Re-evaluating the Command-and-Control Paradigm in Conversational Search Interactions

Johanne R. Trippas
RMIT University
Melbourne, Australia
j.trippas@rmit.edu.au

Luke Gallagher
The University of Melbourne
Melbourne, Australia
gallagher.l@unimelb.edu.au

Joel Mackenzie
The University of Queensland
Brisbane, Australia
joel.mackenzie@uq.edu.au

ABSTRACT

Conversational assistants are becoming prevalent among the wider population due to their simplicity and increasing utility. However, the shortcomings of these tools are as renowned as their benefits. In this work, we present a "first look" at an extensive collection of conversational queries, aiming to identify limitations and improvement opportunities specifically related to information access (i.e., search interactions). We explore over 600,000 Google Assistant interactions from 173 unique users, examining usage trends and the resulting deficiencies and strengths of these assistants. We aim to provide a balanced assessment, highlighting the assistant's shortcomings in supporting users and delivering relevant information to user needs and areas where it demonstrates a reasonable response to user inputs. Our analysis shows that, although most users conduct information-seeking tasks, there is little evidence of complex information-seeking behaviour, with most interactions consisting of simple, imperative instructions. Finally, we find that conversational devices allow users to benefit from increased naturalistic interactions and the ability to apply acquired information in situ, a novel observation for conversational information seeking.

CCS CONCEPTS

• **Information systems → Query log analysis**.

KEYWORDS

Conversational Information Seeking; Log Analysis; Dataset

ACM Reference Format:
Johanne R. Trippas, Luke Gallagher, and Joel Mackenzie. 2024. Re-evaluating the Command-and-Control Paradigm in Conversational Search Interactions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3627673.3679588

1 INTRODUCTION

Log analysis — a method for examining computer-generated records (logs) — has a well-established application in information retrieval and system-user behaviour analysis [2, 9, 19, 26, 46, 48]. More recently, log analysis has been applied to investigating and comparing written and voice queries in the context of *web search* [21, 22]. Despite the growth in conversational information-seeking research, there is a notable lack of investigation into queries with commercially available conversational assistants [52, 58]. Similarly, while interaction logs have been investigated for voice assistance [4, 10, 20], these studies have focused on an interaction viewpoint and provide limited investigation for conversational information-seeking interactions. Furthermore, these studies were often limited to short-term studies on geographically narrow user bases, potentially skewing the understanding of the system's adaptability and long-term utility.

Our study aims to address this research gap by conducting an analysis of over 600,000 Google Assistant interactions from 173 users,[1] spanning more than two years. We conduct an in-depth examination of user demographics and interaction patterns, with a specific focus on complex information seeking behaviour. This extensive data demonstrates that despite Google Assistant's versatile capabilities, the majority of interactions (more than 60%) predominantly consist of simple, one-turn *command-and-control* exchanges. This dominant usage for straightforward interactions emphasises the limitations (whether real or perceived) of state-of-the-art conversational assistants in managing more complex tasks, highlighting that these systems are not yet fully conversational [6, 39]. As such, our analysis is vital in better understanding the gap between the devices' potential multi-functionality and their actual usage in the wild. Furthermore, our exploration of the albeit limited set of complex information seeking tasks uncovers a range of behavioural patterns that provide insights for developing more user-centric conversational assistants, bridging the gap between intended functionalities and real-world performance, and guiding future advancements in conversational AI technology. Ultimately, this research highlights limitations of conversational AI systems and informs their development trajectory, ensuring that they evolve to meet complex and varied user needs. Our work has the following key contributions:

- To the best of our knowledge, we present the most comprehensive analysis of a large-scale conversational interaction log from an information-seeking perspective.
- We illustrate that people still mainly use Google Assistant for non-informational *command-and-control* inputs tasks, as well as simple informational questions regarding daily routine.
- We highlight the gap between the potential for information-seeking conversations (based on system capabilities) and their real-world usage. For instance, an ideal (smart) assistant would seamlessly help to plan and arrange a holiday, although, in practice, we are still far from this goal [23].
- We specify that conversational seeking interactions have unique features. For instance, the ability to fluidly switch contexts and reformulate queries reflects the dynamic nature of human inquiry,

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679588

[1] Note, we use users, participants, and respondents interchangeably.

and; the unique query method of interleaving information needs within a session, where users alternate between unrelated tasks.
- We release the first freely available Conversational Information Retrieval Query Log — CIRQL — from in-the-wild interactions.[2]

To sum up, while Google Assistant users still primarily engage in home automation tasks, our research indicates that there is *some* evidence of complex information seeking behaviour; however, these instances are simplistic, and lag behind advancements in the theory of conversational search [44].

2 RELATED WORK

Log Analyses. Log analysis, the process of examining and interpreting computer-generated records, is commonly used to understand system and user behaviour, or commonly used user actions [29]. Advantages of log analysis include identification and resolution of system issues and improved system performance and user experience analysis [10, 26, 52]. However, since log analyses involve handling large volumes of data, there is potential for false positives in anomaly detection and privacy concerns related to sensitive data [49]. The impact of log analysis is substantial in improving system performance and understanding user behaviour [19, 48, 51, 59].

Many studies have used log analysis to investigate behavioural data, providing insights into the interaction patterns of users with systems like voice assistants. Furthermore, these studies improve our understanding of how users utilise these applications, which is crucial for identifying their strengths and weaknesses [10, 52]. Log analyses are often used for a comprehensive and data-driven approach to understanding the intricacies of user interactions with information systems. These analyses can provide valuable insights into the system's limitations and strengths, essential for targeted improvements and technological advancements.

While prior work has offered insight into conversational interactions [10, 20, 22], further work is needed to determine how users interact with conversational assistants on a longer-term basis, especially for information-seeking. In this work, we aim to identify the features, limitations, and strengths of Google Assistant users, and to provide a deeper analysis on their use for addressing complex information needs. We also aim to examine the relationship between user interaction patterns and information retrieval effectiveness. Insights from our study are thought to help inform the development of more user-centric conversational assistants.

Crowdsourcing Logs. Crowdsourcing logs from Google Assistant devices involves collecting and analysing user-generated data from a large user base to understand interaction patterns and behaviours with these systems [3, 42, 60]. This approach provides insights into how users utilise voice assistants and web search functionalities. The advantages of crowdsourcing logs include accessing diverse user data, which can lead to more comprehensive and varied insights [60]. It enables rapid data collection,the study of user behaviours (or preferences) at scale, which can ultimately aid processes of system improvement and customisation. However, disadvantages include potential quality issues with the data due to varying levels of user expertise and engagement. There is also a risk of bias if the sample of workers do not represent the population of users in question; and privacy concerns may arise from the collection of personally identifiable information. Nevertheless, crowdsourcing logs in this way is one current approach that is generally available to entities across a wide range of resource budgets. The benefit of this is to improve system quality and user experience (or value), and the dissemination of such knowledge.

Prior work conducted a lab study whereby digital voice assistants (i.e. Google Assistant) were adapted for crowdsourcing tasks [24]. The authors report that task efficacy via voice interaction was comparable to the traditional user interface. Our work is in line with the idea of collecting user data from a conversational setting for a specific task, but rather than logs of crowd workers, the frame is to understand information-seeking in the task of web search for Google Assistant enabled devices.

Command-and-control Interactions. Conversational assistants, exemplified by well-known systems like Siri, Alexa, and Google Assistant, work predominantly on a command-and-control paradigm, where user inputs are typically explicit queries or commands [25, 39, 58]. Recent advancements focus on enhancing these assistants beyond the command-and-control paradigm to a more interactive and context-aware experience. For instance, researchers are exploring how conversational agents can leverage acoustic-based activity recognition to become more aware of the user's context [1, 13]. Other research has suggested to shift from traditional command-and-control to multi-modal interactions, blending structured dialogue with information-seeking and questioning-answering [17, 47]. Additionally, the role of conversational cues in conversations with assistants is being examined, indicating a potential for more nuanced interactions beyond simple commands [53, 58]. This emerging trend suggests a future where conversational assistants could offer more dynamic, context-sensitive, and user-centric experiences, moving away from the traditional command-and-control model towards complex session-oriented information seeking [34].

3 METHODOLOGY

3.1 Data Collection Setup

To collect natural conversational interactions, we conduct an online crowdsourcing study [3, 51, 60]. We collect data from respondents (i.e., workers) on the Prolific crowdsourcing platform who have identified themselves as Google Assistant users. The workers were asked to fill out a survey with questions about *(i)* eligibility and willingness to download and upload Google Assistant interaction logs (i.e., screener), *(ii)* demographics information, *(iii)* technology usage, *(iv)* Google and non-Google Assistant ownership and usage and *(v)* interaction upload and review. From this task, two distinctive kinds of data were obtained: (1) the self-reported *survey*; and (2) the interaction *logs*. As such, all of the data is collected from *in-the-wild* users issuing real queries, allowing for reduced bias as compared to lab or crowdsourcing studies. We provide an overview of the collected data with descriptive, time, and session analyses, summarising the characteristics and utility of the data [12, 27, 57].

[2] https://github.com/JTrippas/CIRQL

Crowd Workers. Crowd workers on Prolific could access our task if they met specific requirements:[3] residing in the United States, United Kingdom, Ireland, Australia, Canada, or New Zealand; owning internet-enabled home assistants or smart hubs like Amazon Echo, CastleHub, or Google Home; being over 18 years old; fluent in English; having completed more than 1,000 previous submissions; and maintaining an approval rate exceeding 95%. We inquired further to confirm the eligibility of participants. We asked them whether they possessed Google Assistant-enabled products such as Google Home, Google Nest Audio, or Google Nest Wifi Point with Google Assistant built-in. Additionally, we assessed their ability to download logs associated with their Google account and their willingness to upload these logs.

Workers were paid 2.98 GBP for a submission. The reward was estimated based on the average completion time for all the pilot tasks (15 minutes) and the Australian minimum hourly wage. A maximum time of 56 minutes was set by Prolific's estimated completion time. Workers could only take part in the study once. No further attention checks were in place; empty logs (and corresponding survey data) were discarded. Collection of interaction logs was completed with RMIT ethics approval. Crowd workers were employed through 17 November to 18 December 2023. Workers conducted a compulsory review of their interaction logs to opt-out any entries they cared to omit from the study — via a browser-based offline interface — prior to submission over the web.

3.2 Demographics Questionnaire

We collected self-reported demographic information through a Qualtrics survey. The average participant age was 37 years, ranging from 20 to 72 years, with a standard deviation of 10 years. Most participants (67.63%) described themselves as male, while 30.64% as female and 1.73% as non-binary/third gender. The majority (71.10%) of participants were based in the United States, with the United Kingdom following at 25.43%. Canada and Australia saw similar participant representation, both at 1.73%.

Language Skills and Education Levels. Nearly all participants reported having native English language skills (95.38%), 3.47% reported being fluent, and 1.16% that they had advanced and moderate English skills. Education levels reported the majority of participants held a university degree (i.e., graduate or professional degree (18.50%), and bachelor's degree (47.40%)), indicating a high educational accomplishment sample. The third largest group (15.03%) had pursued some university education without completing a degree. The sample exhibited representation from participants with secondary and vocational or similar educational backgrounds (7.51% and 9.83%, respectively). Partial secondary education (1.16%) and primary education (0.58%) were less common, while no participants reported less than primary education.

Household Size. This dataset's distribution of household members (people who stay in the household at least half of the time) shows that the number of occupants in a household ranges between one and eight, with a trend toward smaller households of four occupants or less. Single-person households represent 16.76%, and over a third (35.84%) of households consist of only two members. Households with three or four members account for 18.50% and 19.65%, respectively. While households with five (7.51%) or more members (1.74%) constitute a smaller proportion.

Search Engine Usage and Search Skills. The majority of the participants use search engines more than six times a day (61.27%). The second largest group (31.21%) indicated they use search engines two to five times daily. Participants indicating using a search engine once a day account for 3.47%. The remaining participants (4.04%) indicated that they used search engines between once and six times a week. The self-assessment distribution of individuals' search skills shows that the majority considers their skills as moderately good (54.34%), followed by those who perceive their skills as extremely good (38.73%), slightly good (4.62%), and neither good nor bad (2.31%). No respondents rated their skills as bad in any category.

Google Assistant Satisfaction and Ownership. Overall, the system garnered a generally positive response, with a majority of users expressing some level of satisfaction with Google Assistant. The majority of users (84.97%) reported being either extremely (12.14%), moderately (55.49%), or slightly satisfied with the system (17.34%). Only a small proportion of users (10.10%) were either neutral or expressed some degree of dissatisfaction.

The majority of participants (64.74%) reported having one device in their household, with 24.28% reporting two and 10.99% more than three. Google-branded devices, such as Google Home were reported most followed by the Google Nest Mini. Self-reported alternatives to the Google-branded Assistant speakers included Google Pixel phones, Lenovo Smart Clocks, and Insignia Voice with Google Assistant Speaker. Respondents identified where their speakers were located, and the data indicates a predominant inclination towards shared home-based spaces. The majority of respondents favour bedrooms (32.16%) and living rooms (31.37%). Kitchens are preferred by 21.96%, while work offices, other locations (this includes the hallway or cupboard), bathrooms, dining rooms, and garages are less favoured, accounting for 5.88%, 4.71%, 2.35%, 1.18%, and 0.39% respectively. Respondents not only use Google Assistant, but 61.85% of the respondents indicated using other non-Google voice-enabled systems. The majority of the time, respondents indicated that they are using Amazon Alexa (76.64%) and Apple Siri (42.06%). There was 2.80% indicating other Assistants which were Bixby (two mentions) and Samsung Galaxy Home Mini (one mention). Of the 107 respondents who reported that they use other voice-enabled systems, 34.58% indicated that they use the system daily, 14.95% between 4–6 times a week, 17.76% between 2–3 a week, and 28.04% once a week. Five respondents (4.67%) indicated that they never use the other voice-enabled systems.

3.3 Pre-processing Google Assistant Data

An interaction log for a user is a series of Google Assistant events that were obtained by asking respondents to export their activity history, from their Google Account, in JSON format. Figure 1 is an example of an entry from the activity history. It shows various attributes about the event that occurred, and depending on the event, different attributes may be present or omitted. For brevity, we describe the pertinent attributes to our discussion. An event datum has a timestamp (*time*), the event input (*title*), and optional event

[3]Note, these are existing screener sets from Prolific, and we further specify our sample within our Qualtrics survey.

```
{ "header": "Assistant",
  "title": "Said what are autistic traits",
  "titleUrl": "https://www.google.com/search?q=what+are+autistic+
     traits",
  "subtitles": [{ "name": "Signs of autism in adults - NHS",
    "url": "https://www.nhs.uk/conditions/autism/signs/adults/" }],
  "time": "2022-01-15T11:57:57.848Z",
  "products": ["Assistant"],
  "activityControls": ["Web & App Activity"] }
```


Figure 1: An example log entry collected from respondents. This particular instance shows a user turn (*title*) and corresponding system response turn (*subtitles*).

Table 1: General statistics of the collected data in terms of users, sessions, and interactions.

Item	Frequency
Respondents	173
Total interactions	627,978
Total sessions	292,098
One-interaction sessions	161,902
Utterance-interactions	392,919
CIRQL queries (conversational search)	50,190

output (*subtitles*). The example in Figure 1 is a "conversation event", the user's dialogue is the event input and the assistant's response is the event output. In this case the output is a search result, but it may also consist of the system's dialogue. The prefix term "*Said*" from the event input provides us with the context of the action that took place. From these prefix terms, we extract a set of "event types" as a proxy for the event context. There were 27 *prefix terms* identified, and then categorised into 11 logical classes as the different kinds of interactions that took place with the Google Assistant device. These categories are summarised in Table 2. Invalid or inconsistent events, such as duplicates and irrelevant submissions were discarded, including respondents having 20 or fewer interaction events. We derive sessions of user interactions by denoting a session boundary after there was 15 minutes of inactivity [21, 50, 52].[4]

4 INTERACTION ANALYSIS

In this section, we provide some basic interaction descriptives to summarise the dataset and explore common usage patterns.

4.1 Overview and Usage Patterns

Respondents in the study submitted a total of 627,978 interactions from around 292,098 sessions; Table 1 provides a brief overview. Figure 2 shows in decreasing order, the frequency of assistant interactions as a function of the respondents. A long tail exists where participants have in total, fewer than 5,000 interactions, contrasted with the more active participants having upwards of 30,000 total interactions. Overall, this amounts to each respondent, on average, submitting 3,630 interactions with a standard deviation of 5,648.[5]

[4]Session denotation includes all interaction events and is not limited to only the "user utterance" events.

[5]Respondents or households – we are unable to reliably detect unique device users.

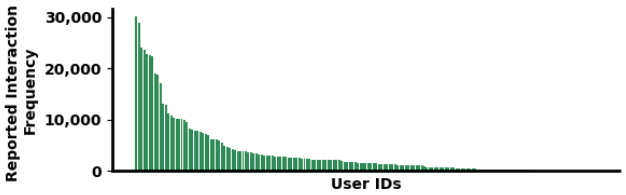

Figure 2: Frequency of the number of interactions per user.

Temporal behaviour. We now describe some of the temporal statistics in different time units of months (Figure 3), days, and hours (Figure 4). The volume of interactions on a month-by-month basis is shown in Figure 3. Evidently, the most active respondents (left side of the figure) have over 500 monthly interactions, and are consistent in their use across the three-year collection period. This contrasts with inconsistent, low-activity respondents (right side), who often go months without interactions.

Drilling down to day-of-the-week usage, we see that interactions are evenly distributed over each day of the week (an average of 90,685±1,227 interactions per day Monday to Friday, inclusive). In contrast to the findings of Bentley et al. [10], we observed fewer interactions on both Saturday (88,364), and Sunday (86,188) as compared to the rest of the week.

We also investigated the interaction distribution across different hours of the day, see Figure 4. The normalised percentages reveal patterns in user engagement, with higher percentages during evening and night hours on weekdays and weekends. Specifically, there is a noticeable peak in activity around 21:00 on weekdays and 17:00 on weekends and a consistent decrease in the early morning hours. Weekday percentages exhibit a slightly different distribution than weekends, including peak usage during mid-day, perhaps suggesting distinct user behaviour patterns during these periods.

Sessions. User session data (annotated in the pre-processing phase, see Section 3.3) can give insights into common usage patterns. There were 292,098 user sessions, amounting to an average of 1,688 sessions per user (standard deviation 2,169). Figure 5 reports the distribution of sessions demonstrating the heavy-tailed distribution (many short sessions, very few long sessions). The majority of the sessions (55.42% or 161,902) contained a single interaction, with 64,889 (22.21%) two-interaction sessions, and 24,567 (8.41%) three-interaction sessions. Hence, only 13.95% of sessions contain four or more interactions. These aggregate results indicate a predominant user inclination towards isolated assistant engagements.

Ignoring single interaction sessions, we find that the average session duration is just 37 seconds. Figure 6 plots the duration of each session grouped by the total number of interactions, and demonstrates that as the interactions within a session increase, so too does duration. The longest session was a clear outlier, and lasted for almost 3 hours; the user was observed to be playing various games with the device, with over 500 recorded interactions across the session. Other long sessions were observed to include multiple rounds of question/answering interactions as well as requests for various songs to be played by the assistant.

Interaction Classes. Next, we investigate the frequency *interaction classes* (i.e., the label from Google Assistant specifying the user

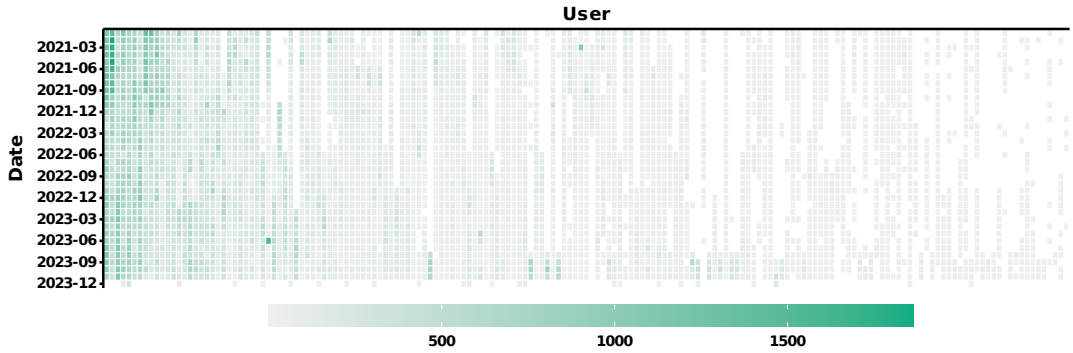

Figure 3: Heatmap of per-month user interactions across the entire collection, sorted according to their total interaction volume.

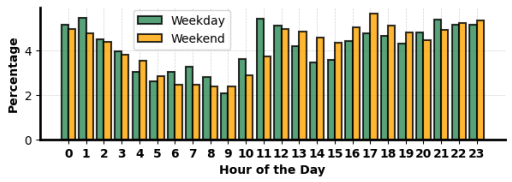

Figure 4: Normalised session frequency by hour of day.

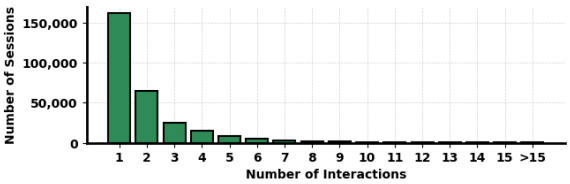

Figure 5: Distribution of sessions by interaction count.

input type) across the entire dataset, enhancing our understanding of user behaviour, input, and preferences. A summary of the *interaction classes* are given in Table 2. The interaction class labels denote the context of the logged event in a user's activity history. The most popular class, *Command*, makes up nearly two thirds of all interactions; these interactions consist of user utterances, which include queries (*"how much electricity does an electric blanket use"* or *"what time is it?"*), as well as commands such as *"stop"* or *"turn off the light"*. The remaining interactions, which are less relevant to our analysis, include system log entries (e.g., the smart assistant being activated and subsequently timing out, alarms being triggered), user utterances within specific contexts like games or other applications, and a relatively high percentage of unknown voice commands (7.17%). The latter indicates potential weaknesses in the assistant's voice recognition and points to possible user difficulties in articulating comprehensible commands.

4.2 Task and Utterance Themes

Next, we turn our attention to user interactions from the *Command* interaction class, since these represent the user's verbalised input such as queries and commands. We employ a suite of techniques to analyse these themes.

Table 2: *Interaction class* description and their frequency.

Class	Description	Freq. (%)
Command	An utterance	392,919 (63%)
Notification	Log notification received, opened, etc.	72,219 (12%)
Timeout	Log wake word activation and timeout	69,919 (11%)
Error	Log "unknown voice command"	45,002 (7%)
Alarm	Log set/rang an alarm	22,168 (4%)
Task	Log assistant task performed	16,711 (3%)
App	Utterances while using an app/game	4,461 (1%)
UI	Non-utterance interactions	3,011 (<1%)
Messaging	Log sent a message/email/call	1,107 (<1%)
Calendar	Log calendar item viewed/added	454 (<1%)
Other	Item redacted	7 (<1%)
Total		627,978

***N*-Grams**. Our first approach is to compute and analyse commonly occurring *N*-grams from user spoken inputs. Table 3 presents the top 10 most common unigrams, bigrams, and trigrams from user input in the *Command* class after applying Porter stemming. This data, comprising 2,199,655 words, reveals "*the*" as the most frequent unigram (119,486), indicative of its prevalence in English texts. Other common words such as "*turn*", "*light*", "*on*", or "*off*", point to interactive or command-driven language usage. Similarly, the prevalence of bigrams like "*turn on/off*" (37,999 and 35,634), or trigrams like "*time is it*" (10,625), and "*a timer for*" (8,178) reflect an orientation towards simple routine tasks. These interactions suggest a functional, directive, command-and-control nature of user-assistant interactions and imply that users primarily engage with the assistant for simple, repetitive tasks. Unsurprisingly, the frequent occurrence of unigrams starting with "*plai*" (50,442 in unigram) — formed from words such as "*play*" — suggests that entertainment-related commands are expected, indicating a substantial aspect of the assistant's role in providing leisure. The unigram "*what*" (47,840) and the bigram "*what the*" (18, 822) might reflect user queries or information needs. Overall, this analysis corroborates that Google Assistant is extensively used for simple and routine tasks [10].

Query Classification. To better understand the intent behind the *Command* interaction class (i.e., utterances that include queries), we employed a state-of-the-art multi-label classifier from Kubis et al.

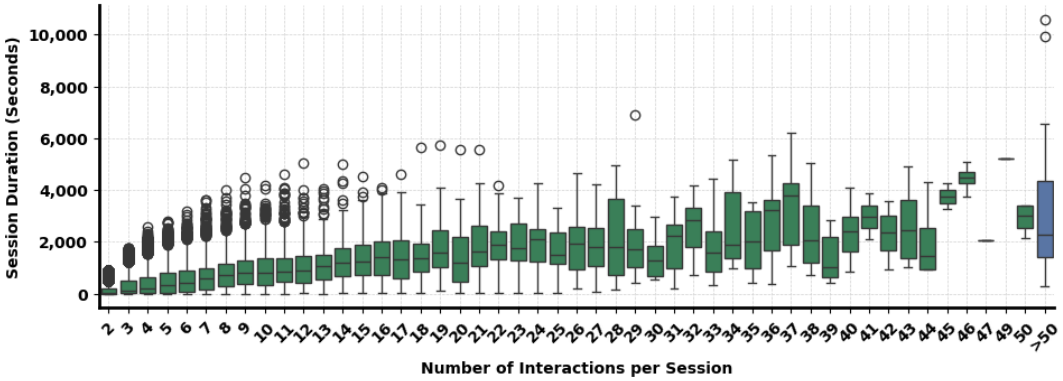

Figure 6: The x-axis is the number of interactions in a session. The y-axis is session duration measured in seconds. The median is shown by the horizontal line in each group. Sessions having more than 50 interactions are aggregated in the rightmost box (blue). Session duration is broadly correlated with the number of interactions within a session.

Table 3: The ten most frequently occurring unigrams, bigrams, and trigrams from 392,919 user utterances. Totals of each ngram set are 2,199,655 unigrams, 1,573,534 bigrams, and 1,065,570 trigrams.

Uni.	Freq.	Bi.	Freq.	Tri.	Freq.
the	114,684	turn off	37,914	turn on the	22,373
turn	102,973	turn on	35,585	turn off the	22,089
light	75,973	on the	25,847	what time is	10,625
on	70,805	off the	23,175	time is it	10,162
off	59,968	the light	20,554	live room light	9,559
plai	50,442	what the	18,822	the live room	8,563
what	47,840	turn the	15,477	a timer for	8,178
set	33,681	timer for	14,533	set a timer	7,826
for	33,422	live room	14,111	what the weather	7,680
is	27,496	is it	13,490	set an alarm	7,138

Table 4: The highest frequency classes across the entire set of user *Command* interactions (i.e., utterances) with examples. The Other class consists of several low-frequency classes such as Cooking, Calendar, Transport, News, Email, and Recommendation.

Class	Frequency (%)	Examples
IOT	103,990 (26.5%)	turn on the lights; turn off the tv
Play	77,118 (19.6%)	play music; play white noise
Audio	66,818 (17.0%)	stop; pause; set the speaker volume...
Alarm	21,900 (5.6%)	set morning alarm; cancel timer
Weather	21,395 (5.4%)	what's the temperature; weather today
QA	21,273 (5.4%)	what does a dinosaur sound like
General	19,069 (4.9%)	good morning; tell me a joke
Datetime	16,788 (4.3%)	what time is it; what's the date today
Music	14,207 (3.6%)	skip; next song; what song is this
Other	30,361 (7.7%)	next ingredient; tell me the news

[33].[6] Based on XLM-RoBERTa [15], the classifier was trained on the MASSIVE dataset [18], which is tailored for training and evaluating models on intelligent assistant-based natural language understanding tasks. Table 4 shows the distribution of assigned class labels across all user utterances. We see IOT (internet-of-things), Play, and Audio classes corresponding to over 63% of all utterances, indicating a majority of *command-and-control* interactions. Bentley et al. [10] reported similar findings, with 40% of their interactions related to the *audio* domain — as compared to 40.2% of our interactions (across Play, Audio , and Music categories) — and 6% for both *weather* and *alarm* respectively (compared to 5.4% and 5.6% in our data). This analysis is a further contribution to the simplistic *command-and-control* nature of voice enabled digital assistants.

5 CONVERSATIONAL QUERYING ANALYSIS

From our analysis so far, we can see that the vast majority of interactions in the logs are simple, transactional, imperative exchanges. In this section, we aim to better understand the extent in which users engage in complex information seeking tasks — such as extended information seeking sessions with evolving information needs [32, 34] — and to determine what sort of complex human-device interactions exist beyond the imperative paradigm.

5.1 Identifying Conversational Queries

Given the large amount of command-and-control queries in our log, our next task is to identify the subset of *conversational search* [44, p. 4] queries for further analysis. That is, we wish to understand if the collected assistant data presents any new evidence of *conversational information seeking* [58, p. 15] interactions that meet the multi-turn mixed-initiative definitions put forth by prior work [41, 44].

Starting with the 392,919 *Command* interaction class utterances, we used a systematic, iterative, filtering approach to remove entries that are not genuine queries. As a first step, we removed non-English input, emojis, addresses, and phone numbers. We identified and categorised non-query entries, such as requests for weather information, commands to set timers, play music, control home appliances, adjust volume, add items to a shopping list, or broadcast messages. After each iteration, we verified the data to exclude all non-query interactions. This process resulted in a final dataset comprising mainly of genuine queries. The disadvantage is that it is time-consuming and complex, requiring significant manual effort, and the possibility of introducing bias.

[6] https://huggingface.co/cartesinus/xlm-r-base-amazon-massive-domain

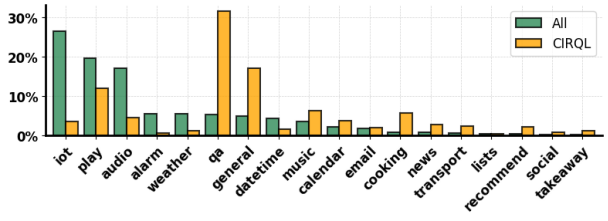

Figure 7: Comparison between the proportion of each category on the entire set of *Command* utterances and the iterative filtered CIRQL dataset as classified by the state-of-the-art classifier [33].

CIRQL Dataset. Our identified query sample, called CIRQL (Conversational Information Retrieval Query Log), consists of 50,190 queries, representing just 12.7% of the original 392,919 *Command* interaction class utterances. These queries were extracted from 31,885 sessions, accounting for 11% of the total sessions. Figure 7 shows the large distribution shift in the categories due to the filtering process, with CIRQL containing a much larger proportion of QA and General queries as compared to the unfiltered data, as expected. Similarly, the filtering process greatly reduces the proportion of IOT, Play, and Audio classes, which predominantly represent command-and-control interactions.

5.2 Conversational Query Characteristics

Query length. The queries vary in length from between 1 to 89 words, with an average length of 5.59 words with a 3.38 words standard deviation. The query subset contains 3,331 one-word queries, representing 6.64% of the total identified queries.

To understand how our queries compare to natural questions, we compare our subset with the MS MARCO dataset [5], which consists of QA-style queries derived from Microsoft Bing logs. The query lengths in MS MARCO range from 1 to 75 words. The standard deviation of the query lengths in the MS MARCO dataset is slightly lower, at 2.65 words, suggesting a more consistent distribution of query lengths around the average of 6.37 words. The proportion of one-word queries in the MS MARCO dataset is notably lower, amounting to just 35 queries or 0.004%. This comparative analysis highlights differences in the distribution and characteristics of query lengths between the two datasets. Our query subset shows greater variability in query length, and a greater proportion of one-word queries. While the distribution of Microsoft Bing queries can be sampled in a controlled manner, it is not quite the case for the query set we identify. Another criterion that must be acknowledged is that different input modalities come with different constraints, for example, a spoken query has no notion of a backspace key leading to inflated one-word queries in spoken contexts. Another angle here, shown by prior work [16] draws a connection that spoken queries are more verbose than written queries.

Interrogative words. We investigate the frequency of the first words in our CIRQL subset, highlighting the prominence of interrogative words. Common first words like "*how*" (7,995 occurrences) and "*what*" (7,326 occurrences) align with the function of interrogative words in eliciting specific information, and mimic QA querying patterns like those observed in MS MARCO. These words often guide search engines to interpret user intent [9, 30]. Other interrogative words such as "*when*", "*who*", "*where*", "*why*", and "*which*" frequently appear, indicating their role in seeking information. In fact, these interrogative words form the first term in 42.8% of the 50,190 sampled queries.

It is of interest to note non-interrogative first words such as "*can*", "*is*" "*do*", and "*show*" also feature in our query subset (> 4,500 occurrences, 9.1%), often starting requests for actions or confirmations (e.g., "*can you show …*", or "*is it …*"). Words like "*i*", "*the*", "*no*", and "*ok*" reflect a mix of command or statement-based queries, and is perhaps some evidence of discourse exhibiting natural form.

Table 5 shows the 15 most common bigrams and trigrams from the first two and three words in the dataset. These results indicate a preference for queries that seek specific information such as definitions, quantities, and of duration. The frequency pattern given by these phrases an example that helps gives some indication to the extent of information-seeking capabilities users explore. diversity in query patterns, including more detailed trigrams, highlights varied user interests and information needs. The frequency pattern given by these phrases gives some weight in support of our query filtering process from Section 5.1, and the trigrams in particular are of an *informational* form [11].

Table 5: The 15 most common bigrams and trigrams in user input from the CIRQL dataset, with their respective percentages.

Bigram	Freq. (%)	Trigram	Freq. (%)
what is	2,904 (5.79%)	how do you	1,052 (2.10%)
how many	2,510 (5.00%)	what is the	909 (1.81%)
what's the	1,530 (3.05%)	how much is	515 (1.03%)
how long	1,386 (2.76%)	how many days	467 (0.93%)
how do	1,337 (2.66%)	how many calories	462 (0.92%)
how much	1,120 (2.23%)	how long does	412 (0.82%)
can you	915 (1.82%)	what are the	349 (0.70%)
what does	776 (1.55%)	how old is	319 (0.64%)
what are	571 (1.14%)	what is a	268 (0.53%)
when is	448 (0.89%)	how do i	251 (0.50%)
show me	399 (0.79%)	what sound does	215 (0.43%)
how old	381 (0.76%)	how far is	202 (0.40%)
where is	348 (0.69%)	how long is	189 (0.38%)
do you	345 (0.69%)	how long do	179 (0.36%)
how far	299 (0.60%)	how much does	170 (0.34%)

System responses. We investigate the system responses to user queries. While the system could often respond sensibily, there was a volume of queries — about 3,000 in total — in which the system did not retrieve any information or reverted back to a traditional search, with *"Sorry, … "* or *"Here at the top results … "* responses. While these responses are expected, they may contribute to a perceived lack of satisfaction or effectiveness in the system's ability to handle diverse and complex queries. This perception, together with the reinforcement of the system's conversational information seeking limitations, can undermine user confidence and trust. It highlights the need for improving response quality to better meet user expectations and information needs.

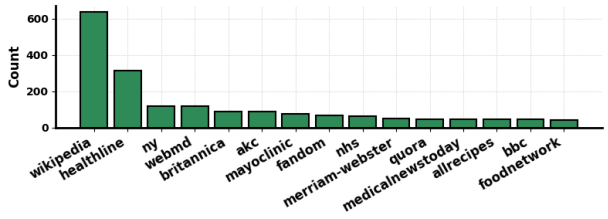

Figure 8: The number of responses from the top-15 most common domains used by the system during response generation.

5.3 Qualitative Session Exploration

Our final task is to analyse the sessions from the CIRQL dataset, to discover key trends of in-the-wild conversational information seeking. To achieve an in-depth understanding of user search behaviour, we categorise and manually review the 50,190 queries. We use a qualitative approach to inspect detailed intent behind user queries [29].

Methodology. From our filtered conversational query set — CIRQL, as described in Section 5.1 — we identify sessions that contain complex information seeking behaviour for further analysis. As a first step, we extract the 7,616 responses from 4,229 unique sessions that contain a URL embedded along with the assistant response, indicating that the response was retrieved from the given webpage (descriptive statistics of URLs are discussed below). Our intuition is that interactions that result in the assistant returning information from a webpage are highly likely to be information seeking interactions; we use this as a proxy to identify relevant sessions for further analysis. Next, we manually examined all sessions containing at least five URL-bearing responses, assuming such interactions might indicate more complex information-seeking behaviours. Three authors performed open coding during these sessions, systematically categorising the data into themes and patterns [31]. They collaboratively discussed the coding process and used an agreement protocol, which included meetings to resolve discrepancies and achieve consensus. This protocol involved iterative rounds of coding and cross-validation to ensure reliability and consistency. They coded all sessions, ultimately discarding five due to offensive language and lack of clear information-seeking objectives, as these appeared random and aimed at eliciting reactions, possibly as limitation testing. We found that most sessions containing URL-bearing responses were indeed related to information seeking; in the interest of brevity, we summarise the key insights from our analysis.

Descriptives. Of the entire set of almost 300,000 sessions, fewer than two percent involve an agent response retrieved from the web, highlighting a potential lack of information-seeking behaviour. These sessions were derived from 109 users, indicating that the remaining 64 users had no interactions of this type. From the 7,616 responses bearing a URL, there were a total of 6,861 unique URLs from 3,076 unique domains. Figure 8 shows the distribution of the top-15 response domains; `wikipedia.org` was the most commonly utilised response domain, followed by `healthline.com`, `ny.gov`, `webmd.com`, and `britannica.com`.

Overall, only 26 (of 173) users contributed to the 70 information seeking sessions we analysed, and more than half of those sessions can be attributed to just five users. This trend indicates that most users seek "single-use" information, infrequently performing various unique or less common searches. Meanwhile, a smaller group of "curious" users are more engaged and consistently seek out new information through multiple searches. Table 6 reports statistics on the 70 information-seeking sessions that were analysed in detail. The longest session had 584 queries, of which only 14 responses carried a URL; this particular session went for almost three hours. On average, each of the 70 sessions had 24 queries, 6 of which had responses carrying a URL.

Evolving Information Need. One complex interaction pattern was evidenced through a user's knowledge seemingly evolving through the duration of the interaction [7, 56]; this pattern was observed in the majority of the sessions analysed (57 out of 70). Users would often ask related follow-up questions and occasionally ask the assistant to "continue" or answer in the affirmative if the agent prompted the user to decide to continue or not. These sessions demonstrate evolving information needs as users adapt their queries while gathering information and contextual understanding, characteristic of exploratory search [40, 55]. This process involves iterative seeking, browsing, and refining, leading to a gradual buildup of understanding and evolving queries [8]. In addition, some sessions highlight the unique aspect of conversational search, which is the ability to use relevant information in situ. A user may simultaneously interact the device, while applying the newfound knowledge to the (physical) task (or activity) at hand. We believe this to be a novel finding from the CIRQL logs.

Context Switching. Another pattern that emerged was the use of context switching within a session [45]. For example, a user may issue a query relating to a given topic, and then switch to a completely different topic or task within the same session, without *switching* back to the original topic. This was observed in 33 unique sessions. Mixed intents could include the user switching between information seeking tasks (different topics); or entirely different tasks (listening to the news, setting alarms, playing games). The observation of context switching within conversational information seeking represents a deviation from the traditional linear, topic-specific search, although a refined analysis would be required to strengthen any further claims [28].

Interleaving and Search Temporality. Unique temporal patterns in conversational information seeking sessions were also observed, previously reported for non-conversational interactions [32, 34]. One such pattern was interleaved intents (four sessions), where users momentarily diverted from their primary information-seeking task to engage in a different activity, before returning to their original task (contrasted with context switching, where users did *not* return to their original task). Certain users were also observed exploring similar or repeated topics across extended periods, encompassing weeks, months, or even years (i.e., successive searches over multiple sessions [14, 37]). Although infrequent (or difficult to detect), these recurring search patterns provide insights into individual user preferences. This could perhaps enhance personalisation and recommendation features in future pro-active assistants [54].

Focused Tasks. Contrasted with sessions involving context switching and interleaving was a set of focused sessions involving a single

Table 6: Summary of the 70 sessions from 26 users containing at least 5 queries with a URL-bearing response.

Session Attribute	Mean	Std.	Min.	Med.	Max.
Queries (URL response)	6.40	2.02	5.00	6.00	14.00
Queries (Total)	24.11	68.75	5.00	13.00	584.00
Duration (Minutes)	19.44	21.72	2.06	13.95	165.23
Sessions (Per user)	2.69	2.74	1.00	1.00	11.00

information need (20 in total). While these sessions were still characterised by evolving knowledge and follow-up questioning, the queries were always related a single underlying topic [38].

Conversational Interactions. Our analysis demonstrates the use of mixed-initiative sessions (50 in total), where both user and assistant take turns, creating a collaborative exchange. In addition, our findings suggest that conversational session boundaries may differ from traditional web search sessions. Unlike web search, where sessions are often defined by a series of related queries within a short timeframe (e.g., 15 minutes), queries within conversational sessions reported a broader range of topics. More importantly, conversations typically require continuous engagement between parties and do not tolerate long pauses; hence a significant gap in dialogue would typically signal a session boundary in our data.

Other Patterns. Various other use patterns were observed, including query reformulation (39 sessions); an entire session consisting of arithmetic expressions; evidence of multi-device interaction; sessions seemingly dedicated to testing the limits and constraints of the assistant; instances of anthropomorphism, and users becoming frustrated with their assistant. These different patterns may further illustrate user reliance on the assistant for quick tasks, or unmet expectations or misunderstandings in the interaction process while avoiding more complex interactions. These interactions will be further investigated in our future work.

6 DISCUSSION AND CONCLUSION

In this paper, we examined long-term conversational interaction behaviours from a diverse set of Google Assistant users. The observed usage patterns reveal three distinctive trends: *(i)* a prevalence of *command-and-control* interactions, *(ii)* a focus on routine tasks, and *(iii)* limited evidence of conversational search interactions. We found that most interactions with the assistant are shallow, imperative instructions such as setting alarms, turning on lights, or adding to a shopping list. These simple interactions suggest that users are neither prompted nor inclined to use the assistant's advanced capabilities, aligning with the findings of Bentley et al. [10]. Additionally, our observations indicate that users primarily use Google Assistant for "low-bandwidth" or routine tasks with straightforward interpretations, such as obtaining factual information, rather than integrating it into complex decision-making processes.

This gap emphasises how large language models can potentially make multi-turn interactions more natural and aligned with user interests [43]. Enhancing assistants with ever-changing large language model capabilities may improve the user experience by facilitating more complex and dynamic interactions, thereby making the assistant appear more intelligent and responsive. The digital assistant we examined already relies on sophisticated distributed search infrastructures for processing user utterances, routing requests, and delivering precise responses. Therefore, advancing towards more fluid, human-like interactions is not a distant goal, and is already being illustrated by emerging large language model-based technology tailored to conversational settings [35, 36].

On the other hand, there were a few information-seeking sessions, we observed as having various information-seeking behaviours that align with those documented in past literature, such as berry-picking behaviour, evolving information needs, and context switching. These prior patterns can frame the strategies a user might employ to find information, model the nature of their information needs, as transition between different topics or tasks occur within a single session. The nature of transient voice interactions, characterised by their immediacy and conversational flow, still presents unique challenges and opportunities for interactive information retrieval. In this view, session boundaries could form from the natural flow of an intelligent assisted dialogue.

The system response sources emphasise a limited range of domains, with Wikipedia as the most common source. This dominance of a few key websites reflects a concentrated reliance on specific authoritative sources for information retrieved by conversational assistants. Nevertheless, this could lead to a lack of diverse perspectives in information access and highlights the need to broaden the range of sources conversational assistants use. In addition, the limitations of the system were exacerbated by a tendency of reverting to traditional search result lists, or apologising that the system could not handle the user request. These system limitations may underscore constraints of the system (or years of user conditioning in web search) and emphasise that there are many open challenges before the ideal conversational agent (system) is practical.

Several factors may influence the limited use of conversational assistants for complex information-seeking. Perhaps users perceive these assistants as suitable for straightforward tasks while overlooking their potential for handling more complex queries. Alternatively, users may not always want to share their information needs through voice with a conversational assistant if sensitive information is read out. Additionally, habitual usage patterns may have established these tools as quick solutions for routine tasks, reinforcing their role in providing basic information. Furthermore, conversational assistant interaction design, favouring brief communication, may be a barrier to articulating and processing complex information needs.

Limitations. Our focus on a specific audience segment introduces a limitation in terms of generalisability. In addition, the system could sometimes not detect interactions in the files, resulting in empty files. Google Assistant received updates over the data collection period, potentially leading to inconsistencies in the data. Our filtering approach may have also resulted in the loss of some nuances in the utterances, such as short responses like "yes" or "continue". Finally, although we employed various techniques (both qualitative and quantitative) to explore and filter our data, the potential bias introduced through our filtering processes must be acknowledged.

Acknowledgements. This work was supported in part by the Australian Research Council's Discovery Projects Scheme (project DP200103136) and partially funded by The University of Melbourne. The third author was supported by a Google Research Scholar grant.

REFERENCES

[1] Rebecca Adaimi, Howard Yong, and Edison Thomaz. 2021. Ok google, what am I doing? Acoustic activity recognition bounded by conversational assistant interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–24.

[2] Eytan Adar, Jaime Teevan, and Susan T Dumais. 2008. Large scale analysis of web revisitation patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1197–1206.

[3] Omar Alonso and Maria Stone. 2014. Building a query log via crowdsourcing. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 939–942.

[4] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 3 (2019), 1–28.

[5] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, and Tri Nguyen an. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. *arXiv preprint* abs/1611.09268 (2016).

[6] Saminda Sundeepa Balasuriya, Laurianne Sitbon, Andrew A. Bayor, Maria Hoogstrate, and Margot Brereton. 2018. Use of voice activated interfaces by people with intellectual disability. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*. 102–112.

[7] Marcia J Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* 13, 5 (1989), 407–424.

[8] Nicholas J Belkin. 1980. Anomalous states of knowledge as a basis for information retrieval. *Canadian journal of information science* 5, 1 (1980), 133–143.

[9] Michael Bendersky and W Bruce Croft. 2009. Analysis of long queries in a large scale search log. In *Proceedings of the 2009 workshop on Web Search Click Data*. 8–14.

[10] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.

[11] Andrei Broder. 2002. A taxonomy of web search. *SIGIR Forum* 36 (2002), 3–10.

[12] Wolfgang Büschel, Anke Lehmann, and Raimund Dachselt. 2021. Miria: A mixed reality toolkit for the in-situ visualization and analysis of spatio-temporal interaction data. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[13] Justin Chan, Thomas Rea, Shyamnath Gollakota, and Jacob E Sunshine. 2019. Contactless cardiac arrest detection using smart devices. *NPJ digital medicine* 2, 1 (2019), 52.

[14] Chun Wei Choo, Brian Detlor, and Don Turnbull. 1998. A Behavioral Model of Information Seeking on the Web–Preliminary Results of a Study of How Managers and IT Specialists Use the Web.. In *American Society for Information Science (ASIS)*.

[15] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.

[16] Fabio Crestani and Heather Du. 2006. Written versus spoken queries: A qualitative and quantitative comparative analysis. *Journal of the American Society for Information Science and Technology* 57, 7 (2006), 881–890.

[17] Yashar Deldjoo, Johanne R Trippas, and Hamed Zamani. 2021. Towards multimodal conversational information seeking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1577–1587.

[18] Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 4277–4302.

[19] Adam Fourney, Richard Mann, and Michael Terry. 2011. Characterizing the usability of interactive applications through query log analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1817–1826.

[20] Radhika Garg, Hua Cui, and Yash Kapadia. 2021. "Learn, Use, and (Intermittently) Abandon": Exploring the Practices of Early Smart Speaker Adopters in Urban India. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[21] Ido Guy. 2016. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 35–44.

[22] Ido Guy. 2018. The characteristics of voice search: Comparing spoken with typed-in mobile web search queries. *ACM Transactions on Information Systems (TOIS)* 36, 3 (2018), 1–28.

[23] Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. 2014. Supporting complex search tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 829–838.

[24] Danula Hettiachchi, Zhanna Sarsenbayeva, Fraser Allison, Niels van Berkel, Tilman Dingler, Gabriele Marini, Vassilis Kostakos, and Jorge Goncalves. 2020. "Hi! I am the Crowd Tasker" Crowdsourcing through digital voice assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[25] Matthew B Hoy. 2018. Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly* 37, 1 (2018), 81–88.

[26] Bernard J Jansen and Amanda Spink. 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management* 42, 1 (2006), 248–263.

[27] Peiling Jiang, Fuling Sun, and Haijun Xia. 2023. Log-it: Supporting Programming with Interactive, Contextual, Structured, and Visual Logs. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[28] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. 699–708.

[29] Steve Jones, Sally Jo Cunningham, Rodger McNab, and Stefan Boddie. 2000. A transaction log analysis of a digital library. *International Journal on Digital Libraries* 3 (2000), 152–169.

[30] Emilia Kacprzak, Laura M Koesten, Luis-Daniel Ibáñez, Elena Simperl, and Jeni Tennison. 2017. A query log analysis of dataset search. In *Proceedings of the 17th International Conference on Web Engineering*. 429–436.

[31] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.

[32] Alexander Kotov, Paul N. Bennett, Ryen W. White, Susan T. Dumais, and Jaime Teevan. 2011. Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 5–14.

[33] Marek Kubis, Paweł Skórzewski, Marcin Sowański, and Tomasz Zietkiewicz. 2023. Back Transcription as a Method for Evaluating Robustness of Natural Language Understanding Models to Speech Recognition Errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 11824–11835.

[34] Yuan Li, Rob Capra, and Yinglong Zhang. 2020. Everyday cross-session search: how and why do people search across multiple sessions?. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 163–172.

[35] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive Conversational Agents in the Post-ChatGPT World. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3452–3455.

[36] Guan-Ting Lin, Cheng-Han Chiang, and Hung yi Lee. 2024. Advancing Large Language Models to Capture Varied Speaking Styles and Respond Properly in Spoken Conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. To appear.

[37] Shin-jeng Lin and Nick Belkin. 2005. Validation of a model of information seeking over multiple search sessions. *Journal of the American Society for Information Science and Technology* 56, 4 (2005), 393–415.

[38] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. 2011. Identifying task-based sessions in search engine query logs. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. 277–286.

[39] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5286–5297.

[40] Garry Marchionini. 1997. *Information Seeking in Electronic Environments*. Cambridge University Press.

[41] Paul Owoicho, Jeffrey Dalton, Mohammad Aliannejadi, Leif Azzopardi, Johanne R Trippas, and Svitlana Vakulenko. 2022. TREC CAsT 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation. In *The Thirty-First Text REtrieval Conference Proceedings (TREC '22)*.

[42] Martin Potthast, Matthias Hagen, Michael Völske, and Benno Stein. 2013. Crowdsourcing interaction logs to understand text reuse from the web. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1212–1221.

[43] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*. 28492–28518.

[44] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. 117–126.

[45] Soo Young Rieh. 2006. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management* 42, 3 (2006), 751–768.

[46] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large web search engine query log. *ACM SIGIR Forum* 33, 1 (1999), 6–12.

[47] Alessandro Speggiorin, Jeffrey Dalton, and Anton Leuski. 2022. TaskMAD: A Platform for Multimodal Task-Centric Knowledge-Grounded Conversational Experimentation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3240–3244.

[48] Amanda Spink, Bernard J Jansen, Dietmar Wolfram, and Tefko Saracevic. 2002. From e-sex to e-commerce: Web search changes. *Computer* 35, 3 (2002), 107–109.

[49] Jan Svacina, Jackson Raffety, Connor Woodahl, Brooklynn Stone, Tomas Cerny, Miroslav Bures, Dongwan Shin, Karel Frajtak, and Pavel Tisnovsky. 2020. On vulnerability and security log analysis: A systematic literature review on recent trends. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*. 175–180.

[50] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. 2011. #TwitterSearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*. 35–44.

[51] Johanne R Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. 2024. What do Users Really Ask Large Language Models? An Initial Log Analysis of Google Bard Interactions in the Wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. 2703–2707.

[52] Johanne R Trippas, Damiano Spina, Mark Sanderson, and Lawrence Cavedon. 2021. Accessing Media Via an Audio-only Communication Channel: A Log Analysis. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–6.

[53] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–7.

[54] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2022. Understanding User Perceptions of Proactive Smart Speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2022), 28 pages.

[55] Ryen W. White. 2016. *Interactions with Search Systems*. Cambridge University Press.

[56] I-Chin Wu. 2011. Toward supporting information-seeking and retrieval activities based on evolving topic-needs. *Journal of Documentation* 67, 3 (2011), 525–561.

[57] Yutong Xie, Zhaoying Pan, Jinge Ma, Luo Jie, and Qiaozhu Mei. 2023. A prompt log analysis of text-to-image generation systems. In *Proceedings of the ACM Web Conference 2023*. 3892–3902.

[58] Hamed Zamani, Johanne R Trippas, Jeff Dalton, Filip Radlinski, et al. 2023. Conversational information seeking. *Foundations and Trends® in Information Retrieval* 17, 3-4 (2023), 244–456.

[59] Junte Zhang and Jaap Kamps. 2010. Search log analysis of user stereotypes, information seeking behavior, and contextual evaluation. In *Proceedings of the Third Symposium on Information Interaction in Context*. 245–254.

[60] Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Emine Yilmaz, Joemon M Jose, and Leif Azzopardi. 2013. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information Retrieval* 16 (2013), 267–305.