to request information) in information seeking conversations. This research has been informed by theories of human-human communication and linguistics. According to research by Wu (2005), user elicitation behavior was found to be influenced by individual differences such as rank, age, and experience as well as by situational factors such as the length of the interaction and the number of utterances. With specific reference to user satisfaction, subsequent studies have developed the idea of elicitation styles, which are defined by linguistic forms,

utterance purposes, and communicative functions (Wu & Liu, 2003, 2011). However, these findings have not been directly applied to the design of conversational search systems.

Recent research has placed emphasis on creating recommendations for system design based on user studies. For instance, in one study, researchers analyzed human-to-human interactions in a laboratory environment and compared them to established search models to guide the development of spoken conversational search systems (Trippas et al., 2018). Additionally, there has been an exploration of the system requirements for intelligent conversational assistants, with a focus on improving the overall UX (Vtyurina et al., 2017).

Recent studies have adopted the idea of computers as social actors and used interpersonal communication theories to develop a taxonomy of social cues for conversational agents (Feine et al., 2019). Ethnomethodology and conversational analysis have been employed to explore the role of conversational interfaces in daily life, particularly in relation to designing request and response structures for embedded social interactions (Porcheron et al., 2018). In addition, the Expectation-Confirmation Theory has been used to propose a PEACE (Politeness, Entertainment, Attentive Curiosity, and Empathy) model for conversational chatbots, emphasizing the importance of adhering to politeness norms (Svikhnushina & Pu, 2022). This research thread, along with IR research, contributes to ongoing discussions on borrowing and re-conceptualizing theories from other fields for the design of conversational search systems.

2.4. 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” (Radlinski & Craswell, 2017, p. 160). An evaluation framework for conversational agents in the aerospace sector was introduced, with the aim of identifying conversational styles for constructing computational models for speech-based agents (Arnold et al., 2019). Additionally, research has been conducted to identify the conversational styles for building computational models at scale for speech-based conversational agents (Thomas et al., 2018). Researchers proposed a natural conversation framework for a conversational agent that can engage in natural conversations, which is based on the conversation analysis (CA) from sociology (Moore & Arar, 2019). The system was designed with a conversational-centric interaction style in mind, with the goal of natural understanding and conversational competence. It consists of 15 types of conversational UX patterns for (1) conversational activities, (2) sequence-level management, and (3) conversation-level management. The technical discussions are based on IBM’S Watson Assistant. These studies suggest the existing methods used to explore the conversational search systems from the perspectives of conversational search system design.

Technical research on conversational search systems has focused on identifying user intent during information-seeking conversations, designing user interfaces for various interaction modes, and providing clarification questions. Neural classifiers are used to identify user intent, with the position of an utterance in a dialogue being the most significant structural feature (Qu et al., 2019). The formulation of tasks as noun phrase ranking problems is used to generate clarification questions from community question-answering websites (Braslavski et al., 2017). Neural models are used to generate clarification questions by taking into account the sequences of interaction purposes (Aliannejadi et al., 2019). A formal model of information-seeking dialogues that includes the query, request, feedback, and answer has been proposed to identify the frequency of sequence patterns (Vakulenko et al., 2019). These studies demonstrate the technical aspects that have been examined regarding conversational search systems, focusing on identifying user intent, designing user interfaces, and providing clarification questions.

Overall, although conversational search systems have received significant research and practical attention, their usefulness has not been thoroughly evaluated from the user’s perspective during the system design process.

2.5. Evaluation of conversational search system

The multidisciplinary nature of conversational search system research has led to varying approaches in their evaluation. In the field of NLP, one of the primary evaluation approaches distinguishes between intrinsic and extrinsic evaluations of machine outputs (Schneider et al., 2010). Intrinsic evaluation assesses the system’s internal outputs, while extrinsic evaluation examines how the system’s use contributes to external outcomes, such as task completion.

Evaluation approaches for conversational search systems vary across different communities. In the IR community, the focus has been on creating test collections to compare system performance, using appropriate evaluation metrics for different types of question-answering tasks. This is an intrinsic approach that evaluates the internal outputs of the system. One example is the Text REtrieval Conference (TREC) Question Answering (QA) track (Voorhees, 2008), which aimed to investigate appropriate evaluation methodologies for question-answering systems through the development of a test collection and assessing fact-based short answer questions. Assessors evaluated the correctness of answers. Another example is the complex, interactive question answering (ciQA) task (Dang et al., 2006), which involved factoid and list questions in interactive systems and moved away from factoid questions and one-shot interactions. The evaluation metrics for ciQA included accuracy and average *F*-score. Recently, the TREC Complex Answer Retrieval (Dietz et al., 2018) focused on answering questions that require multiple text segments, passage task, and entity task, and the evaluation metrics included R*P*rec (Precision at *R*, where *R* is the number of relevant documents for a given query), MAP (Mean Average Precision), and NDCG (Normalized Discounted

Cumulative Gain) to consider the effectiveness of a ranking search system.

When evaluating interactive IR systems, such as conversational search systems, from a user perspective, the approach taken is extrinsic and holistic, with a focus on identifying appropriate measures for successful task completion. For example, an empirical study (Su, 1992) aimed to identify the best evaluation measure for interactive IR performance by analyzing 40 users with genuine information needs interacting with intermediaries and conducting post-search interviews. The results indicated that the value of search results is the most effective measure of search success. Other critical factors that influence measures of search success include interaction and effectiveness. Evaluation methodologies have been developed to assess the usefulness of proposed interactive, analytical question-answering systems that support intelligence analysts in writing up reports. These methodologies measure the system's effectiveness by the quality of reports produced (Kelly et al., 2007; Small & Strzalkowski, 2009; Sun et al., 2011; Wacholder et al., 2007).

In order to evaluate conversational search systems, several factors need to be taken into consideration, including the domain of application (e.g., open or domain-specific), the specific tasks that the system is intended to support (e.g., goal-oriented or non-goal oriented dialog, complex answer retrieval), and the alignment between the task and the metrics used to evaluate the system. The intrinsic approach provides a way to compare system performance in a controlled environment, while the extrinsic approach can demonstrate how the use of a conversational search system contributes to task completion. By considering these factors and selecting appropriate evaluation methodologies, we can gain a better understanding of the strengths and limitations of conversational search systems, and make improvements to enhance their usefulness and effectiveness.

Our study adopts a comprehensive method to investigate both UX and performance when engaging with a prototype conversational agent. Our objective is to bridge the divide between evaluations centered around system-generated metrics and those reliant on human input, as seen in the crowdsourcing platforms used in the Alexa Prize Socialbot Grand Challenge (Dinan et al., 2020).

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

Studies on user information seeking have indicated that various factors, such as domain expertise, work roles, tasks,

and procedures, can affect users' strategies for seeking information and their perceived usefulness of information resources (Earle et al., 2015; Freund, 2013; Hertzum & Simonsen, 2019; Li & Belkin, 2008). Information behavior research has shown that in safety-critical environments, overly conditioned information behaviors can limit methodical information behaviors, leading to a risk of missing crucial steps in the process of projecting the future state of the aircraft and planning ahead (von Thaden, 2008). Additionally, professional pilots' situation awareness can be predicted by individual differences in their level of domain-specific expertise (Cak et al., 2020).

Studies on user behavior have indicated that domain knowledge of users has an impact on their perception of search interfaces and the features of a system, particularly in situations where the search environment is uncertain. User perception in this context refers to their understanding of the search interactions and their interpretation of the search results. For instance, domain experts have found the suggestion feature useful for unfamiliar search tasks (Tang et al., 2013), whereas medical practitioners have been able to find highly relevant documents using semantic components to structure their queries (Lykke et al., 2012). Furthermore, topical knowledge and credibility perception in web searches have been shown to be correlated (Lee & Pang, 2018). These findings suggest that the perceived usefulness of system features is affected by the user's domain knowledge, as well as their familiarity with the search tasks. However, the relationship between domain knowledge and user perception measures in specific domains requires further investigation.

Therefore, within the context of the aerospace domain, our proposed research hypotheses are as follows:

- H1. Search task difficulty is correlated with user perception measures.
- H2. Types of search systems and user perceptions affect user search performance.
- H3. Perceived search task difficulty and user perceptions affect user search performance.

3.3. Research design

In this study, our main focus is on designing and evaluating conversational search systems from a user's perspective. To achieve this, we have decided to adopt a laboratory setting to observe and analyze user interactions with a prototype system. This approach has been selected because it allows us to establish a causal relationship between variables in a controlled environment, and we can use the findings to inform specific design decisions. However, it is important to note that this approach is resource-intensive, time-consuming, and requires a diverse range of expertise (C. Liu et al., 2021). Additionally, the results obtained from laboratory studies may be subject to individual variability, as highlighted by previous research studies (e.g., Steichen et al., 2014; Tang et al., 2013; Wittek et al., 2016). Some examples of these studies include experiments involving pilots co-designing a system in a flight simulator (Aragon & Hearst, 2005) and

touch design experiments with students in a turbulent environment (Cockburn et al., 2017).

The experiment protocol has been approved by the Toulouse University Research Ethics Committee (IRB00011 835-2020-29-03-208). Each 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 (BIGONE – A320/A330 cockpit simulator) within the ACHIL (Aeronautical Computer Interaction Lab) platform of the ENAC (Ecole Nationale de l'Aviation Civile) Research Lab. The setting was intended to create an environment that can elicit the information needs of participants, as suggested in simulated work task situations (Borlund, 2016).

3.5. Search task

We have taken into account the complexity of search tasks by categorizing them as easy or complex, using the search as learning approach in our design process (Urgo et al., 2019). To assess the perceived complexity of the search tasks after using the system, we administered a questionnaire to the users (Li & Belkin, 2008). Specifically, the easy task involves fact-finding (see Figure 1) 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. Besides, there is a need

to reformulate the problem with new words/questions and at least two document units are necessary to find the solution to the problem. In other words, 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 (Flight Crew Operating Manual) document. The very narrow specificity of the aeronautical domain and the particular form of documents, allowed 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

Subjects in the study were presented with tasks using a traditional Graeco-Latin square design (Kelly, 2007). To minimize the effect of presentation order of treatments (Kirk, 2013), the study employed a 2×2 design that included two types of search systems (i.e., e-flight bag and SL) and two types of search tasks (easy and hard). All participants interacted with both systems and both search tasks using a tablet similar to the one used by professional pilots in the cockpit.

3.7. Metrics

3.7.1. Search performance

The tasks defined are pure goal-oriented search task: the user was given five minutes to find the exact answer and locates the procedure used. Classic precision and recall metrics used in IIR 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 TOP1 precision – 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 answer to the situation in the procedure, finding

(8) 2 - SL-A - Cockpit windshield cracked

You are on cruise at FL370.

You notice a crack on the cockpit windshield on the cockpit side. You have to look for the procedure to follow and what is the MAX FL to use.

⋮

Find the appropriate document - Please provide the title here.

Long answer text

What is the maximum FL (flight level) to use?

Short answer text

Figure 1. Example of a simple task.

Table 1. Difficulty level of search tasks, with description and key aspects (FL means “Flight Level”).

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

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

the next procedure, etc.). Since hard tasks had more steps, the task score was the average of scores in [0;1] for each step (1 being the maximum). 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/recall of the system measures 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 one data point per user per search task.

3.7.2. User’s perception of the problem and system

The study collected additional metrics by administering post-search questionnaires after each task and a final exit questionnaire. A five-point Likert scale was used to collect user perceptions of task difficulty, familiarity, system relevance, and usefulness. Each participant completed the questionnaires after each task and system usage.

3.8. Prototype system

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

- A dialog engine (based on RASA platform (Bocklisch et al., 2017) handling the conversation and identifying user’s intents;
- A search engine (based on Solr (Turnbull & Berryman, 2016)) where the documents collection is indexed following the BM25F relevance framework (Robertson & Zaragoza, 2009);
- A QA engine, based on a BERT large model (Devlin et al., 2019), fine-tuned using the FARM framework.² 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 (Rajpurkar et al., 2018); the other is a classification task (i.e., whether the answer to the question is contained or not in the document extract).

3.8.1. The dialog engine

We used the open-source RASA platform for the implementation of the dialog engine (Bocklisch et al., 2017), comprising:

- *Natural language understanding*: Recognizing high-level intent/entities from raw user utterances (e.g. “greeting”, “positive/negative feedback” or “question”);

- *Dialog policy*: Predicting the next best action (an utterance or a custom action) based on current dialog state, including last recognized intent.

Both components above were trained with machine learning pipelines provided in RASA, based on natural language and story examples: the former maps user utterances to predefined intents/entities, the latter gives typical dialog scenarios to learn and generalize from (to avoid building manually a conversation state machine).

The core “skill” of our dialog engine focuses on recognizing any generic question from the user, mapping it to the “question” intent, and predicting the trigger of a custom action (written in Python), which calls the retriever and QA systems described in next sections to provide an answer. Examples of natural language questions were built by combining open QA dataset questions with in-house examples more related to our pilots’ documentation context. Taking inspiration from the real dialogs in the technical problem-solving domain from MSDialog dataset (Qu et al., 2018), we also integrated positive/negative feedback intents to be able to handle user reactions after providing an answer: in case of negative feedback, a custom action is triggered to propose the best answer from the document ranked just below the one currently suggested.

A chitchat “skill” (i.e., another sub-part of our conversational system) was added to the core one, containing more than 50 typical small talk intents and responses to make the dialog appear more human-like: “greeting”, “goodbye”, “thanking”, etc. Some chitchat user utterances might be in question form but are usually learned not to be confused with the generic “question” intent mentioned above with enough training data.

3.8.2. The retriever

This component consists of an IR system, which follows the classic architecture of recent neural QA systems. It allows to filter the overall document collection (1) to exclude non-relevant documents and (2) reduce the considered set of documents to a size that is compatible with the foreseen response time.

First, the document collection has been extracted from its original XML format which includes simultaneously semantic and presentation tagging. The isolation of each procedure has been done at this level as well as the extraction of metadata such as a unique identifier, applicability scope, and classification in the hierarchical ATA chapters.³ This allowed defining the minimal granularity of the collection. The text content has then been pre-processed to eliminate

inconsistent formatting issues and improve the quality of the terms indexed such as resolving abbreviations meanings, adapting the numerical representation of units, and flattening table content to ensure headers and legends are correctly indexed.

We compared different indexing schemes (including the classic tf/idf) and the BM25F from (Robertson & Zaragoza, 2009) allowed to offer the best performances.

3.8.3. The QA system

3.8.3.1. BERT fine-tuning approach. The QA module in the pipeline is an extractive QA approach, where the model extracts a span of text from a document to answer a natural language question. The QA model is obtained by fine-tuning a BERT language model (Devlin et al., 2019) for an extractive QA task on the SQuAD 2.0 dataset (Rajpurkar et al., 2018).

Almost all QA datasets are made of long documents that cannot fit in a standard transformer model. Hence, for a given question and a document to consider, the document is first decomposed into n smaller passages; those passages are provided, with the question to answer, as input to the QA engine; this step results in n predictions, that need to be aggregated to get a final answer for the given question/document pair. The different steps are summarized in Figure 2.

The input in Figure 3 provided to the QA model contains tokens from both the question and the passage, some special tokens, and eventually some padding tokens (that enables to reach the model's maximum sequence length if needed).

Formally, we define a training set instance as a triple (c, s, e) , where c is a context of a given size ($\max_{seq_len} \in \{384, 512\}$ in this study) of wordpiece ids, corresponding to the question, to the passage considered, to some special tokens and eventually some padding. $(s, e) \in [0, \max_{seq_length}]^2$ respectively refer to the start and end of the target answer span when the answer span is contained in the passage; they are both equal to 0 otherwise.

During inference, the predictions for each passage need to aggregate to extract, from the document, the answer (or not) to the question. If the best prediction for each passage is no answer, then the final prediction is no answer. Otherwise, the final prediction is built by picking up the answer span kind prediction with the highest score.

The FARM framework⁴ was used to fine-tune the BERT models on the QA task. More details on the approach can be found in this blog post.⁵

3.8.3.2. Multi-task approach. Multi-task learning (MTL) aims at boosting the overall performance of each task by leveraging useful information contained in multiple related tasks. It has shown great success in NLP. The main idea of MTL is to leverage useful information contained in multiple related tasks to improve the generalization performance of all the tasks (Zhang et al., 2018). Multi-task learning has been successfully used in many applications from machine learning, from NLP (Collobert & Weston, 2008) and speech recognition (Deng et al., 2013) to computer vision (Girshick, 2015), etc.

With the MTL approach, a training set instance becomes a four-tuple (c, s, e, t) where c is a context of a given size ($\max_{seq_len} \in \{384, 512\}$ in this study) of wordpiece ids, corresponding to the question, to the passage considered, to some special tokens and eventually some padding. $(s, e) \in [0, \max_{seq_length}]^2$ respectively refer to the start and end of the target answer span. t is the tag associated to the sample with two possible values: $t = \text{SPAN}$ if the answer to the question is in the passage considered, $t = \text{NO_SPAN}$ otherwise.

Adding this classification head to the network, that tries to predict if it is able to answer a question given the considered passage, may help the overall performance of the QA engine.

During training, the losses of both tasks, QA on one hand and classification, on the other hand, are summed. During inference, as for the question-answering task, for a question and a given document, all results from all the samples of question/passage have to be aggregated to get at the end a unique classification: "SPAN" if the answer is in the document, "NO_SPAN". For this step, we went for a basic approach, which consists of taking the classification tag from the passage with the highest score for the classification head.

3.8.4. Application integration

Additional capabilities to process speech inputs and produce speech outputs have been integrated as an alternative to the traditional textual input. Figure 4 offers an overview of the whole architecture.

The whole system was made available through a reactive web interface enabling conversation and document exploration

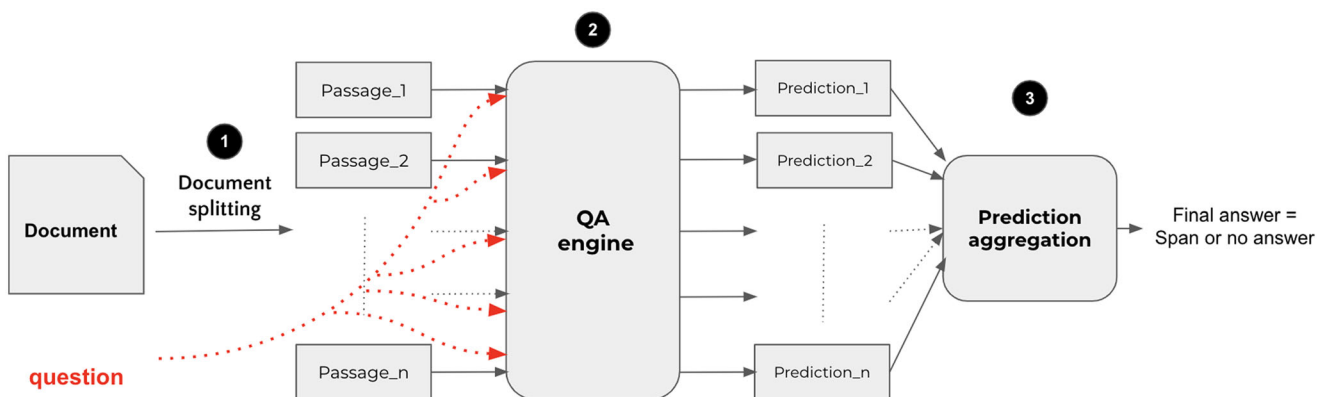


Figure 2. QA engine pipeline.

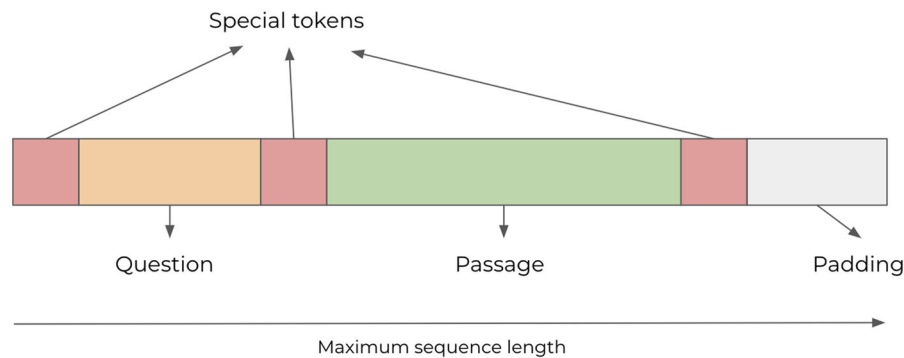


Figure 3. Input to the QA model (with a size equal to the maximum sequence length).

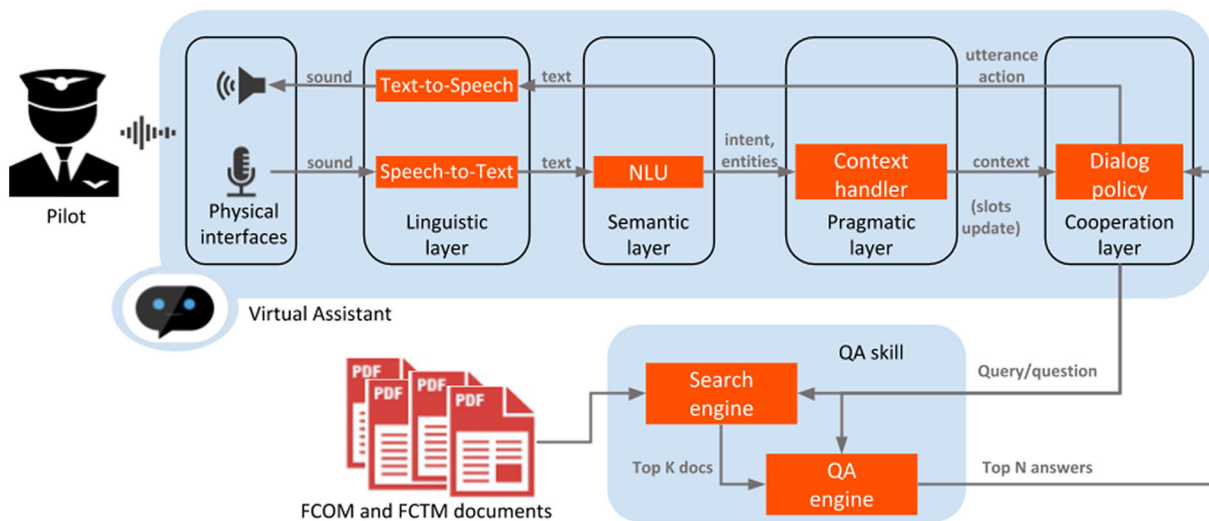


Figure 4. Overview of the prototype architecture.

(see Figure 5). It was deployed in a cloud environment and made accessible to users through a tablet.

The reference system (e-flight bag) was one of the current software used by many pilots in commercial flights. This application is distributed on tablets and customized for each aircraft. Only the library features (for access to documents and procedures) were used in the experiment. Neither the baseline system and the prototype system had the speech recognition feature 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. Even if these students did not have any airline experience yet, they have a good understanding of aircraft technologies and flying physics. At the time of the experiments, they did not have a particular course on a specific aircraft (such as A320).

3.10. Evaluation protocols

We have developed realistic scenarios for user experiments 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 (Borlund, 2016). In designing search tasks, we have considered the complexity of tasks from the perspectives of search as learning (Urgo et al., 2019). Several questionnaires, including a demographic questionnaire (Appendix A), post-search (Appendix B), and exit questionnaire (Appendix C), have been developed to assess the user perceptions during the search process as well as overall perceptions about the whole interaction 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

To examine whether there is a significant relationship between user perception measures and search task difficulty, we employed a logarithmic cross-ratio analysis technique (Fleiss et al., 2003). We chose this technique due to its ability to resist sample selection bias and its successful use in analyzing the relationship between individual differences and user search and gaze behavior in previous research (e.g., Saracevic et al., 1988; Wittek et al., 2016).

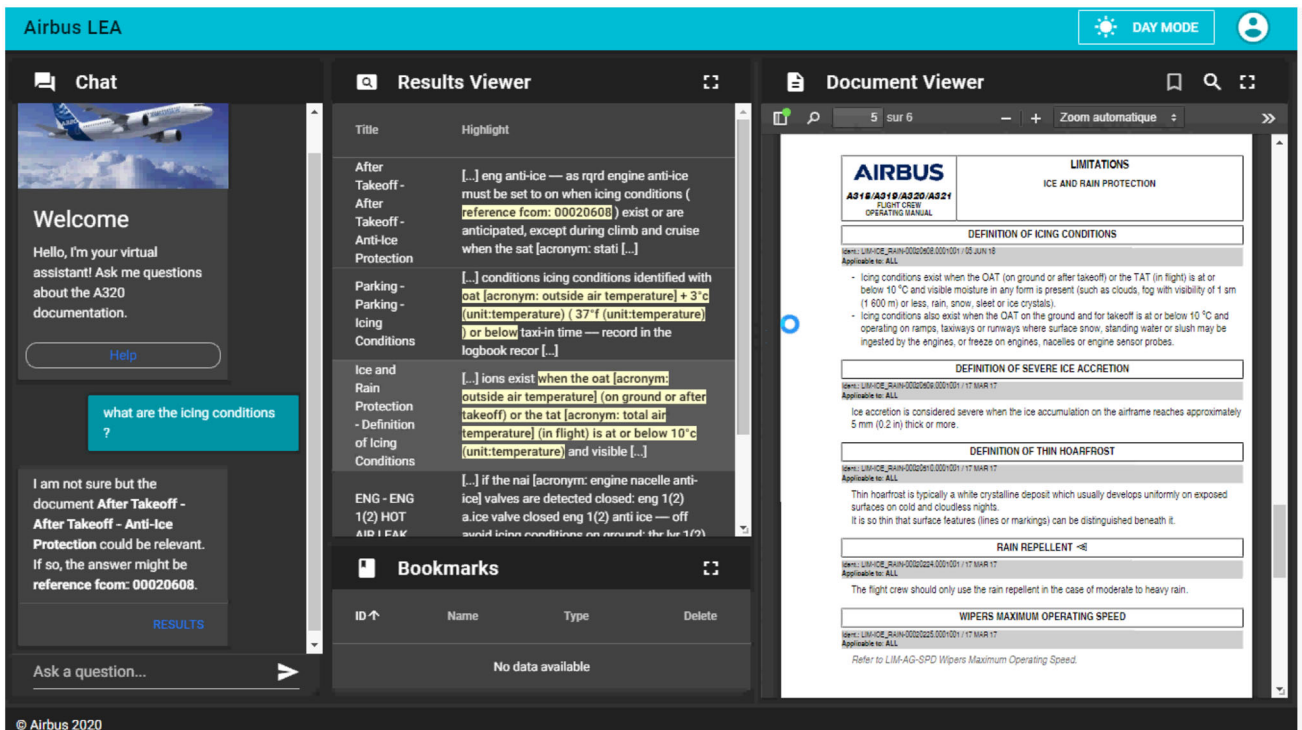


Figure 5. Screenshot of the prototype showing conversation on the left, search result panel in the center, and document view on the right.

To analyze the effects of system and user perceptions on search performance, we employ mixed-effects models that take into account both fixed effects, which are related to experimental conditions, and random effects, which account for individual differences in the sample. Mixed-effects models are particularly useful in examining the impact of random effects of subjects and search tasks, and have been applied in previous IR research to model search topics, gaze behavior, and user characteristics (Baayen et al., 2008; Carterette et al., 2011; Hofmann et al., 2014; C. Liu et al., 2019). We use the lme4 package in R statistical computing software to fit our models (Bates et al., 2015).

We did not find any significant relationship between the order of the search tasks and the time spent on them ($R = 0.012, p = 0.92$). On average, participants spent around four minutes on each search task, with a standard deviation of 1.55 minutes, and within the seven-minute time limit ($M = 4.04, SD = 1.55$). Our full model construction and data fitting considered both fixed effects of system and user perception, as well as the random effects of search task and user. We used an automatic backward model selection to fit the data to a linear mixed model (Kuznetsova et al., 2017), and found that the random intercepts for both search task and user were significant for time spent, with $p < .001$ and $p < .01$, respectively. Thus, we selected a mixed-effects model that controlled for search task and user as random effects. We also conducted diagnostic checks to assess the normality of residuals, outliers, distribution of random effects, and heteroscedasticity. We found that the random intercepts for search task were significant for task score, with ($p < .001$), and thus selected a mixed-effects model with search task as a random effect.

5. Results

5.1. Participant characteristics

The study included 16 pilot school students, the majority of whom were aged between 18 and 25 years. Nearly, all students had flying experience as an amateur or student pilot, with an average of less than 70 flight hours, while three students had general aviation experience for over 5 years. None of the participants had commercial flying experience. Most of the students reported using search engines every day or several times a day, but had limited experience using virtual assistants. Demographic variables were not found to be correlated with search performance or user perception measures. Therefore, the participants were considered homogeneous with respect to demographic variables.

5.2. Search task difficulty and user perception measures

Table 2 shows that user perception measures of the systems were all negatively correlated with search task difficulty, except for the user's familiarity with the search task.

To clarify the results, when the search tasks were perceived as difficult, the systems were less likely to be useful for completing the tasks. This decrease in usefulness was a factor of 0.04, which translates to a 96% reduction in usefulness. The perceived difficulty of the search tasks was found to be consistent with the actual difficulty, indicating that the tasks were designed as intended. The results also showed that perceived task difficulty was a better predictor of user perception than task familiarity, as there was no correlation found between search task familiarity and difficulty. Therefore, while our research hypothesis H1: search task difficulty is correlated with user perception measures is

Table 2. Summary of the relationship between search task difficulty and user perception measures. N search task difficulty and user perception measures = 64, N user perception measures = 64; statistical significance at 95%.

	CutPoint (mean)	Odds ratio	Log odds	Stand. error	t-Value	Stat. signif.
familiarity	1.45	1.00	0.00	0.52	0.00	No
perceived_difficulty	2.86	12.37	2.52	0.60	4.16	Yes
sys_usefulness	3.56	0.04	-3.26	0.76	-4.31	Yes
topic_relevance	3.77	0.05	-3.00	0.75	-4.00	Yes
utility	3.42	0.10	-2.31	0.58	-3.97	Yes
sys_satisfaction	3.48	0.09	-2.37	0.60	-3.96	Yes
process_satisfaction	3.42	0.12	-2.13	0.57	-3.76	Yes

Notes: Familiarity is the user's familiarity with the task topic; perceived difficulty refers to how difficult the search task was; 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.

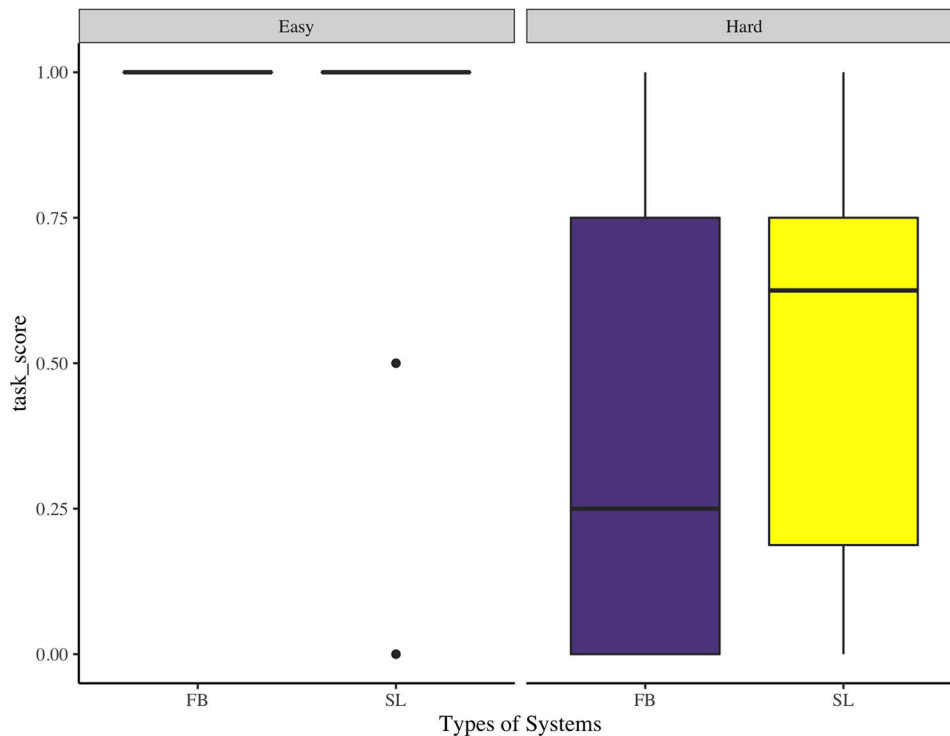


Figure 6. Boxplot of the types of systems and task score by search task difficulty.

partially supported, and the correlation is specifically with user perception measures about the systems and perceived task difficulty, rather than familiarity with the search task.

5.3. Search performance by search task difficulty

Search performance was evaluated through [0;1] scores for each step in the tasks (finding the right procedure, finding the right answer to the situation in the procedure, finding the next procedure, etc.). Since hard tasks had more steps, the score was normalized in [0;1] for each task (1 behind the maximum score). The overall results suggest that there was no significant difference in search performance by task score and time spent (min). However, there were very significant differences in search performance by search task. Figures 6 and 7 indicate that the proposed SL system enhanced task score for hard search tasks, but there was no significant difference by time spent. As expected, search task difficulty had significant effects on search performance. The SL system performed particularly well for hard search tasks.

Table 3. Model selection of fixed and random effects for system and user perception measures by task score.

	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.

5.4. Effect of system and user perception on task score

Table 3 presents the results of backward model selection (Kuznetsova et al., 2017) for the mixed-effects of system and user perception on task scores.

According to the results, the fixed effect of the system did not appear in the selected models, while the perceived

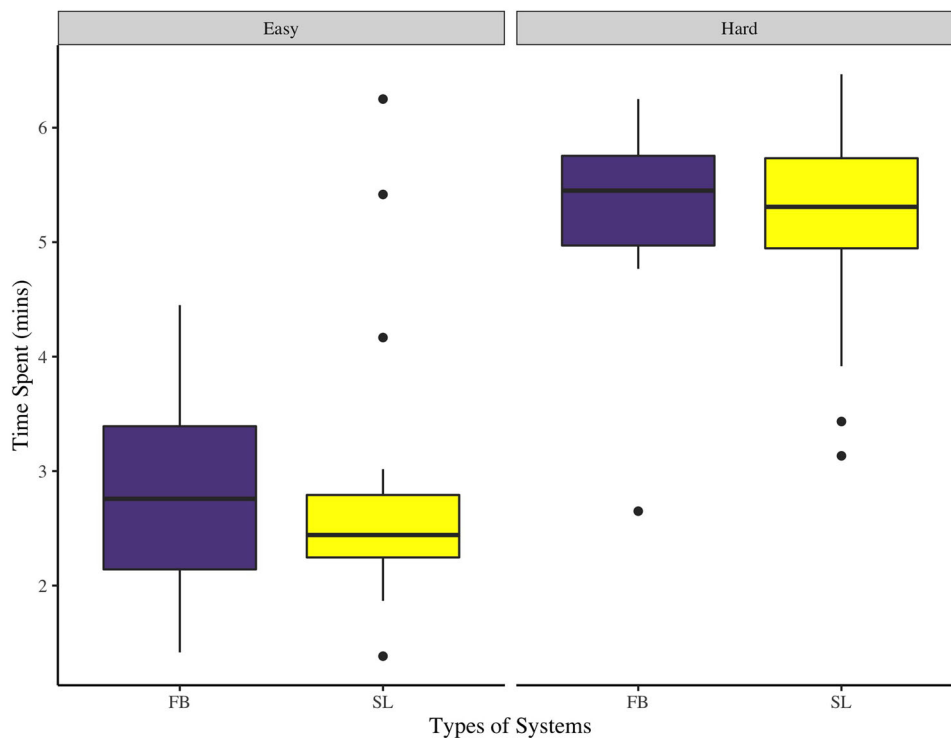


Figure 7. Boxplot of the types of systems and time spent (min) by search task difficulty.

Table 4. 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** (0.09)	-0.01 (0.15)	0.27* (0.15)	0.28* (0.15)	0.44** (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

* $p < .1$.

** $p < .05$.

*** $p < .01$.

usefulness of the system was influenced by the random effect of the user. The relevance of the system's responses to the topic and user satisfaction with the search process were influenced by the random effect of the task. The usefulness of the system's responses to find answers and user satisfaction with system responses were influenced by the random effects of both user and task. In summary, user perception measures were significant as fixed effects, while the random effects of user and task varied in the selected models.

Table 4 reveals that all the user perception measures made significant differences in the task score.

The best model, as determined by the Akaike information criterion (AIC), was model 1, where 65% of the variance was explained by the user-perceived usefulness of the system. Model 2, where 41% of the variance was explained by the relevance of the system's responses to the topic, followed

closely behind. Interestingly, in model 4, user satisfaction with the system's responses had an effect size of 21%. These findings suggest that system designers should prioritize features that enhance user-perceived usefulness (which relates to usability) and relevance of the system's responses to the topic (which relates to effectiveness), rather than solely focusing on user satisfaction as a predictor of search performance.

5.5. Effect of system and user perception on time spent

Table 5 presents the results of model selection for the mixed-effects of the system and user perception on time spent.

The results show that the system did not make any difference in the time spent on user perception measures. All the

perception measures had significant effects on the time spent. The random effects of both task and user were present in the models.

Table 6 shows that the marginal effect size of user perception measures ranged from 9% to 32%, while there was little difference in the total effect size. Overall, the results suggest that the perceived usefulness of the system is the best predictor of user search performance by time spent.

Therefore, our research hypothesis H2, which states that the types of search systems and user perceptions will affect user search performance, is partially supported. Specifically, our findings suggest that the search system is not correlated with user search performance as measured by task score and time spent. Instead, the user's perception of the usefulness of the system in completing the search task is a good predictor of both task score and time spent. Furthermore, the system's responses to the relevance of the topic are also a good predictor of task score.

5.6. Effect of perceived difficulty and user perception on task score

Table 7 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.

To summarize, when participants perceived a search task as difficult, they had more difficulty determining the relevance of

Table 5. Model selection of fixed and random effects for user perception measures by time spent (min).

	Fixed and random effects model
Model 1	sys_usefulness + (1 task) + (1 user)
Model 2	topic_relevance + (1 task) + (1 user)
Model 3	utility + (1 task) + (1 user)
Model 4	sys_satisfaction + (1 task) + (1 user)
Model 5	process_satisfaction + (1 task) + (1 user)

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.

the system's responses to the topic. User perceptions of the system's usefulness and satisfaction had significant effects on task scores, and there were significant interactional effects as well. Results from Tables 4 and 7 indicate that user perception of system usefulness was the best predictor of task score, and there was no correlation between perceived search task difficulty and task score. The constructed mixed-effects models had relatively large effect sizes, indicating that participants were adept at assessing their performance.

5.7. Effect of perceived difficulty and user perception on time spent

Table 8 reveals that the best model was the usefulness of the system, which accounts for 39% of all variances. All the user perception measures had significant effects on time spent.

Table 7. Effect of perceived search task difficulty and user perception on task score.

	Task score		
	Model 1	Model 2	Model 3
perceived_difficulty	-0.33*** (0.10)	-0.33*** (0.08)	-0.35*** (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*** (0.46)	1.66*** (0.37)	1.80*** (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

** $p < .05$.

*** $p < .01$.

Table 6. Effect of system and user perception on time spent (min).

	Time spent (min)				
	Model 1	Model 2	Model 3	Model 4	Model 5
sys_usefulness	-0.63*** (0.11)				
topic_relevance		-0.58*** (0.12)			
utility			-0.48*** (0.10)		
sys_satisfaction				-0.46*** (0.10)	
process_satisfaction					-0.37*** (0.10)
Constant	6.27*** (0.58)	6.23*** (0.66)	5.69*** (0.60)	5.64*** (0.62)	5.32*** (0.69)
N	64	64	64	64	64
Log likelihood	-80.17	-83.29	-83.07	-84.75	-87.82
AIC (Akaike information criterion)	170.34	176.58	176.14	179.51	185.63
ICC (intraclass correlation)	0.69	0.69	0.73	0.73	0.75
R ² (fixed)	0.32	0.23	0.19	0.18	0.09
R ² (total)	0.79	0.76	0.78	0.78	0.77

*** $p < .01$.

Table 8. Effect of perceived search task difficulty and user perception on time spent (min).

	Time spent (min)		
	Model 1	Model 2	Model 3
perceived_difficulty	0.29** (0.12)	0.35*** (0.13)	0.34*** (0.13)
sys_usefulness	-0.42*** (0.13)		
topic_relevance		-0.30** (0.15)	
utility			-0.26** (0.13)
Constant	4.70*** (0.84)	4.19*** (0.96)	3.97*** (0.87)
N	64	64	64
Log likelihood	-78.39	-80.92	-81.01
AIC (Akaike information criterion)	168.78	173.84	174.02
ICC (intraclass correlation)	0.68	0.69	0.71
R ² (fixed)	0.39	0.32	0.29
R ² (total)	0.81	0.79	0.80

** $p < .05$.

*** $p < .01$.

Given the findings of the effect of perceived difficulty and user perception on task score and time spent, our research hypothesis H3: perceived search task difficulty and user perceptions 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 by task score.

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. Additionally, user-perceived search task difficulty and its interactional effect with the relevance of the system's responses to the topic make a significant difference in search performance.

Our research findings indicate that user satisfaction with the system's responses may not be a reliable predictor of user search performance. The Alexa Prize Socialbot Grand Challenge, which aims to advance our understanding of human interactions with socialbots and improve UX, primarily evaluates success based on user satisfaction (Dinan et al., 2020). This includes both automatic metrics from the system and human evaluation through Amazon's Mechanical Turk. However, since these evaluations rely on approximations of user satisfaction, there may be a discrepancy between automatic metrics and human evaluation results. Other research on user interaction with intelligent assistants has shown that using interactional signals as

features can accurately predict user satisfaction with search dialogues (Kiseleva et al., 2016). While user satisfaction has been used as a criterion of search effectiveness for IR systems (Al-Maskari & Sanderson, 2010), our study suggests that for enhancing user search performance in a specific domain, the usefulness of the system in completing the task and the relevance of the system's responses to the topic are better predictors of search success than user satisfaction with the system. This aligns with previous studies (Su, 1992; Wu & Liu, 2003, 2011) that have found the value of search results to be a crucial factor in search success.

Our aim of adopting a holistic approach to understanding UX and performance is to bridge the gap between system-centric and human evaluations. This approach aligns with extrinsic evaluation, which focuses on how the use of the system contributes to external outputs, such as task completion (Schneider et al., 2010). In IIR studies, the user's judgment of usefulness has been proposed and utilized as an evaluation criterion (Cole et al., 2009; Vakkari et al., 2019). Other extrinsic evaluation efforts include an evaluation methodology that assesses the usefulness of an interactive, analytical question-answering system in supporting the writing of intelligence reports (Kelly et al., 2007; Small & Strzalkowski, 2009; Sun et al., 2011; Wacholder et al., 2007). These studies demonstrate how user-oriented evaluations can inform the design of conversational search systems for specific domains.

Our findings suggest that user-perceived usefulness and relevance of the system's responses to the topic can be used as metrics for evaluating current conversational search systems. User perceptions about the system's usefulness in completing the task and the relevance of its responses to the topic are good predictors of search performance. This demonstration of the applicability of the holistic approach used in previous information-seeking conversations (Wu, 2005; Wu & Liu, 2011) to the design and evaluation of conversational search systems in a specific domain can aid the development of such systems in the future.

The study emphasizes the significance of user characteristics such as levels of domain knowledge and self-assessment in professional contexts. Previous research has indicated that the situational awareness of professional pilots can be predicted based on their domain-specific expertise (Cak et al., 2020). Similarly, in the context of conversational search systems, users' perception is related to their understanding of the situation and how search interactions are interpreted. We found that the user-perceived task difficulty is a crucial factor that affects and interacts with the relevance of the topic, the usefulness of system responses, and user satisfaction with search performance, which is consistent with the previous findings that performance quality of the conversational agents is a significant factor for trust building (Rheu et al., 2021). Hence, future studies need to consider both situational awareness (Endsley, 2018, 2021) and user perceptions when developing tools to support conversational agents in professional settings such as supporting pilots in a cockpit.

The main objectives of this study were to develop evaluation protocols for user experiments and analyze user data to inform the design of conversational assistant systems to support pilots in cockpits. The study successfully achieved these objectives. Based on the findings, future research and development should focus on designing system support features that improve relevance judgment, such as snippets in search engine results pages and system feedback for software documentation collection. This work will involve addressing both usability and effectiveness issues in system development. Furthermore, investigating the correlations between system-centric metrics and user task scores is recommended to reconcile the discrepancy between system evaluation and human evaluation, as discussed in Dinan et al. (2020).

The generalizability of the research findings to other settings may be limited due to the homogeneous age and experience of the participants within a specific domain. Furthermore, enhancing the internal validity of the results could be achieved by increasing the sample size. To gain a deeper understanding of user search processes when interacting with a prototype conversational agent, future research can explore reading behavior during information search by examining user search and visual behaviors (e.g., Y.-H. Liu et al., 2022; Spiller et al., 2021; Wittek et al., 2016). Despite these limitations, the mixed-effects models employed in the study reveal a substantial effect size, highlighting the robustness of the experimental design.

Since the design and evaluation of conversational search systems are still emerging, it is advisable to further consider the social and emotional aspects of information-seeking dialogs, including trust-building and adherence to social norms when implementing them in specific domains (Rheu et al., 2021; Shneiderman, 2022; Svikhnushina & Pu, 2022). In future studies aiming to develop tools supporting conversational search systems in professional settings such as assisting pilots in a cockpit, it is crucial to account for both situational awareness (Endsley, 2018, 2021) and user perceptions about the conversational search system. This consideration will contribute to a more comprehensive and effective design of conversational search systems specifically tailored to address the unique requirements of professionals operating within safety-critical environments.

7. Conclusions

In this article, a collaborative research project between academia and industry is presented, which demonstrates a user-centered approach to the design and evaluation of conversational SUIs. This approach takes into account the user's search behavior and individual differences when interacting with the proposed conversational search system, and develops conversational search systems from the user's perspectives. The study finds that perceived search task difficulty and its interaction with the relevance of the system's responses to the topic have a significant impact on search performance. The user's perception of the system's usefulness in successfully completing the search task emerges as a strong predictor for both the task score and time spent.

User satisfaction with the system's responses, which has been widely used as evaluation metrics of conversational search systems, may not be a reliable predictor of user search performance.

Notes

1. <https://developer.amazon.com/alexaprize>
2. <https://github.com/deepset-ai/FARM>
3. https://av-info.faa.gov/sdrx/documents/JASC_Code.pdf
4. <https://github.com/deepset-ai/FARM>
5. <https://towardsdatascience.com/modern-question-answering-systems-explained-4d0913744097>

Acknowledgements

We acknowledge Monika Litvová's contribution to the design of search tasks as Cockpit and Operation Design engineer for AIRBUS. Authors in alphabetical order at AIRBUS AI Research.











Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was funded by Airbus Central Research & Technology. The views expressed herein are those of the authors and are not necessarily those of the Airbus or the authors' affiliated institutions.

ORCID

Ying-Hsang Liu  <http://orcid.org/0000-0001-6504-4598>
 Alexandre Arnold  <http://orcid.org/0000-0002-7850-6541>
 Gérard Dupont  <http://orcid.org/0000-0002-3284-2320>
 Catherine Kobus  <http://orcid.org/0009-0008-2367-1532>
 François Lancelot  <http://orcid.org/0009-0008-0776-0957>
 Géraud Granger  <http://orcid.org/0000-0002-4094-4669>
 Yves Rouillard  <http://orcid.org/0000-0003-4380-1839>
 Alexandre Duchevet  <http://orcid.org/0000-0002-9751-1194>
 Jean-Paul Imbert  <http://orcid.org/0000-0001-5082-1374>
 Nadine Matton  <http://orcid.org/0000-0002-1678-2022>

References

- Aliannejadi, M., Zamani, H., Crestani, F., & Croft, W. B. (2019). *Asking clarifying questions in open-domain information-seeking conversations* [Paper presentation]. 2019: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information, New York. <https://doi.org/10.1145/3331184.3331265>
- Al-Maskari, A., & Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology*, 61(5), 859–868. <https://doi.org/10.1002/asi.21300>
- Anand, A., Cavedon, L., Joho, H., Sanderson, M., & Stein, B. (2020). Conversational search (Dagstuhl Seminar 19461). *Dagstuhl Reports*, 9(11), 34–83. <https://doi.org/10.4230/DagRep.9.11.34>
- Aragon, C. R., & Hearst, M. A. (2005). *Improving aviation safety with information visualization: A flight simulation study* [Paper presentation]. CHI 2005: Proceedings of the Conference on Human Factors in Computing Systems, New York. <https://doi.org/10.1145/1054972.1055033>

- Arnold, A., Dupont, G., Kobus, C., & Lancelot, F. (2019). Conversational agent for aerospace question answering: A position paper. In *Proceedings of the 1st Workshop on Conversational Interaction Systems (WCIS at SIGIR)*. <https://drive.google.com/file/d/1yEw4Kj9c04PScdnaHkwPTiACBJ8AL-G3/view>
- Arvola, P., Vainio, J., Junkkari, M., & Kekäläinen, J. (2012). *Model for simulating result document browsing in focused retrieval* [Paper presentation]. IiiX 2012 – Proceedings of the 4th Information Interaction in Context Symposium: Behaviors, Interactions, Interfaces, Systems, New York, NY, USA. <https://doi.org/10.1145/2362724.2362764>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 51. <https://doi.org/10.18637/jss.v067.i01>
- Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). *Rasa: Open source language understanding and dialogue management*. <https://arxiv.org/abs/1712.05181>
- Borlund, P. (2016). A study of the use of simulated work task situations in interactive information retrieval evaluations: A meta-evaluation. *Journal of Documentation*, 72(3), 394–413. <https://doi.org/10.1108/JD-06-2015-0068>
- Braslavski, P., Savenkov, D., Agichtein, E., & Dubatovka, A. (2017). *What do you mean exactly? analyzing clarification questions in CQA* [Paper presentation]. CHIIR 2017 – Proceedings of the 2017 Conference Human Information Interaction and Retrieval, New York. <https://doi.org/10.1145/3020165.3022149>
- Cak, S., Say, B., & Misirlisoy, M. (2020). Effects of working memory, attention, and expertise on pilots' situation awareness. *Cognition, Technology & Work*, 22(1), 85–94. <https://doi.org/10.1007/s10111-019-00551-w>
- Carim, G. C., Saurin, T. A., Havinga, J., Rae, A., Dekker, S. W., & Henriqson, É. (2016). Using a procedure doesn't mean following it: A cognitive systems approach to how a cockpit manages emergencies. *Safety Science*, 89, 147–157. <https://doi.org/10.1016/j.ssci.2016.06.008>
- Carterette, B., Kanoulas, E., & Yilmaz, E. (2011). *Simulating simple user behavior for system effectiveness evaluation* [Paper presentation]. CIKM 2011 – Proceedings of the International Conference on Information and Knowledge Management, New York. <https://doi.org/10.1145/2063576.2063668>
- Case, D. O., & Given, L. M. (2016). *Looking for information: A survey of research on information seeking, needs, and behavior* (4th ed.). Emerald.
- Cockburn, A., Gutwin, C., Palanque, P., Deleris, Y., Trask, C., Coveney, A., Yung, M., & MacLean, K. (2017). *Turbulent touch: Touchscreen input for cockpit flight displays* [Paper presentation]. CHI '17 – Proceedings of the Conference on Human Factors in Computing Systems, New York. <https://doi.org/10.1145/3025453.3025584>
- Cole, M., Liu, J., Belkin, N., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., & Zhang, X. (2009). Usefulness as the criterion for evaluation of interactive information retrieval. In HCIR '09 – Proceedings of the Third Workshop on Human-Computer Interaction and Information Retrieval (pp. 1–4). <http://cuaslis.org/hcir2009/H CIR2009.pdf>
- Collobert, R., & Weston, J. (2008). *A unified architecture for natural language processing: Deep neural networks with multitask learning* [Paper presentation]. ICML [Paper presentation]. Learning [Paper presentation]. Proceedings of the 25th International Conference on Machine, 2008, New York. <https://doi.org/10.1145/1390156.1390177>
- Dang, H. T., Kelly, D., & Lin, J. (2006). Overview of the TREC 2003 question answering track. In *TREC 2006 – Proceedings of the Fifteenth Text REtrieval Conference* (pp. 54–68). <http://trec.nist.gov/pubs/trec16/papers/QA.OVERVIEW16.pdf>
- Deng, L., Hinton, G., & Kingsbury, B. (2013). *New types of deep neural network learning for speech recognition and related applications: An overview* [Paper presentation]. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada. <https://doi.org/10.1109/ICASSP.2013.6639344>
- Deriu, J., Rodrigo, A., Otegi, A., Echevoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1), 755–810. <https://doi.org/10.1007/s10462-020-09866-x>
- de Sá Siqueira, M. A., Müller, B. C. N., & Bosse, T. (2023). When do we accept mistakes from chatbots? The impact of human-like communication on user experience in chatbots that make mistakes. *International Journal of Human-Computer Interaction*, 1–11. <https://doi.org/10.1080/10447318.2023.2175158>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019 – Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Dietz, L., Verma, M., Radlinski, F., & Craswell, N. (2018). TREC complex answer retrieval overview. In *TREC 2018 – Proceedings of the Twenty-Seventh Text REtrieval Conference* (pp. 1–13). Nist. <https://trec.nist.gov/pubs/trec26/papers/Overview-CAR.pdf>
- Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I. V., Lowe, R., Prabhumoy, S., Black, A. W., Rudnicky, A. I., Williams, J., Pineau, J., Burtsev, M., & Weston, J. (2020). The second conversational intelligence challenge (ConvAI2). In S. Escalera & R. Herbrich (Eds.), *The NeurIPS '18 competition: From machine learning to intelligent conversations* (pp. 187–208). Springer International Publishing. https://doi.org/10.1007/978-3-030-29135-8_7
- Earle, R. H., Rosso, M. A., & Alexander, K. E. (2015). *User preferences of software documentation genres* [Paper presentation]. Proceedings of the Annual International Conference on the Design of Communication, New York. <https://doi.org/10.1145/2775441.2775457>
- Endsley, M. R. (2018). Expertise and situation awareness. In A. M. Williams, A. Kozbelt, K. A. Ericsson, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (2nd ed., pp. 714–742). Cambridge University Press. <https://doi.org/10.1017/9781316480748.037>
- Endsley, M. R. (2021). A systematic review and meta-analysis of direct objective measures of situation awareness: A comparison of SAGAT and SPAM. *Human Factors*, 63(1), 124–150. <https://doi.org/10.1177/0018720819875376>
- Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, 132, 138–161. <https://doi.org/10.1016/j.ijhcs.2019.07.009>
- Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Assessing significance in a fourfold table* (3rd ed.). John Wiley & Sons.
- Freund, L. (2013). A cross-domain analysis of task and genre effects on perceptions of usefulness. *Information Processing & Management*, 49(5), 1108–1121. <https://doi.org/10.1016/j.ipm.2012.08.007>
- Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1440–1448). IEEE.
- Gutwin, C., Cockburn, A., & Gough, N. (2017). *A field experiment of spatially-stable overviews for document navigation* [Paper presentation]. CHI '17 – Proceedings of the Conference on Human Factors in Computing Systems, New York, NY, USA. <https://doi.org/10.1145/3025453.3025905>
- Harper, D. J., Koychev, I., Sun, Y. X., & Pirie, I. (2004). Within-document retrieval: A user-centred evaluation of relevance profiling. *Information Retrieval*, 7(3/4), 265–290. <https://doi.org/10.1023/B:INRT.0000011207.45988.bb>
- Hearst, M. A. (2011). 'Natural' search user interfaces. *Communications of the ACM*, 54(11), 60–67. <https://doi.org/10.1145/2018396.2018414>
- Hertzum, M., & Simonsen, J. (2019). How is professionals' information seeking shaped by workplace procedures? A study of healthcare clinicians. *Information Processing & Management*, 56(3), 624–636. <https://doi.org/10.1016/j.ipm.2019.01.001>

- Hofmann, K., Mitra, B., Radlinski, F., & Shokouhi, M. (2014). *An eye-tracking study of user interactions with query auto completion* [Paper presentation]. CIKM 2014 – Proceedings of the 2014 ACM International Conference on Information and Knowledge Management, New York. <https://doi.org/10.1145/2661829.2661922>
- Kelly, D. (2007). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2), 1–224. <https://doi.org/10.1561/1500000012>
- Kelly, D., Wacholder, N., Rittman, R., Sun, Y., Kantor, P., Small, S., & Strzalkowski, T. (2007, May). Using interview data to identify evaluation criteria for interactive, analytical question-answering systems. *Journal of the American Society for Information Science and Technology*, 58(7), 1032–1043. <https://doi.org/10.1002/asi.20575>
- Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Brooks/Cole.
- Kiseleva, J., Williams, K., Hassan Awadallah, A., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016). *Predicting user satisfaction with intelligent assistants* [Paper presentation]. SIGIR 2016 – Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, For Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/2911451.2911521>
- Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 25(9), 1248–1258. <https://doi.org/10.1093/jamia/ocy072>
- Lee, H., & Pang, N. (2018). Understanding the effects of task and topical knowledge in the evaluation of websites as information patch. *Journal of Documentation*, 74(1), 162–186. <https://doi.org/10.1108/JD-04-2017-0050>
- Letondal, C., Vinot, J.-L., Pauchet, S., Boussiron, C., Rey, S., Becquet, V., & Lavenir, C. (2018). *Being in the sky: Framing tangible and embodied interaction for future airliner cockpits* [Paper presentation]. TEI 2018 – Proceedings of the 12th International Conference on Tangible, Embedded, and Embodied Interaction, New York. <https://doi.org/10.1145/3173225.3173229>
- Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6), 1822–1837. <https://doi.org/10.1016/j.ipm.2008.07.005>
- Liu, C., Liu, Y.-H., Gedeon, T., Zhao, Y., Wei, Y., Yang, F., & Zhangs, F. (2019). The effects of perceived chronic pressure and time constraint on information search behaviors and experience. *Information Processing & Management*, 56(5), 1667–1679. <https://doi.org/10.1016/j.ipm.2019.04.004>
- Liu, C., Liu, Y.-H., Liu, J., & Bierig, R. (2021). Search interface design and evaluation. *Foundations and Trends in Information Retrieval*, 15(3–4), 243–416. <https://doi.org/10.1561/15000000073>
- Liu, Y. H., Arnold, A., Dupont, G., Kobus, C., & Lancelot, F. (2020). Evaluation of conversational agents for aerospace domain. In *CEUR Workshop Proceedings (CIRCLE 2020: Joint Conference of the Information Retrieval Communities in Europe)*. http://ceur-ws.org/Vol-2621/CIRCLE20_21.pdf
- Liu, Y.-H., & Belkin, N. J. (2008). *Query reformulation, search performance, and term suggestion devices in question-answering tasks* [Paper presentation]. IiiX'08: Proceedings of the 2nd International Symposium on Information Interaction in Context, New York, NY. <https://doi.org/10.1145/1414694.1414702>
- Liu, Y.-H., Thomas, P., Gedeon, T., & Rusnachenko, N. (2022). *Search interfaces for biomedical searching: How do gaze, user perception, search behaviour and search performance relate?* [Paper presentation]. ACM SIGIR Conference on Human Information Interaction and Retrieval, New York, NY, USA. <https://doi.org/10.1145/3498366.3505769>
- Logacheva, V., Malykh, V., Litinsky, A., & Burtsev, M. (2020). ConvAI2 dataset of non-goal-oriented human-to-bot dialogues. In S. Escalera & R. Herbrich (Eds.), *The NeurIPS '18 Competition. The Springer Series on Challenges in Machine Learning* (pp. 277–294). Springer International Publishing. https://doi.org/10.1007/978-3-030-29135-8_11
- Lykke, M., Price, S., & Delcambre, L. (2012). How doctors search: A study of query behaviour and the impact on search results. *Information Processing & Management*, 48(6), 1151–1170. <https://doi.org/10.1016/j.ipm.2012.02.006>
- Mollashahi, E. S., Uddin, M. S., & Gutwin, C. (2018). *Improving revisitation in long documents with two-level artificial-landmark scrollbars* [Paper presentation]. Proceedings of the 2018 International Conference on Advanced Visual Interfaces, New York, NY, USA. <https://doi.org/10.1145/3206505.3206554>
- Moore, R. J., & Arar, R. (2019). *Conversational UX design: A practitioner's guide to the natural conversation framework*. ACM.
- Myers, C. M. (2019). *Adaptive suggestions to increase learnability for voice user interfaces* [Paper presentation]. Proceedings of the International Conference on Intelligent User Interfaces IUI '19 Companion, New York. <https://doi.org/10.1145/3308557.3308727>
- Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). *Voice interfaces in everyday life* [Paper presentation]. CHI '18 – Proceedings of the Conference on Human Factors in Computing Systems, New York. <https://doi.org/10.1145/3173574.3174214>
- Qu, C., Yang, L., Croft, W. B., Trippas, J. R., Zhang, Y., & Qiu, M. (2018). *Analyzing and characterizing user intent in information-seeking conversations* [Paper presentation]. SIGIR 2018 – Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA. <https://doi.org/10.1145/3209978.3210124>
- Qu, C., Yang, L., Croft, W. B., Zhang, Y., Trippas, J. R., & Qiu, M. (2019). *User intent prediction in information-seeking conversations* [Paper presentation]. CHIIR 2019 – Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, New York. <https://doi.org/10.1145/3295750.3298924>
- Radlinski, F., & Craswell, N. (2017). *A theoretical framework for conversational search* [Paper presentation]. CHIIR 2017 – Proceedings of the 2017 Conference Human Information Interaction and Retrieval, New York. <https://doi.org/10.1145/3020165.3020183>
- Rajpurkar, P., Jia, R., & Liang, P. (2018). *Know what you don't know: Unanswerable questions for SQuAD* [Paper presentation]. ACL 2018 – Proceedings of the Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Melbourne, Australia. <https://doi.org/10.18653/v1/P18-2124>
- Rheu, M., Shin, J. Y., Peng, W., & Huh-Yoo, J. (2021). Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human-Computer Interaction*, 37(1), 81–96. <https://doi.org/10.1080/10447318.2020.1807710>
- Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/1500000019>
- Saracevic, T., Kantor, P., Chamis, A. Y., & Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science*, 39(3), 161–176. [https://doi.org/10.1002/\(SICI\)1097-4571\(198805\)39:3<161::AID-AS12>3.0.CO;2-0](https://doi.org/10.1002/(SICI)1097-4571(198805)39:3<161::AID-AS12>3.0.CO;2-0)
- Schneider, A., Van Der Sluis, I., & Luz, S. (2010). Comparing intrinsic and extrinsic evaluation of MT output in a dialogue system. In *Proceedings of the 7th International Workshop on Spoken Language Translation* (pp. 329–336). <https://aclanthology.org/2010.iwslt-papers.16>
- Schwartz, M., Hash, C., & Liebrock, L. M. (2010). Term distribution visualizations with Focus+Context. *Multimedia Tools and Applications*, 50(3), 509–532. <https://doi.org/10.1007/s11042-010-0479-1>
- Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.
- Small, S., & Strzalkowski, T. (2009). HITIQA: High-quality intelligence through interactive question answering. *Natural Language Engineering*, 15(1), 31–54. <https://doi.org/10.1017/S1351324908004890>
- Spiller, M., Liu, Y.-H., Hossain, M. Z., Gedeon, T., Geissler, J., & Nürnberger, A. (2021). Predicting visual search task success from

- eye gaze data as a basis for user-adaptive information visualization systems. *ACM Transactions on Interactive Intelligent Systems*, 11(2), 1–25. <https://doi.org/10.1145/3446638>
- Steichen, B., Conati, C., & Carenini, G. (2014). Inferring visualization task properties, user performance, and user cognitive abilities from eye gaze data. *ACM Transactions on Interactive Intelligent Systems*, 4(2), 1–29. <https://doi.org/10.1145/2633043>
- Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Information Processing & Management*, 28(4), 503–516. [https://doi.org/10.1016/0306-4573\(92\)90007-M](https://doi.org/10.1016/0306-4573(92)90007-M)
- Sun, Y., Kantor, P. B., & Morse, E. L. (2011). Using cross-evaluation to evaluate interactive QA systems. *Journal of the American Society for Information Science and Technology*, 62(9), 1653–1665. <https://doi.org/10.1002/asi.21585>
- Svikhnushina, E., & Pu, P. (2022). PEACE: A model of key social and emotional qualities of conversational chatbots. *ACM Transactions on Interactive Intelligent Systems*, 12(4), 1–29. <https://doi.org/10.1145/3531064>
- Tang, M.-C., Liu, Y.-H., & Wu, W.-C. (2013). A study of the influence of task familiarity on user behaviors and performance with a MeSH term suggestion interface for PubMed bibliographic search. *International Journal of Medical Informatics*, 82(9), 832–843. <https://doi.org/10.1016/j.ijmedinf.2013.04.005>
- Thomas, P., Czerwinski, M., McDuff, D., Craswell, N., & Mark, G. (2018). *Style and alignment in information-seeking conversation* [Paper presentation]. CHIIR 2018 – Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, New York. <https://doi.org/10.1145/3176349.3176388>
- Trippas, J. R., Spina, D., Cavedon, L., Joho, H., & Sanderson, M. (2018). *Informing the design of spoken conversational search* [Paper presentation]. CHIIR 2018 – Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, New York. <https://doi.org/10.1145/3176349.3176387>
- Turnbull, D., & Berryman, J. (2016). *Relevant search: With applications for Solr and Elasticsearch*. Manning Publications.
- Urigo, K., Arguello, J., & Capra, R. (2019). *Anderson and Krathwohl's two-dimensional taxonomy applied to task creation and learning assessment* [Paper presentation]. ICTIR 2019 – Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, New York. <https://doi.org/10.1145/3341981.3344226>
- Vakkari, P., Völske, M., Potthast, M., Hagen, M., & Stein, B. (2019). Modeling the usefulness of search results as measured by information use. *Information Processing & Management*, 56(3), 879–894. <https://doi.org/10.1016/j.ipm.2019.02.001>
- Vakulenko, S., Revoredol, K., Di Ciccio, C., & de Rijke, M. (2019). QRFA: A data-driven model of information-seeking dialogues. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, & D. Hiemstra (Eds.), *Advances in information retrieval. ECIR 2019. Lecture notes in computer science* (Vol. 11437, pp. 541–557). Springer International Publishing. https://doi.org/10.1007/978-3-030-15712-8_35
- von Thaden, T. L. (2008). Distributed information behavior: A study of dynamic practice in a safety critical environment. *Journal of the American Society for Information Science and Technology*, 59(10), 1555–1569. <https://doi.org/10.1002/asi.20842>
- Voorhees, E. M. (2008). Evaluating question answering system performance. In T. Strzalkowski & S. M. Harabagiu (Eds.), *Advances in open domain question answering* (pp. 409–430). Springer Netherlands. https://doi.org/10.1007/978-1-4020-4746-6_13
- Vtyurina, A., Savenkov, D., Agichtein, E., & Clarke, C. L. A. (2017). *Exploring conversational search with humans, assistants, and wizards* [Paper presentation]. CHI EA '17: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, New York. <https://doi.org/10.1145/3027063.3053175>
- Wacholder, N. (2008). *How should users access the content of digital books?* [Paper presentation]. Proceedings of the 2008 ACM Workshop on Research Advances in Large Digital Book Repositories, New York, NY, USA. <https://doi.org/10.1145/1458412.1458424>
- Wacholder, N., Kelly, D., Kantor, P., Rittman, R., Sun, Y., Bai, B., Small, S., Yamrom, B., & Strzalkowski, T. (2007). A model for quantitative evaluation of an end-to-end question-answering system. *Journal of the American Society for Information Science and Technology*, 58(8), 1082–1099. <https://doi.org/10.1002/asi.20560>
- Wittek, P., Liu, Y.-H., Darányi, S., Gedeon, T., & Lim, I. S. (2016). Risk and ambiguity in information seeking: Eye gaze patterns reveal contextual behavior in dealing with uncertainty. *Frontiers in Psychology*, 7, 1790. <https://doi.org/10.3389/fpsyg.2016.01790>
- Wu, M.-M. (2005). Understanding patrons' micro-level information seeking (MLIS) in information retrieval situations. *Information Processing & Management*, 41(4), 929–947. <https://doi.org/10.1016/j.ipm.2004.08.007>
- Wu, M.-M., & Liu, Y.-H. (2003). Intermediary's information seeking, inquiring minds, and elicitation styles. *Journal of the American Society for Information Science and Technology*, 54(12), 1117–1133. <https://doi.org/10.1002/asi.10323>
- Wu, M.-M., & Liu, Y.-H. (2011). On intermediaries' inquiring minds, elicitation styles, and user satisfaction. *Journal of the American Society for Information Science and Technology*, 62(12), 2396–2403. <https://doi.org/10.1002/asi.21644>
- Zamani, H., Dumais, S., Craswell, N., Bennett, P., & Lueck, G. (2020). *Generating clarifying questions for information retrieval* [Paper presentation]. Proceedings of the Web Conference 2020, New York. <https://doi.org/10.1145/3366423.3380126>
- Zhang, Y., Wei, Y., & Yang, Q. (2018). Learning to multitask. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 5776–5787). Curran Associates Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/aeefb050911334869a7a5d9e4d0e1689-Paper.pdf

About the authors

Ying-Hsang Liu is a Researcher at the Department of ALM in Uppsala University, Sweden. He holds a PhD in Information Science from Rutgers University. His research lies at the intersections of knowledge organization, interactive information retrieval, and human information behavior, with particular emphasis on search interface design and evaluation.

Alexandre Arnold is an AI Senior Research Scientist at Airbus Central R&T in the Artificial Intelligence domain, specializing in Reinforcement Learning. He has been involved in multiple projects related to Natural Language Processing, including as a research project leader for LEA (LEarning Assistant) to ease the creation of customized chatbots.

G erard Dupont is Head of Machine Learning at Mavenoid, an innovative Swedish startup. He holds a PhD in Computer Science from the Universit  de Rouen. During his tenure at Airbus, he was an active participant in various international research projects in the areas of language understanding and large-scale information retrieval.

Catherine Kobus is a Research Scientist at Airbus Central R&T working on research projects around Natural Language Processing and Speech recognition. She holds a PhD in Computer Science from University of Avignon. Before joining Airbus, she worked in other groups like Orange, Nuance Communications and Systran.

Fran ois Lancelot is a Research Engineer at Airbus Central R&T working on research projects around Natural Language Processing and Speech recognition.

G raud Granger is a Research Engineer at ENAC, French Civil Aviation University. He holds a PhD in Computer Science since 2002, and continues his work in artificial intelligence, mathematics and statistics for civil aviation.

Yves Rouillard is an Engineer and Instructor at ENAC (French National School of Civil Aviation). He has contributed to research projects on cockpit simulators, onboard digital assistant and eye interaction.

Alexandre Duchevet is a cockpit operation engineer and a PhD student. After three years in the aeronautical industry, he joined the ENAC (French National School of Civil Aviation) to work on the benefits of artificial intelligence in cockpits. His main research topics are human factors and smart assistants in aviation.

Jean-Paul Imbert holds a PhD in computer science, neuroscience and human factors. He holds a position of research engineer at ENAC (French National School of Civil Aviation) in a team dedicated to applied human system interaction and human factor research in Air Traffic Control and Cockpit operations.

Nadine Matton is an associate professor at ENAC (French National School of Civil Aviation) and also affiliated to the CLLE laboratory (Cognition Langues Langage Ergonomie). She received her PhD in cognitive psychology in 2008 at the University of Toulouse. Her research interests focus on cognitive skills in aviation.

Appendix A. Demographic Questionnaire

(1) What is your native language?

- English
- French
- Chinese
- Spanish [Database]
- Other

(2) How many years of general aviation experience do you have?

- Not at all
- Less than five years
- Five to ten years
- Ten to fifteen years
- More than fifteen years

(3) Have you ever had any flying experiences as an amateur or student pilot?

- Yes
- No

(4) If you had flying experiences, how many hours?

(5) Have you ever had any flying experiences as a professional pilot?

- Yes
- No

(6) If you had flying experiences in a commercial airline, how many hours?

(7) How often do you search for information using a search engine, such as Google, Bing, or Baidu?

- Not at all
- Several times a month
- Several times a week
- Every day [Database]
- Several times a day or more

(8) How often do you use a virtual assistant, such as Siri, Alexa, Duer or OK Google?

- Not at all
- Several times a month
- Several times a week
- Every day [Database]
- Several times a day or more

(9) Are you a student?

- Yes, an undergraduate
- Yes, a postgraduate
- No

(10) If you are a student, what is your specific area of study?

(11) How old are you?

- Younger than 18
- 18 or older and not yet 25
- 25 or older and not yet 35
- 35 or older and not yet 45
- 45 or older

Appendix B. Post-Search Questionnaire

(1) How familiar were you with the topic of the task?

- 1 Not at all
- 2 Slightly
- 3 Fairly [Database]
- 4 Very
- 5 Extremely

(2) How difficult was the search task?

- 1 Not at all
- 2 Slightly
- 3 Fairly [Database]
- 4 Very
- 5 Extremely

(3) How useful was the system in completing the task?

- 1 Not at all
- 2 Slightly
- 3 Fairly [Database]
- 4 Very
- 5 Extremely

Appendix C. Exit Questionnaire

(1) What factors affected the level of difficulty?

(2) How could a conversational assistant be designed to make this kind of search task easier?

(3) Any other suggestion or comment?

(4) How relevant were the search results?

- 1 Not at all
- 2 Slightly
- 3 Fairly [Database]
- 4 Very
- 5 Extremely

(5) How useful were the search results?

- 1 Not at all
- 2 Slightly
- 3 Fairly [Database]
- 4 Very
- 5 Extremely

(6) How satisfied were you with the search results?

- 1 Not at all
- 2 Slightly
- 3 Fairly [Database]
- 4 Very
- 5 Extremely

(7) How satisfied were you with the search process?

- 1 Not at all
- 2 Slightly
- 3 Fairly [Database]
- 4 Very
- 5 Extremely

(8) How satisfied were you with the system interaction?

- 1 Not at all
- 2 Slightly
- 3 Fairly [Database]
- 4 Very
- 5 Extremely