The Effects of Domain Expertise on a User's Conversational Search

Rohan Ray Sobha[1] , **Claudia Hauff**[1]
[1]Delft University of Technology

Abstract

This paper delved into the effects of domain expertise on a user's conversational search, because as the use and acceptance of voice assistants increase the need for conversational search agent that can accommodate to a human characteristic such as domain expertise. Accommodating to disadvantaged users of web search as earlier works showed that users with low literacy and low spatial visualization abilities are strongly affected in their searches compared to users who do not suffer from these impairments. Prior research into domain expertise demonstrated the influence it has had on the querying behavior of users in web search. They found that domain experts included more domain specific jargon in their messages, made longer queries and spent less time per search task. This paper examined these findings in a conversational search setting. Contrary to these findings, no significant relation between the domain expertise level and any of these results could be established. However, conducting the experiment to assess these findings has provided insight into how users respond to a conversational search study such as this one.

List of Terms

Conversational Search (CS)
Interactive Information Retrieval (IIR)
Web Search (WS)
Conversational Search Agent (CSA)
Domain Knowledge/Domain Expertise

1 Introduction

With the increase in use and acceptance of voice-assistants [38], the field of conversational search (CS) becomes more important. This branch of interactive information retrieval (IIR) is mostly concerned with the interaction of humans and search engines in a conversational setting. This means that the presentation of information needs to be adjusted accordingly to fit a conversation format. For example, if a user asks about a description of the Netherlands, the expected response should be limited to a summary of a couple of sentences and not citing the Wikipedia page [1] for the Netherlands.

Research into CS ranges from building a theoretical framework [27] to designing CS systems [33]. Most of this work is focused on the technical details and the inner workings of CS, whereas there has not been much research conducted describing any relationship between characteristics of humans (e.g. age, gender, level of expertise) and how this influences the interaction between humans and Conversational Search agents. Vtyurina et al. [35] delve into establishing the main expectations from a CS system, but from the system's point of view, rather than accounting for the user's needs. This paper will focus on level of expertise as the main human characteristic of the user. Since these works do not tackle relevant points necessary to introduce the impact domain expertise has on search in general, let us consider a more conventional approach to search: web search (WS).

Before elaborating on WS research, let us consider why studying the effects of domain expertise is useful and why applying this knowledge is essential in a CS environment. This requires looking at predictive searching. Concerning search engines, Google is known for its predictive searching ability [12]. Predictive search benefits from derived data to build a user profile without any input from the user. Gauging one's expertise level for a particular domain (i.e. subject) and distinguishing the difference between a novice and expert can enhance this predictive search ability by providing context. This is essential in a CS setting as the user assumes that a chat bot can understand the context in which they search for information [35].

It is important to regard CS, because WS cannot accommodate all users as it lacks the ability to adapt to some disabilities present in some users. For example, those users who may possess knowledge about their own domain, but are inhibited from using a search engine effectively. Two of these disabilities that CS can address are low literacy [11, 36] and low spatial visual awareness [8] as they have been demonstrated to affect these users' searches in a traditional WS environment.

In contrast to CS, WS research has been conducted on the influence of domain expertise on the interaction of users with WS systems. Firstly, Allen [1] describes the influence of

[1]https://en.wikipedia.org/wiki/Netherlands

knowledge level on the number of retrieved items in a WS. Secondly, Hsieh-Yee [16] outlines how search experience and domain knowledge affect search tactics between beginners and experts. Lastly, White et al. [37] examined the effects of domain expertise, but in the context of WS.

This paper will close the gap between CS and the effects of domain expertise in WS by reproducing some of the results from White et al. [37] with a slightly different methodology as will be illustrated in the third section. They found that domain experts create more technical queries (R1), create longer queries in terms of tokens and characters (R2) and generate more successful queries than novices (R3). Since R3 is quite subjective, this paper will assume that domain experts send more queries on average per unit of time.

Based on the findings R1-R3, the following research question will be formulated: **How does domain expertise affect user's conversational search queries?** Since this question cannot be answered using solely one metric, the results R1-R3 will be transformed into three sub-questions Q1-Q3 each answering to one metric:

Q1 Do domain experts create sophisticated queries with more technical jargon?

Q2 Do domain experts create longer queries in terms of the number of tokens and characters?

Q3 Do domain experts send more queries than novices per unit of time?

As the reader may have noticed by now, this paper relies on terminology used in IIR, which they may be unfamiliar with. To familiarize the reader, the second section will define the important concepts by delving into related works in conversational search and domain expertise. The third section will elaborate on the setup of the online lab study and how the participants of this study have been categorized into experts and novices for each of the domains. The domains and what these entail will be defined in this section too. The fourth section will serve as a reflection on ethical procedures taken during this study and will expand on the ability to reproduce this study. The fifth section will illustrate the results and how these were derived and calculated. The sixth section is dedicated towards discussing the results by highlighting any confounding factors and providing an explanation how these results came to be. The seventh and last section will draw a conclusion and describe possible future works.

2 Background

This section will define the relevant concepts for this paper and motivate their importance in this paper. Each subsection will highlight a key concept.

2.1 Conversational Search

As stated by Zhang et al. [40] CS is fairly new, but its concepts are rooted in some earlier works such as the designed MERIT system by Belkin et al. [2] which uses a "script-based conversational interaction for effective search" [40, p.178].

Radlinski and Craswell [27] also contributed to defining conversational search by constructing a theoretical framework. What sets these works apart, is their approach: where [2] used "case-based reasoning to select the next steps and offer users choices" [27, p.118], Radlinski and Craswell [27, p.118] assumed "a simpler conversational interface (such as a chat) where users enter text to agent actions also consisting of simple statements."

Hence, this paper uses this definition to describe CS: *CS is a type of search which embodies a simpler conversational interface where users enter text to agent actions also consisting of simple statements.*

Despite the novelty of CS, researchers have conceived expectations [35] for a CS system and designed one [33]. These expectations came with their share of design implications that Vtyurina et al. [35] recommended. Two of these will be highlighted here. One of them is context. Context is regarded as crucial as this can influence **Q2** and **Q3**. Context allows users to write shorter queries than they normally would and may result in either more or less queries, because context communicates to the user that a conversational search agent (CSA), a chat bot, understand the user. This could possibly make the user more interested to keep talking or the user may find their answer in less messages. The other is the use of feedback as too much feedback will overwhelm and discourage the user to use the system. However, too little feedback results in misunderstanding and frustration from the user.

From these recommendations a design can be created. Trippas et al. [33] designs a spoken conversational search system. In their paper they examine the way in which users write their queries and ways in which these can be intercepted.

Trippas et al. illustrates how one utterance (i.e. query) can have so-called multiple moves. These moves are clauses in the sentence of a user in which each clause has one specific aim. Whether that is to navigate, provide feedback or to inquire more information. By rewriting these multiple moves into one coherent query, it enables the system to understand the user's speech and obtain the correct information from the given source.

A CS can intercept utterances by incorporating Shannon's model of communication [29]. This entails that one needs to take any noise from the sender and receivers side into account when creating a CS system. The eventual goal is to create a system where the user and the CSA understand each other. However, the aim of this paper is not to create such a system, but rather to use such a system to measure the performance of experts and novices and their domain knowledge in a CS environment.

2.2 Domain Expertise

Allen [1] defined domain knowledge as knowledge of the subject area. Kiestra et al. [19] stated that there is variation between the amount of domain knowledge between end-user give a specific subject area. The amount of domain knowledge is referred to as the level of domain expertise in this paper and the subject area is called a domain in this paper. More specifically, the distinction is made between novices and expert and what constitutes as one or the other is clarified in section 3.2. Which domains have been chosen and the motivation is explained in section 3.1. Allen's definition will be used throughout this paper.

The motivation for including domain expertise as independent variable is demonstrated by Downing et al. [8]. They show results where this variable has been included and they discovered the existence of a correlation between domain expertise and the number of articles found. This correlation becomes more evident when a user possesses a higher spatial visualization ability. Hence, there is reason to assume domain expertise may result in distinguishable results between experts and novices.

3 Methodology

This section will first illustrate the domains for which the research questions will be answered. Secondly, classification of users into their expert and novice domain will be demonstrated. Thirdly, the collection of survey questions used to classify will be defined as well as the assigned search tasks. Fourthly, the setup of the chat bot will be elaborated upon. Lastly, the metrics for each research question will be defined.

3.1 Establishing domains

IIR articles have explored all kinds of domains. There even exists a comprehensive list of domains in the TREC benchmark campaign [30]. From all these domains, these three were selected:

1. Finance
2. Medicine
3. Computer Science

The motivation for picking these three domains is because these domains represent an variety of sciences. White et al. [37] also considers these fields are also beneficial to the general public. Choosing three domains was also mentioned in Kiestra et al. [19], but the authors do not provide any justification for the amount of domains.

White et al. [37] covered four domains in their setup: the aforementioned three and the *Legal* domain. This domain was left out intentionally as any search tasks about the Legal domain tend to be very specific per legal system and region. This would require pre-selecting participants familiar with English legal terms and common law as opposed to civil law.

Moreover, the three domains [17, 25, 5] generate large amounts of data and finding this information quicker without having to consult any corpuses is beneficial to users.

As will be explained in section 3.3 and 3.4, both the questions and search tasks were assembled bearing in mind that these were not centered around one specific topic within any of these domain. Naturally, neither the questions nor the tasks can cover the domain as whole and certain sub-groups are ignored as a result of a limitation to maintain overhead and feasibility.

3.2 Classifying novices and experts

Determining who is a domain expert and who is not, can be approached in many ways. These ways can be categorized into two main approaches. One approach is classifying participants by tracking their WS behavior [37]. The other approach that White et al. [37] incorporate is to assess users' domain knowledge by using URL-filters to classify users into either an expert or novice in each of the three domains. The scope and time frame of this study however cannot afford the luxury of using these techniques such as observing the user's behavior over a longer period of time.

Which of the following topics is unrelated to computer science?

- Designing Systems
- Technical Support
- Blockchain
- Big Endian - Little Endian
- I do not know

Figure 1: Example of survey question

Instead, this classification is determined by the outcome of a survey [2]. The motivation behind choosing these questions and their sources will be discussed in section 3.3. To minimize the effect of the difficulty of the questions, each participant was classified as an **expert** if they scored s one standard deviation above the mean for that specific domain (i.e. $s \geq \mu + \sigma$). If they score lower than this, they were classified as a **novice** (i.e. $s < \mu + \sigma$).

The approach is to set up a 3×2 factorial design as described by Kelly [18]. This refers to the three domains cited in section 3.1 and the expertise level (i.e. expert or novice) of each participant.

In this study, participants were treated as within-subjects, since all participants will fill in the same survey and undergo the same session with the chat bot as described in subsection 3.6.

3.3 Defining survey questions

As mentioned in section 3.2, survey questions were used to classify the participants into experts and novices. Some of these questions came from different sources varying from TED-ed [3, 7, 28, 6] to Kiplinger [20]. Others were recalled from the author's high school and university curriculum and his personal experience. Admittedly, the TED-ed questions are the most reputable as these TED-ed lessons are used for complementary education [34]. Questions from Kiplinger and the author are second and least reputable. In total, there are thirty questions, ten for each domain. When choosing these questions, it was ensured that not all questions for one domain were also about one topic, since these domains are quite broad. For reference, the full survey is provided in the appendix after the references.

3.4 Defining search tasks

As will be described in section 3.6, upon completion of the survey, the participants were asked to perform search tasks using the setup mentioned in section 3.5. The set of search tasks is available on the page starting with the header **Search Tasks**. These tasks were mostly gather using this website [3] and the respective sources for these are:

[2] https://forms.gle/CCJ44v2wY8rQYncJ6

[3] https://ils.unc.edu/searchtasks/index.html

[15, 13, 26, 22, 24, 4]. Some of these tasks were devised by the author. These tasks are trustworthy as they are able to elicit different queries, because none of them can be answered with one simple answer. However, it must be noted that finding a good answer or asking the right questions was not accounted for as this does not influence any of the metrics defined in **Q1**, **Q2** or **Q3**. In total, fifteen search tasks were composed, five for each domain.

3.5 Setting up the conversational search agent

To set up the conversational search agent (CSA), Macaw [39] was used. One of the lab members forked the repository [4] and deployed it inside a Docker [5] environment together with a MongoDB [6] installation.

Although Macaw has been used as a tool in this setup, it is essential to explain what it consist of in order to gain a deeper understanding of methodology as whole. Zamani and Craswell [39] describe Macaw as a Conversational Information Seeking system that centers around four main components. These components are illustrated in figure 2.

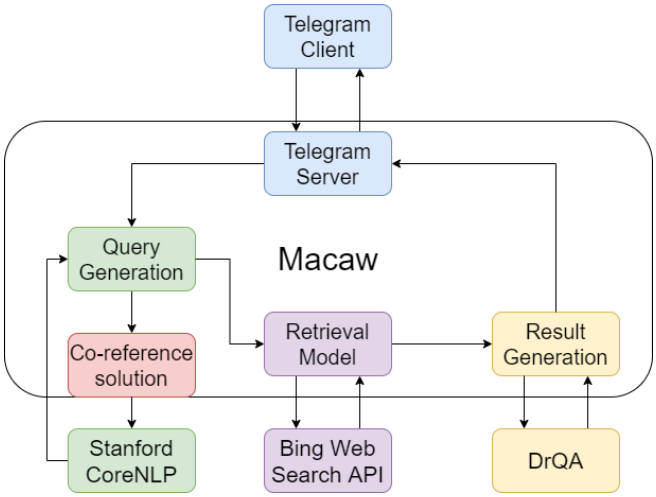


Figure 2: Layout of the CSA

The first component is Co-Reference Resolution. This refers to the fact that Macaw stores previous queries of the user to infer context from queries. This is driven by the Stanford CoreNLP [23]. The second component is Query Generation which parses the user's query and optionally expands it if co-reference resolution is enabled. In this research this component was enabled as a convenience towards our participants. The third component is the Retrieval Model. To allow users to make queries to the web effectively, the Bing Web Search API [7] was called to return search results from online. This API was chosen, because Macaw was already configured for this specific search engine. The last component is the Result Generation which bears responsibility for presenting documents and other search results in an appropriate manner to the user. Since question and answering is valued in Conversational Search research, the authors included the DrQA [8] model to generate concise answers and understand questions.

Macaw can be interacted with using not only the I/O, but also by creating a Telegram Bot [9] as it was included in the package. As this is more comfortable for people who are not used to a shell-like environment, this bot was configured and programmed.

3.6 Performing the experiment

Each participant was recruited from Prolific.[10] Participants were selected based on their educational background in Finance, Medicine and Computer Science. They were able to enter the experiment by clicking on a provided link within the Prolific environment that redirected them to the survey mentioned in section 3.3. Each of them were asked to create a Telegram account on the first page of the survey to be able to continue on to the second part of the experiment. After completing this survey, they were given a link with the set of instructions. They are found on the page with the header **Instructions**. The participants were only shown this page, so they did not know in advance which search tasks were available or which one they would be assigned to.

Once they found the chat bot inside their Telegram client of choice and they entered the **/start**, they were presented with a screen similar to figure 3. For each of the three domains, the participants were randomly assigned one exploratory search task. The python script threw a dice [11] between zero and four and assigned them a task based on the outcome. They were instructed to complete one search task, as defined in section 3.4, for one domain and proceed to the next domain by entering the **/next_domain** command as shown in figure 4.

As stated earlier, this research is mainly a within-subject research [18] in the sense that users complete search tasks in all domains, regardless of their level of expertise. However, as participants only complete one task per domain, they are not exposed to the other questions. When each participant had completed these tasks, they could enter the **/finish** command, as shown in figure 5 and they would receive a message to contact the researchers via Telegram to receive their completion code. Hence, each submission was manually checked to ensure that every participant completed the task successfully.

3.7 Delineating metrics

As mentioned in the introduction, this paper addresses one research question using three metrics. This subsection will outline these metrics and describe in detail what will be measured.

To answer **Q1**, several domain dictionaries need to be consulted to verify that queries contain specific jargon. To this

[4] https://github.com/roynirmal/macaw_docker
[5] https://www.docker.com/
[6] https://www.mongodb.com/
[7] https://docs.microsoft.com/en-us/azure/cognitive-services/bing-web-search/
[8] https://github.com/facebookresearch/DrQA
[9] https://core.telegram.org/bots/api
[10] https://www.prolific.co/
[11] Not a real dice, but Random.randrange()

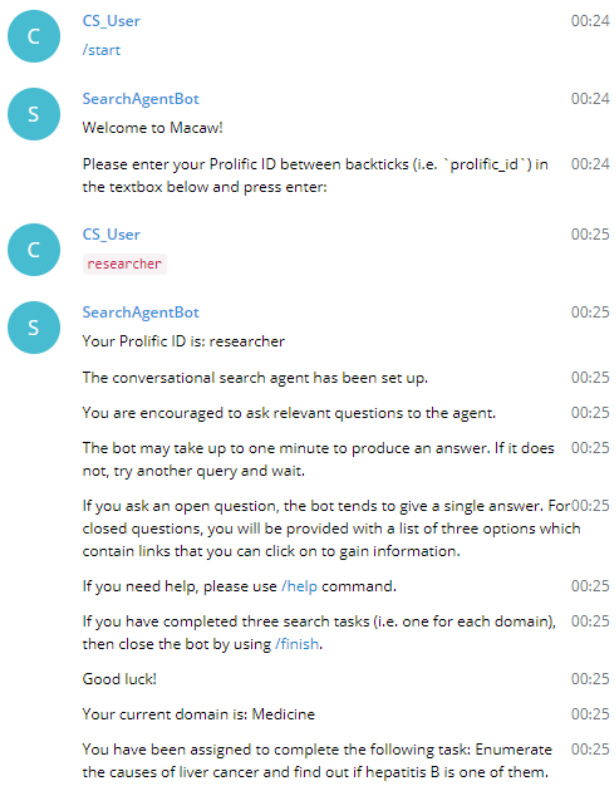


Figure 3: Starting the CSA

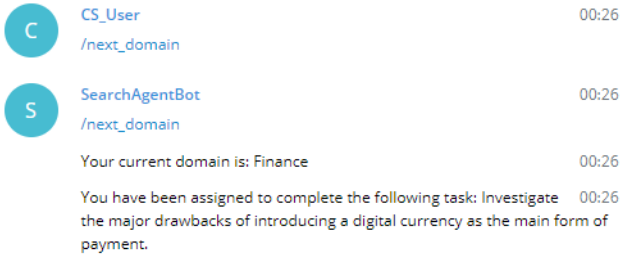


Figure 4: Switching to the next domain

end, these domain dictionaries are the same lexicons that White et al. [37] refer to in table 1. The messages of each user are stored in a database. Every message will be manually examined for the presence of domain specific lexicon in the aforementioned source. The listing below shows shows how this was stored. The ellipsis indicates that the line/object repeats itself, since there are three domains for each user.
To address **Q2**, a simple Python script can be constructed to aggregate several statistics regarding the length of each message. For example, one could calculate per user the mean and the variance of the lengths to derive properties about these data points to either support or reject the hypothesis.

[12]https://medlineplus.gov/encyclopedia.html
[13]https://www.anz.com.au/support/help/glossaries-terms/
[14]https://www.acm.org/publications/computing-classification-system/1998/ccs98

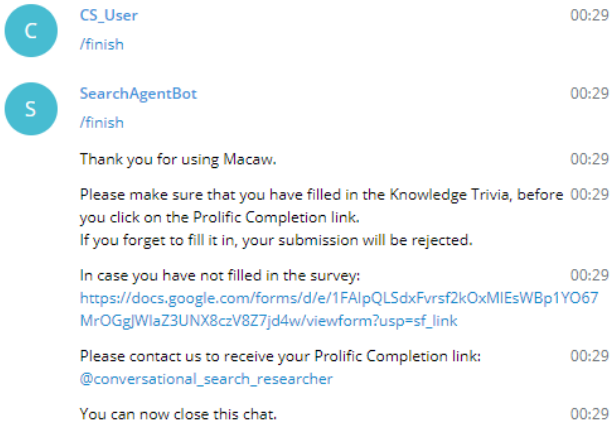


Figure 5: Finish the search session

```
user: {
    domain: int,
    ...
}
```

Listing 1: Storing jargon count

Furthermore, multiple statistical tests can be conducted on these data points. Most notably, ANOVA and pairwise t-testing are common [18] in the field of IIR. Due to the factorial setup described in section 3.2, this paper will opt for pairwise t-testing as will be further disclosed in detail in section 5. The listing below shows how this is stored for each user.

```
user: {
    domain: int[],
    ...
}
```

Listing 2: Storing length of message/domain/user

For **Q3** it is significant to measure the time it took for each participant to perform a search task. To calculate the time they spent, subtracting the timestamps between two messages is the key to reveal the duration of one's search task. Then the mean is calculated to display the average time a user spent on one search task and thus on one domain.

4 Responsible research

This section will discuss three integral parts of scientific research within Computer Science. Namely, research integrity, the three general perspectives on computer science and the feasibility with regards to reproducing this methodology. The first subsection deals with distinguishing right from wrong [10]. The second subsection attempts to frame this particular research within the three main branches of computer science. That especially determines the way this research should be considered and executed. The last subsection provides insight into the feasibility of reproducing the methodology used

Domain	Lexicon	Entries
Medicine	MedlinePlus Medical Encyclopedia [12]	3,535
Finance	ANZ Glossary of terms [13]	2,476
Computer Science	ACM Comp. Class. System (1998). [14]	1,361

Table 1: Domain Lexicons and number of entries
Authors: White et al. [37]

```
user: {
    domain: {
        nr_messages: int,
        session_length: float,
        time_between_messages: float,
        knowledge_level: str
    }, ...
}
```

Listing 3: Storing session time and time between messages

in this paper.

4.1 Research integrity

Research in Conversational Search supports hypotheses by means of facilitating an experiment. This paper is no exception. More specifically, dealing with research in general requires upholding *research integrity*. There are five principles prescribed by the Netherlands Code of Conduct [21] that form the basis of research integrity.

The first one deals with honesty. In this research, the methodology is accessible to anyone. This is achieved by publishing the code and data after the research period is over. The reason for this is that the code-base has case specific data hard coded which are privacy sensitive or prone to licensing issues with external APIs. It is essential that data is stripped away of any unique identifiers that can be traced back to the individuals, before its release.

The second principle revolves around exercising scrupulousness. This entails finding the method of executing the experiment given the limitations and the scope of one's research. As explained in the methodology, there were some different approaches suggested than those described here, because of a difference in the scope and (time) resources available. Nevertheless, an endeavor was made to carefully consider some of these methods first.

The third principle is all about being transparent. The source of ideas and methods need to be clearly referred to. Most of this work builds upon the results achieved by White et al. [37]. Their results serve as the hypotheses for this paper and is paid attention to.

The fourth, but not the last, principle stresses the importance of independence as a researcher. The choice of method and the level of scrutinizing data should not depend on the funding of external third-parties. This research has been solely funded by Delft University of Technology which avoids any conflict of interest.

The fifth and last principle concerns bearing responsibility. This includes cooperating with fellow research peers to discuss methods and analysis of data. Encouraging this behavior minimizes the odds of performing unethical research, because listening to another peer may aid oneself to reflect on seemingly small acts of misconduct in research. This encouragement was addressed by having weekly if not semi-weekly meetings to discuss progress and struggles that my peers and I faced.

4.2 Research in computer science

There are three views that provide different ways to think about research in computer science and are also applicable to IIR or Conversational Search. Ontology can be characterized as the field that invents new ways to classify information and attach names to concepts. Epistemology is often defined as the science of obtaining knowledge. Lastly the methodology describes the way in which a way in which this knowledge can be attained.

Moreover, there are three different kinds of research fields in computer science with each their own school of thought as Eden [9] states.

Firstly, there is the rationalist paradigm "which takes the discipline to be a branch of mathematics, the tenets of which have been common among scientists investigating various branches of theoretical computer science" - Eden [9, p.143].

Secondly, the technocratic paradigm is defined as "a branch of engineering, proponents of which dominate the various branches of software engineering, including software design, software architecture, software, maintenance and evolution, and software testing." - Eden [9, p.148]

Third and lastly, "the scientific paradigm contends that computer science is a branch of natural (empirical) science, on a par with 'astronomy, economics, and geology', the tenets of which are prevalent in various branches of AI, evolutionary programming, artificial neural networks, artificial life, robotics and modern formal methods." - Eden [9, p.154]

Of these three paradigms, the scientific paradigm acts as a perfect classifier for field of IIR and CS. Any paper of Trippas [31, 32] will illustrate the necessity of empiricism. In conversational search, one cannot build upon certain axioms and prove lemmas. To be able to provide an answer, data of participants needs to be gathered both qualitatively and quantitatively.

Furthermore, as mentioned in section 3.5, Macaw relies indirectly on the use of Machine Learning algorithms to return results. This means that it is not feasible, if not impossible, to logically deduce certain outcomes given these two facts. One could argue that one could predict every possibility within a certain probability, but would be rather unfruitful since there is no statistical significance.

4.3 Reproducibility

In section 3 the methodology was thoroughly discussed including an overview of other methods used in different papers. In order to reproduce the methods used, it is important that one has prior basic programming knowledge. Besides that, all the necessary repositories and instructions have been

provided [15] online.

As mentioned in section 3.5, a modified repository [16] was used to configure the chat bot using Macaw as its backbone.

However, weeks were spent trying to debug Macaw as it did not work with the provided instructions. Some bugs were in the core code and some key parameters were not documented properly, which caused delays and frustration as our peer group struggled to obtain basic functionality.

In addition, preferably one runs this software on a dedicated server with at *least* 4GB of RAM as the Macaw Telegram bot described in 3.6 would often become unresponsive if there was a lack of memory with no errors or warnings issued.

All in all, Macaw has potential in the field of Conversational Search, but some of the results shown in are not really conversational as it really depends on the data set and the supporting frameworks how Macaw performs. A full example of a conversation is provided on the last three pages before the references. The developers of Macaw offer a generous amount of freedom in terms of parameters which exactly made things the more confusing for a developer who is not proficient in that field.

5 Results

This section will present the results of this paper in the light of the three sub questions (i.e. **Q1**, **Q2** and **Q3**) that support the main research question of whether domain expertise affects conversational search. Each subsection will first explain how each metric was calculated.

Afterwards, the results will be presented using graphs and tables. The header row of table 3, 4 and 5 consists of: the metric and corresponding unit, the number of participants and each statistic (mean = μ, median, standard deviation = σ, sum = Σ).

Finally, each result will be subject to Welch's t-test and the corresponding t-statistic and p-value will be evaluated to the common scientific standard. This standard entails that the critical p-value is 0.05, so if the p-value is lower than this value, the null hypothesis is rejected. Otherwise, it remains which indicates that the data is inconclusive. The header row of table 2 is made up of: the corresponding sub question, t-statistic and lastly the p-value. An explanation for each of the results will be provided in next section, section 6.

5.1 More jargon

To answer **Q1**, the number of words belonging to technical jargon had to be counted per domain for each user as mentioned in listing 1. Moreover these results were then analyzed to assess whether a correlation between the expertise level and the presence of jargon exists. The sources for identifying jargon are mentioned table 1. As illustrated in figure 6 novices mentioned more domain jargon than experts in all domains.

However applying the t-test on **Q1**, shows that no $p \not\leq 0.05$ for any domain as illustrated in table 2. This implies that the null-hypothesis - *expertise level does not indicate a higher usage of jargon in queries* - cannot be rejected.

[15] https://github.com/microsoft/macaw

[16] https://github.com/roynirmal/macaw_docker

5.2 More characters

Answering **Q2** calls for experts writing longer queries on average per message. Calculating this metric given the format provided in listing 2 is feasible as the length of each message can be expressed in the number of characters. Plotting the average query length per user for each domain and expertise level results in a graph illustrated in figure 7 and table 4. The graph demonstrates no strong indication that novices or experts make longer queries. Applying the t-test indeed shows that $p > 0.05$ for every domain illustrated in table 2. Thus, the null-hypothesis - *expertise level exerts no influence on the query length* - cannot be rejected.

5.3 More time

To answer **Q3**, the time between each message was logged as shown in listing 3. Afterwards, the average time users took between messages was compared between the novices and experts as depicted in table 5 and figure 8. The associated bar graph in figure 8 does not exhibit any consistent trend. Applying the t-test confirms this as shown in table 2. This indicates the possibility that the expertise level and time spent on a task may be unrelated. Hence, the null-hypothesis - *expertise level is not an indicator of the time spent* - cannot be rejected.

Q1	t-statistic	p-value
Medicine	0.64	0.54
Finance	-0.16	0.88
Computer Science	0.48	0.63
Q2	t-statistic	p-value
Medicine	1.12	0.28
Finance	-0.04	0.96
Computer Science	0.59	0.56
Q3	t-statistic	p-value
Medicine	0.55	0.60
Finance	-1.26	0.22
Computer Science	1.01	0.38

Table 2: p-value and t-statistic for **Q1**, **Q2** and **Q3**

No. words	experts	μ	median	σ	Σ
Medicine	6	3.83	2.0	5.60	23
Finance	4	1.75	1.0	2.22	7
Comp. Sci.	4	2.50	2.5	0.58	10

(a) Experts

No. words	novices	μ	median	σ	Σ
Medicine	19	2.32	2.0	2.11	44
Finance	21	1.95	2.0	2.42	41
Comp. Sci.	21	2.24	2.0	2.10	47

(b) Novices

No. words	all	μ	median	σ	Σ
Medicine	25	2.68	2.0	3.21	67
Finance	25	1.92	1.0	2.34	48
Comp. Sci.	25	2.28	2.0	1.93	57

(c) All participants

Table 3: **Q1**: Presence of domain specific lexicon

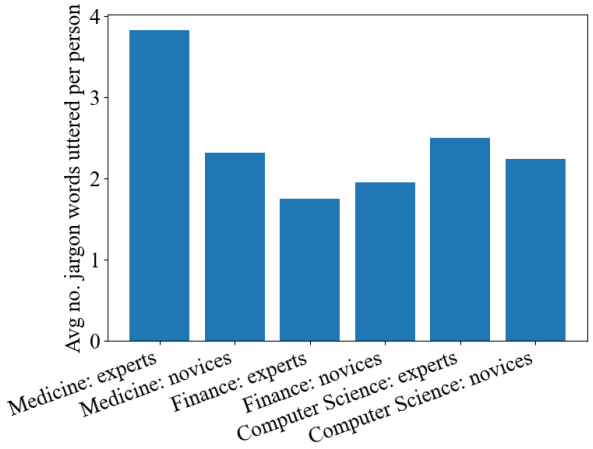


Figure 6: **Q1**: Presence of domain specific vocabulary

No. characters per message	messages	μ	median	σ
Medicine	19	61	49.0	75.00
Finance	7	69	51.0	56.77
Comp. Sci.	8	89	78.0	67.77

(a) Experts

No. characters per message	messages	μ	median	σ
Medicine	60	41	39.0	28
Finance	39	71	54.0	88
Comp. Sci.	46	71	47.5	129

(b) Novices

No. characters per message	messages	μ	Med	σ
Medicine	79	45.61	39	44.50
Finance	46	70.35	53.5	83.26
Comp. Sci.	54	73.93	50	121.17

(c) All participants

Table 4: **Q2**: Length per message

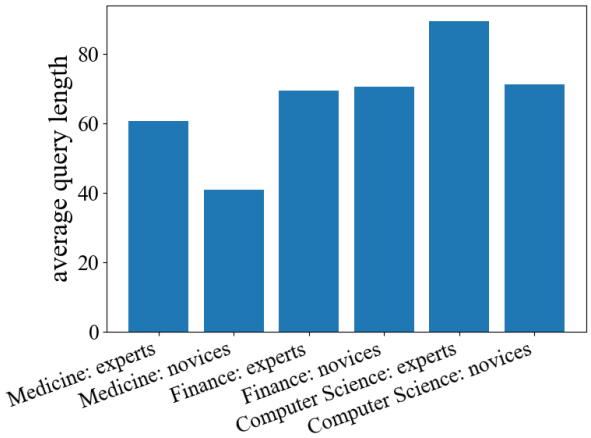


Figure 7: **Q2**: Average query length

6 Discussion

This section will discuss not only the results, but also reflect on the methods used to achieve these results and on the paper's position within other papers of domain expertise. For example, examining possible confounding factors and issues that occurred during the experimental setup. These refer to subsections 3.2 and 3.6 respectively. Firstly, this section will discuss the results with respect to each metric. Secondly, it will elaborate on how the participants were chosen. Thirdly, an argument will be provided explaining why sanitizing the data was essential. Lastly, the position of this paper will be assessed with respect to other papers that discussed domain expertise.

6.1 Writing more jargon

The aforementioned trend mentioned in 5.1 may seem surprising under the assumption that domain experts are more familiar with domain specific vocabulary. The metric used to answer **Q1** merely refers to the number of times a specific word appeared in the queries of users. The distinct set of vocabulary was not taken into account.

As a consequence, users could repeat the same question with the same terminology over and over and this was registered each time. Upon closer examination of the message logs, it is remarkable that experts tend to have a larger command of their domain's vocabulary, but are less likely to repeat the same specific words. The learning effect that came with asking questions was also disregarded, because most users only asked one or two questions to the chat bot as will be elaborated upon in section 6.2.

6.2 Writing longer queries

The results presented in figure 7 are skewed. Among the participants there were some people who engaged in a conver-

Time elapsed between messages in s	experts	μ	median	σ
Medicine	6	736	288	1077
Finance	4	125	63	167
Computer Science	4	443	362	454

(a) Experts

Time elapsed between messages in s	novices	μ	median	σ
Medicine	19	479	190	655
Finance	21	308	81	546
Computer Science	21	210	184	201

(b) Novices

Time elapsed between messages in s	all	μ	median	σ
Medicine	25	541	191	759
Finance	25	279	81	507
Computer Science	25	247	184	259

(c) All participants

Table 5: **Q3**: Time elapsed between messages

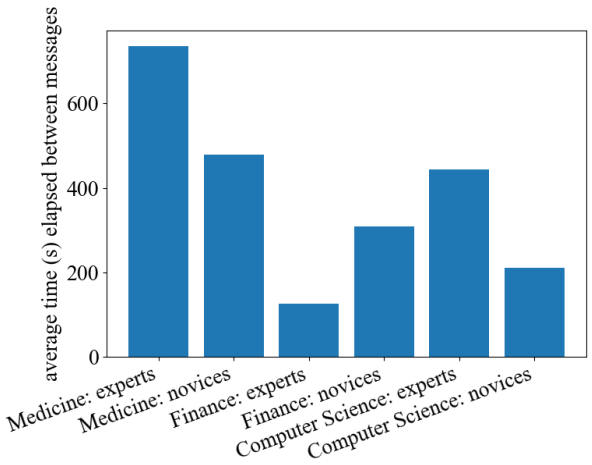


Figure 8: **Q3**: Average time elapsed between messages

sation with the Macaw chat bot, because of the *possibly misleading* name of Conversational Search.

Whenever such a user sent a message to the Macaw bot, it was not internally programmed to handle such conversation. Moreover, recordings of people telling the bot what they learned in a message or two are the cause of some of the long messages.

The perhaps unjustified assumption was made that people would **always** ask questions to the bot, because the instructions specified so. 'You are encouraged to ask questions' was misinterpreted by many, because of its friendly tone. Some considered it as a mere invitation to interact rather than a format specified in which the chat bot could process data

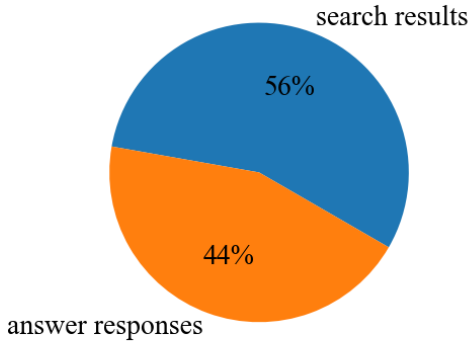


Figure 9: Macaw responding with search results or answers

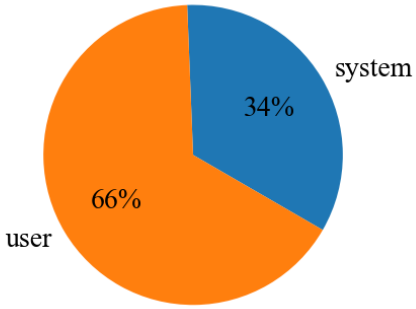


Figure 10: Distribution of messages between user and system

easier.

6.3 Sending more queries

The users were free to click on a link and browse the web page to obtain info for as long as they wanted. A possible confounding factor could have been the retention time of the suggested web pages by the chat bot. This retention time is the time users were reading web pages instead of searching for other pages containing information.

Moreover, figure 9 shows that 44% of the responses from the bot were single line responses and the majority were search results presented to the user. The representation of an answer to a user's query may also affect the likeliness of a user sending another query. For example, if the answers are clear and useful to the user, then they may ask less or different questions related to a different topic. On the other hand, unclear answers usually result in users repeating the same words time and time again.

Furthermore, figure 10 demonstrates that approximately 2

out of 3 messages were sent by the user and that the other messages were sent as replies from the bot. During the experiment, the bot froze once and needed to be restarted. As a result, some participants had to repeat this part of the experiment and thus were able to send more queries. It was considered to omit these messages, since they were not part of a complete submission of a user. However, since the metrics are concerned with the messages themselves and the data is not bound to any context, it was decided to keep this data in the published results.

6.4 Classifying participants

The difficulty of the survey questions was not of equal level for each domain. It suffered from bias from the researcher's side as a BSc Computer Science Engineering student tends to know more. As there was no ready-made survey covering these three domains, search task databases and several websites were consulted to collect a set of appropriate questions. Admittedly, some of these questions were made up, because to keep consistency every domain needed to have ten questions.

Furthermore, the sample size was too small to derive statistically significant properties, because trends were observed. Increasing the sample size may lead to finding the hypothesized results. As it stands now, we cannot reject the null-hypothesis for any of the three research questions.

6.5 Sanitizing data

Mentioning section 4.3, the chat bot froze on test day. A quick restart fixed most problems, but since the Telegram chat bot was mostly state-based it remembered the state of everyone who used it. Unbeknownst it had skipped a domain, because the command **/next_domain** was sent twice. These lines were manually removed from the document as they are not part of the queries that a user can send.

Despite the instructions to enter the full Prolific ID upon starting the bot, some people only filled in half their ID. This made it difficult to query their data afterwards even though they completed their experiment. Using regular expressions and parts that they did fill in, retrieving the documents from participants became possible again.

In another case, someone pressed **send** too early and send a **?** afterwards. This is unfortunately registered as two messages of which several statistics will be automatically calculated.

All in all, this does not imply that the raw privacy protected data will be concealed. This data will be published in the near future. However, sanitizing was crucial to be able to process the data fairly.

6.6 Domain expertise

To address the elephant in the room, the research question and the corresponding metrics were heavily inspired by the work of White et al.. They found that for every message containing words from domain lexicons experts generate 50% more than novices. This refers to **Q1** and remains inconclusive in this paper, but the research of White et al. [37] was conducted in a traditional WS setting. In another WS setting, Hembrooke et al. [14] found domain novices created shorter and less complex queries compared to their expert counterparts.

Due to some confounding effects as mentioned earlier this is inconclusive. All in all, the findings of these papers could not be reproduced in a conversational search setting.

7 Conclusions and Future Work

This paper's aim was to provide an answer to how domain expertise affects user's conversational search queries. Due to the reasons explained in the discussion, none of the results for the given sub questions **Q1**, **Q2** and **Q3** are statistically significant. As explained in the last subsection (6.6) of the discussion, other works in the field of domain expertise were not in line with the results from this paper.

All things considered, domain expertise may affect user's conversational search queries, but not according to the metrics defined in the introduction. This means that looking into the relation between domain expertise and CS remains fruitful as it can help users in the long run to find information easier.

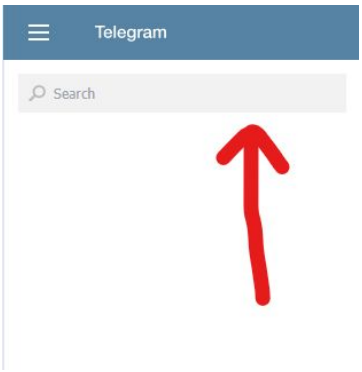
As for future work, the difference between domain expertise and search expertise could be considered in the context of CS. Search expertise is revolved around how knowledge and experience with a particular way of searching influences search measurement. In CS this means that users who use voice assistants daily could be considered experts in CS from a search expertise perspective, since they know how to formulate their queries implicitly regardless of their domain knowledge.

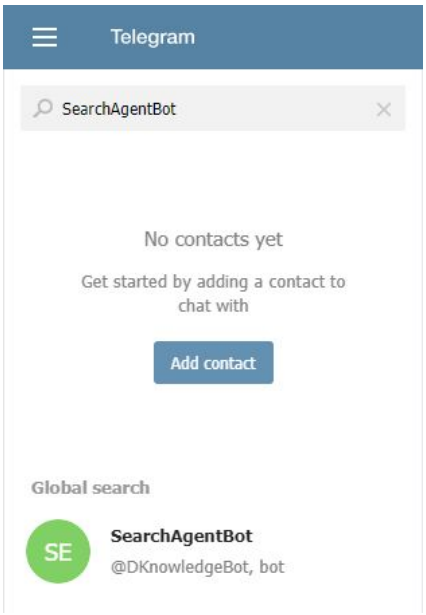
To conclude, researching how to cater to certain user groups, such as age, gender, cultural background, visual impairments, domain expertise and search expertise among others, in CS studies would benefit this field as it will increase awareness and encourage the adoption of this promising searching paradigm.

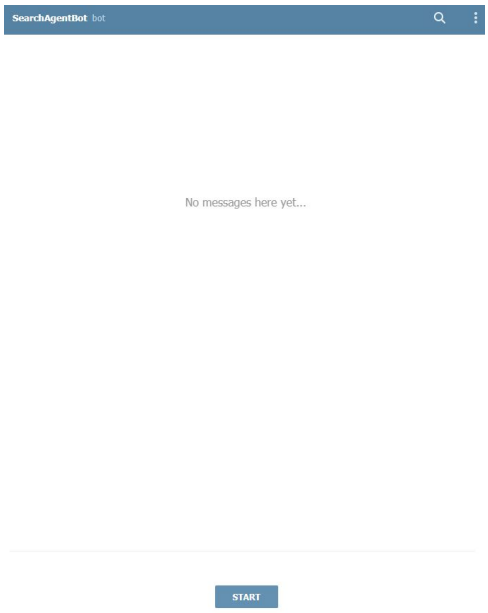
Instructions

There are three domains: medicine, finance and computer science

1. Please open your Telegram client (Desktop app/Mobile app/etc.) or web.telegram.org
2. Type in the search bar SearchAgentBot (or: @DKnowledgeBot).







3. Type **/start** to start the bot and configure the settings below:
4. Type your Prolific ID between backticks. Like this: `prolific_id`
5. You will be automatically assigned a domain with a corresponding task.
6. Start chatting with the bot. Open questions often elicit a single response. Whereas closed questions elicit a list of three links on which you can click to gain more information.
7. Please wait up to one minute if the bot seems unresponsive. If you do not receive an answer, please try another query.
    a. If that does not help, contact us on Telegram: @conversational_search_researcher
    b. Alternatively, send an email over Prolific.
8. When you are finished with your chosen search task of a domain of your choice, please type **/next_domain**.
9. You will then move on to the next domain and a task will be assigned to you. Go to step 5, until you have visited three domains.
10. *If you queried one task per domain (i.e. you completed three search tasks),* please type **/finish** to end your session.
11. If you have followed these instructions:
    a. Your completion code for Prolific has been given to you in one of the messages after using the /finish command.
    b. Contact us on Telegram: @conversational_search_researcher
    c. Or send an email to: 5ebbf521e21fc001c14b28a8@email.prolific.co

Search Tasks

Medical: (https://ils.unc.edu/searchtasks/search_show_task.php?aid=510&no=2)
(https://ils.unc.edu/searchtasks/search_show_task.php?aid=31&no=1)

1. Find out about the influence of long-distance running on intervertebral disc narrowing in men.

2. Find several natural family planning methods and look up on which one is the best.

3. Research the correlation between smoking pregnant women and the possibility of having smaller babies.

4. Enumerate the causes of liver cancer and find out if hepatitis B is one of them.

5. Inquire how spit (sputum crytoloty) can be used to possibly identify lung cancer.

Finance:
(http://www.annualreviews.org/doi/10.1146/annurev.soc.29.010202.100037)
(http://www.tandfonline.com/doi/full/10.1080/13662719400000002)

1. Inquire whether it is wiser to buy financial instruments (e.g. bond, stocks, obligations) or sell them (imagine you already have them) in times of crisis.

2. Research the influence of unemployment on the value of consumer goods.

3. Investigate the major drawbacks of introducing a digital currency as the main form of payment.

4. Analyze the effects that a learning- and knowledge economy brings to a nation's prosperity.

5. Investigate how a circular economy can lead to sustainable business models.

Computer Science:
(https://link-springer-com.tudelft.idm.oclc.org/article/10.1007/s10677-016-9745-2)
(https://ils.unc.edu/searchtasks/search_show_task.php?aid=742&no=1)

1. Research how Blockchain can keep our data safe and decentralized (not all data in one place).

2. Investigate the relation between use of voice assistants and the accuracy of queries.

3. Find whether all calculable problems can be solved (decided) given enough time.

4. Investigate the issues that come with unsupervised learning.

5. Research what ethical principles are applied when developing software used in (semi-mechanical) machines (e.g. robots, self-driving cars, vending machines).

C CS_User 00:24

/start

S SearchAgentBot 00:24

Welcome to Macaw!

Please enter your Prolific ID between backticks (i.e. `prolific_id`) in the textbox below and press enter: 00:24

C CS_User 00:25

`researcher`

S SearchAgentBot 00:25

Your Prolific ID is: researcher

The conversational search agent has been set up. 00:25

You are encouraged to ask relevant questions to the agent. 00:25

The bot may take up to one minute to produce an answer. If it does not, try another query and wait. 00:25

If you ask an open question, the bot tends to give a single answer. For closed questions, you will be provided with a list of three options which contain links that you can click on to gain information. 00:25

If you need help, please use /help command. 00:25

If you have completed three search tasks (i.e. one for each domain), then close the bot by using /finish. 00:25

Good luck! 00:25

Your current domain is: Medicine 00:25

You have been assigned to complete the following task: Enumerate the causes of liver cancer and find out if hepatitis B is one of them. 00:25

C CS_User 00:26

What causes liver cancer?

S SearchAgentBot 00:26

Liver Cancer Risk Factors), exactly how these may lead normal liver cells

C CS_User 00:26

How can hepatitis B cause liver cancer?

S SearchAgentBot 00:26

cancerous tumors

C CS_User 00:26

/next_domain

S SearchAgentBot 00:26

/next_domain

Your current domain is: Finance 00:26

You have been assigned to complete the following task: Investigate the major drawbacks of introducing a digital currency as the main form of payment. 00:26

C CS_User 00:26

What is a digital currency?

S SearchAgentBot 00:26

Digital Currency
What Is Digital Currency

C CS_User 00:27

How can it be used?

S SearchAgentBot 00:27

as an alternative to money

C CS_User 00:27

Where can digital currency be used?

S SearchAgentBot 00:27

England, Sweden

C CS_User 00:27

/next_domain

S SearchAgentBot 00:27

/next_domain

Your current domain is: Computer Science 00:27

You have been assigned to complete the following task: Find whether all calculable problems can be solved (decided) given enough time. 00:27

C CS_User 00:28

Can all calculable problems be solved?

S SearchAgentBot 00:28

Retrieved document list (click to see the document content):

C

CS_User 00:28

What are calculable problems?

S

SearchAgentBot 00:28

the Cambridge English Corpus and from sources on the web

C

CS_User 00:28

When can problems be decided?

S

SearchAgentBot 00:28

in polynomial time

C

CS_User 00:29

What is an example of a calculable problem?

S

SearchAgentBot 00:29

Cambridge English Corpus

C

CS_User 00:29

/finish

S

SearchAgentBot 00:29

/finish

Thank you for using Macaw. 00:29

Please make sure that you have filled in the Knowledge Trivia, before you click on the Prolific Completion link. 00:29
If you forget to fill it in, your submission will be rejected.

In case you have not filled in the survey: 00:29
https://docs.google.com/forms/d/e/1FAIpQLSdxFvrsf2kOxMIEsWBp1YO67MrOGgJWIaZ3UNX8czV8Z7jd4w/viewform?usp=sf_link

Please contact us to receive your Prolific Completion link: 00:29
@conversational_search_researcher

You can now close this chat. 00:29

References

[1] B. Allen. Topic Knowledge and Online Catalog Search Formulation. *The Library Quarterly*, 61(2):188–213, Apr. 1991. ISSN 0024-2519. doi: 10.1086/602333. URL https://www.journals.uchicago.edu/doi/abs/10.1086/602333. Publisher: The University of Chicago Press.

[2] N. J. Belkin, C. Cool, A. Stein, and U. Thiel. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9(3):379–395, Jan. 1995. ISSN 0957-4174. doi: 10.1016/0957-4174(95)00011-W. URL http://www.sciencedirect.com/science/article/pii/095741749500011W.

[3] E. Bryce. How do the lungs work?, May 2020. URL https://ed.ted.com/lessons/what-do-the-lungs-do-emma-bryce. Library Catalog: ed.ted.com.

[4] C. Carpineto and G. Romano. A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval. *Machine Learning*, 24(2):95–122, Aug. 1996. ISSN 1573-0565. doi: 10.1023/A:1018050230279. URL https://doi.org/10.1023/A:1018050230279.

[5] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang. Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. *IEEE Access*, 5:8869–8879, 2017. ISSN 2169-3536. doi: 10.1109/ACCESS.2017.2694446. Conference Name: IEEE Access.

[6] R. Coffin. What causes an economic recession?, May 2020. URL https://ed.ted.com/lessons/what-causes-an-economic-recession-richard-coffin. Library Catalog: ed.ted.com.

[7] M. Davis. What causes opioid addiction, and why is it so tough to combat?, May 2020. URL https://ed.ted.com/lessons/what-causes-opioid-addiction-and-why-is-it-so-tough-to-combat-mike-davis.

[8] R. E. Downing, J. L. Moore, and S. W. Brown. The effects and interaction of spatial visualization and domain expertise on information seeking. *Computers in Human Behavior*, 21(2):195–209, Mar. 2005. ISSN 0747-5632. doi: 10.1016/j.chb.2004.03.040. URL http://www.sciencedirect.com/science/article/pii/S0747563204000913.

[9] A. H. Eden. Three Paradigms of Computer Science. *Minds and Machines*, 17(2):135–167, July 2007. ISSN 1572-8641. doi: 10.1007/s11023-007-9060-8. URL https://doi.org/10.1007/s11023-007-9060-8.

[10] J. Fieser. Ethics | Internet Encyclopedia of Philosophy, May 2020. URL https://www.iep.utm.edu/ethics/.

[11] A. C. Graesser, H. Li, and C. Forsyth. Learning by Communicating in Natural Language With Conversational Agents. *Current Directions in Psychological Science*, 23(5):374–380, Oct. 2014. ISSN 0963-7214. doi: 10.1177/0963721414540680. URL https://doi.org/10.1177/0963721414540680. Publisher: SAGE Publications Inc.

[12] O. Hansson, D. Black, J. M. Wiley, M. Tungare, Z. Mahkovec, B. J. McMahan, B. A. Gomes, J. J. Effrat, J. R. Wright, and M. K. Wichary. Predictive query completion and predictive search results, Apr. 2014. URL https://patents.google.com/patent/US8706750B2/en. Library Catalog: Google Patents.

[13] R. B. Haynes, K. A. McKIBBON, C. J. Walker, J. Mousseau, L. M. Baker, D. Fitzgerald, G. Guyatt, and G. R. Norman. Computer Searching of the Medical Literature. *Annals of Internal Medicine*, 103(5):812–816, Nov. 1985. ISSN 0003-4819. doi: 10.7326/0003-4819-103-5-812. URL https://www.acpjournals.org/doi/10.7326/0003-4819-103-5-812. Publisher: American College of Physicians.

[14] H. A. Hembrooke, L. A. Granka, G. K. Gay, and E. D. Liddy. The effects of expertise and feedback on search term selection and subsequent learning. *Journal of the American Society for Information Science and Technology*, 56(8):861–871, 2005. ISSN 1532-2890. doi: 10.1002/asi.20180. URL http://www.onlinelibrary.wiley.com/doi/abs/10.1002/asi.20180. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20180.

[15] W. Hersh. TREC 2002 interactive track report. In *The Eleventh Text REtrieval Conference, TREC 2002*, pages 283–293, 2002.

[16] I. Hsieh-Yee. Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science*, 44(3):161–174, 1993. ISSN 1097-4571. doi: 10.1002/(SICI)1097-4571(199304)44:3<161::AID-ASI5>3.0.CO;2-8.

[17] A. Joorabchi, M. English, and A. E. Mahdi. Text mining stackoverflow. *Journal of Enterprise Information Management*, Mar. 2016. ISSN 1741-0398. doi: 10.1108/JEIM-11-2014-0109. URL https://www-emerald-com.tudelft.idm.oclc.org/insight/content/doi/10.1108/JEIM-11-2014-0109/full/html. Publisher: Emerald Group Publishing Limited.

[18] D. Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends® in Information Retrieval*, 3(1):1–224, 2009. ISSN 1554-0669. doi: 10.1561/1500000012. URL http://dx.doi.org/10.1561/1500000012.

[19] Kiestra, M. Stokmans, and J. Kamphuis. End-users searching the online catalogue: the influence of domain and system knowledge on search patterns. *The Electronic Library*, 12(6):335–343, Jan. 1994. ISSN 0264-0473. doi: 10.1108/eb045321. URL https://doi.org/10.1108/eb045321. Publisher: MCB UP Ltd.

[20] E. Kiplinger. Test Your Financial Fluency, Jan. 2015. URL http://www.kiplinger.com/quiz/saving/T065-S001-test-your-financial-fluency/index.html. Library Catalog: www.kiplinger.com.

[21] KNAW, NFU, NWO, TO2-Federatie, V. Hogescholen, and VSNU. Nederlandse gedragscode wetenschap-

pelijke integriteit, 2018. URL https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:110600. Medium: application/pdf type: dataset.

[22] B.-a. Lundvall and B. Johnson. The Learning Economy. *Journal of Industry Studies*, 1(2):23–42, Nov. 1994. ISSN 1320-6095. doi: 10.1080/13662719400000002. URL http://www.tandfonline.com/doi/full/10.1080/13662719400000002.

[23] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-5010. URL https://www.aclweb.org/anthology/P14-5010.

[24] S. Nyholm and J. Smids. The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice*, 19 (5):1275–1289, Nov. 2016. ISSN 1572-8447. doi: 10.1007/s10677-016-9745-2. URL https://doi.org/10.1007/s10677-016-9745-2.

[25] J. Patel, S. Shah, P. Thakkar, and K. Kotecha. Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, 42(4): 2162–2172, Mar. 2015. ISSN 0957-4174. doi: 10.1016/j.eswa.2014.10.031. URL http://www.sciencedirect.com/science/article/pii/S0957417414006551.

[26] W. W. Powell and K. Snellman. The Knowledge Economy. *Annual Review of Sociology*, 30(1):199–220, Aug. 2004. ISSN 0360-0572, 1545-2115. doi: 10.1146/annurev.soc.29.010202.100037. URL http://www.annualreviews.org/doi/10.1146/annurev.soc.29.010202.100037.

[27] F. Radlinski and N. Craswell. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 117–126, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 978-1-4503-4677-1. doi: 10.1145/3020165.3020183. URL https://doi-org.tudelft.idm.oclc.org/10.1145/3020165.3020183.

[28] A. Rosenthal and P. Thordarson. Which is better: Soap or hand sanitizer?, May 2020. URL https://ed.ted.com/lessons/which-is-better-soap-or-hand-sanitizer-alex-rosenthal-and-pall-thordarson.

[29] C. E. Shannon. *The Mathematical Theory of Communication, by CE Shannon (and Recent Contributions to the Mathematical Theory of Communication), W. Weaver*. University of illinois Press, 1949.

[30] TREC. Text REtrieval Conference (TREC) Data, Mar. 2020. URL https://trec.nist.gov/data.html.

[31] J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. A conversational search transcription protocol and analysis. In *Proceedings of SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17), CAIR*, 2017.

[32] J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. How Do People Interact in Conversational Speech-Only Search Tasks: A Preliminary Analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 325–328, New York, NY, USA, 2017. ACM. doi: 10.1145/3020165.3022144.

[33] J. R. Trippas, D. Spina, L. Cavedon, H. Joho, and M. Sanderson. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 32–41, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-4925-3. doi: 10.1145/3176349.3176387. URL https://doi-org.tudelft.idm.oclc.org/10.1145/3176349.3176387.

[34] G. Villias. TED-Ed lessons & TED-Ed clubs: Educational activities to amplify students' voices. 19:19252, Apr. 2017. URL http://adsabs.harvard.edu/abs/2017EGUGA..19192V. Conference Name: EGU General Assembly Conference Abstracts.

[35] A. Vtyurina, D. Savenkov, E. Agichtein, and C. L. A. Clarke. Exploring Conversational Search With Humans, Assistants, and Wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pages 2187–2193, Denver, Colorado, USA, May 2017. Association for Computing Machinery. ISBN 978-1-4503-4656-6. doi: 10.1145/3027063.3053175. URL http://doi.org/10.1145/3027063.3053175.

[36] W. M. Watanabe, A. C. Junior, V. R. Uzêda, R. P. d. M. Fortes, T. A. S. Pardo, and S. M. Aluísio. Facilita: reading assistance for low-literacy readers. In *Proceedings of the 27th ACM international conference on Design of communication*, SIGDOC '09, pages 29–36, Bloomington, Indiana, USA, Oct. 2009. Association for Computing Machinery. ISBN 978-1-60558-559-8. doi: 10.1145/1621995.1622002. URL http://doi.org/10.1145/1621995.1622002.

[37] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 132–141, Barcelona, Spain, Feb. 2009. Association for Computing Machinery. ISBN 978-1-60558-390-7. doi: 10.1145/1498759.1498819. URL http://doi.org/10.1145/1498759.1498819.

[38] J. R. Williamson. *User experience, performance, and social acceptability: usable multimodal mobile interaction*. PhD, University of Glasgow, 2012. URL https://eleanor.lib.gla.ac.uk/record=b2922742.

[39] H. Zamani and N. Craswell. Macaw: An Extensible Conversational Information Seeking Platform. *arXiv:1912.08904 [cs]*, Dec. 2019. URL http://arxiv.org/abs/1912.08904. arXiv: 1912.08904.

[40] Y. Zhang, X. Chen, Q. Ai, L. Yang, and W. B. Croft. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 177–186, Torino Italy, Oct. 2018. ACM. ISBN 978-1-4503-6014-2. doi: 10 . 1145 / 3269206 . 3271776. URL https : / / dl . acm . org/doi/10.1145/3269206.3271776.

Influence of Domain Knowledge on Conversational Search

This study consists of two parts:
I. Knowledge Trivia
II. Interacting With a Conversational Search Agent

It takes approximately 20 minutes to finish this study completely.

Pre-requisite: A Telegram account
- An Android smartphone with Android 4.1 or higher/An iPhone with iOS 9.0 or higher
- Download the Android app: https://play.google.com/store/apps/details?id=org.telegram.messenger
- Download the iOS app: https://apps.apple.com/app/telegram-messenger/id686449807
- Open the app and follow the provided steps to create your account.

To contact the researchers, if you have any questions:
In Telegram, search for the username: @conversational_search_researcher
Alternatively, send an email to: 5ebbf521e21fc001c14b28a8@email.prolific.co

Part I will assess your knowledge in three domains which will be elaborated on later. In this part, one fills out the form and answers the given questions.
Afterwards, you move on to part II where you will be interacting with a chatbot.

YOU NEED TO COMPLETE BOTH PARTS TO HAVE YOUR SUBMISSION APPROVED.

If you need help, feel free to contact us. Please read each section carefully.
*Required

1. Please fill in your Prolific ID *

Withdrawing from this study can be done in several ways:

1. Refusing to provide consent.

2. By sending us an email in Prolific stating that you no longer wish to participate.

2. I agree that I have read the following ways to withdraw from this study: *

Tick all that apply.

- [ ] I agree

Consent

This section is to provide consent for us to process your data.

We store the following:

- Your prolific ID
- Your answers to the Trivia quiz
- Your chat session with the bot (i.e the texts you type and the replies from the bot)

We do not store:
- Your name
- Your phone number

3. Do you provide consent to have us process your data anonymously for research purposes? *

Mark only one oval.

- ◯ Yes, I hereby provide consent.
- ◯ No, I do not provide consent. *Skip to section 9 (Refusal to Provide Consent)*

Start of Knowledge Trivia

In this trivia, you will be presented with three categories: Medicine, Finance and Computer Science.

Your payment does not depend on the number of correct answers given so please do not look up the answers to the questions in this trivia and only answer the questions based on your own knowledge.

Medicine

This section contains questions in the field of medicine. These questions do not represent the whole field of medicine. These questions were selected with the intent to be understood by people who understand basic anatomy and easy medicinal terminology. If you do not know an answer, please fill in 'I do not know', rather than guessing. As mentioned earlier: your payment does not depend on the number of correct answers given, so only answer the questions using your own knowledge.

4. What is an Epipen? *

Mark only one oval.

- A device to inject adrenaline to mitigate an anaphylactic reaction.
- A luxury pen that is used to write in space.
- A syringe to inject heroin.
- I do not know

5. How many pairs of chromosomes does a healthy person have? *

Mark only one oval.

- 20
- 23
- 40
- 46
- I do not know

6. Which organ in your body excretes insulin? *

Mark only one oval.

- Kidneys
- Liver
- Pancreas
- Stomach
- I do not know

7. Hand sanitizer contains high concentrations of alcohols such as __ *

Mark only one oval.

- Methanol
- Ethanol
- Isopropanol
- None of the above
- I do not know

8. Which of these compounds can be extracted directly from opium? *

Mark only one oval.

- Fentanyl
- Morphine
- Hydrocodone
- Naloxone
- I do not know

9. What happens as the body develops a tolerance for opioids? *

Mark only one oval.

- Receptors become less responsive to opioids.
- Smaller doses of opioids are equally effective.
- The body increases its number of opioid receptors.
- Noradrenaline levels increase.
- I do not know

10. What controls the movement of the lungs? *

Mark only one oval.

- The ribcage
- One's heart rate
- The diaphragm
- The autonomous nervous system
- I do not know

11. From largest to smallest, the airways are structured as follows: *

Mark only one oval.

- Trachea, bronchi, bronchioles, alveoli
- Alveoli, bronchi, trachea, bronchioles
- Trachea, bronchioles, bronchi, alveoli
- Bronchi, alveolus, bronchioles, trachea
- I do not know

12. What is the biggest danger when staying in a room without proper ventilation? *

Mark only one oval.

- Oxygen levels in a room drop over time, because humans exhale CO2 and consume O2. One could die from a lack of oxygen.
- Bacteria and fungi grow optimally without air. One could could get poisoned by inhaling these organisms.
- As combustion is not perfect, CO levels increase in a closed space. If one inhales this gas, it binds to hemoglobin and organs are deprived of oxygen.
- One might go insane without air, because fresh air is necessary to think properly.
- I do not know

13. What is the relationship between the mitochondria and the host cell in terms of ectosymbiosis? *

Mark only one oval.

- ( ) No ectosymbiosis, but endosymbiosis.
- ( ) Mutualistic
- ( ) Commensalistic
- ( ) Parasitic
- ( ) I do not know

Finance

This section contains questions in the field of finance. These questions do not represent the whole field of finance. These questions were selected with the intent to be understood by people who understand basic financial terminology. If you do not know an answer, please fill in 'I do not know', rather than guessing. As mentioned earlier: your payment does not depend on the number of correct answers given, so only answer the questions using your own knowledge.

14. Which tax scheme is the least common? *

Mark only one oval.

- ( ) Automation Tax
- ( ) Sales Tax
- ( ) Income Tax
- ( ) Wealth Tax
- ( ) I do not know

15. What is the term for decreasing purchasing power per unit of money as a result of an increase in money supply? *

Mark only one oval.

- Constipation
- Deflation
- Inflation
- Economic Collapse
- I do not know

16. Which of the following things can be considered as both an asset and a liability? *

Mark only one oval.

- A purchased house (Home Ownership)
- Buying a Netflix subscription
- Camera to take professional photos
- Desktop PC to play games recreationally
- I do not know

17. Which is not an example of a shock that can prompt a recession? *

Mark only one oval.

- Earthquake
- Regulation
- Technological Innovation
- Inflation
- I do not know

18. Is it wiser to put your money into one investment or multiple investments? *

Mark only one oval.

- ◯ One investment
- ◯ Multiple investments
- ◯ One or multiple investments are equally risky
- ◯ Do not invest at all, because there is less risk involved in spending your money at a casino.
- ◯ I do not know

19. Bob is unemployed, single, childless, in good health and drives his beaten car from the US to Mexico. Which insurance could he most likely do without? *

Mark only one oval.

- ◯ Life insurance
- ◯ Health insurance
- ◯ Car insurance
- ◯ Travel insurance
- ◯ I do not know

20. What is an annuity? *

Mark only one oval.

- ◯ An investment that has no definite end and a stream of cash payments that continues forever
- ◯ A stream of cash flows that start one year from today and continue while growing by a constant growth rate
- ◯ A series of equal payments at equal time periods and guaranteed for a fixed number of years
- ◯ A series of unequal payments at equal time periods which are guaranteed for a fixed number of years
- ◯ I do not know

21. Which of the following formulas correctly calculates the amount of money one has in their savings account (S) after n years? The current amount (X), an interest rate (r) are given constants. Assume that the current amount X does not change within n years (i.e. no deposits, no withdrawals). *

Mark only one oval.

- S = X / (1+r)^n
- S = X / (1+r)*n
- S = X * (1+r)^n
- S = X * (1+r)*n
- I do not know

22. Which statement about the denominations of the US Dollar (USD) and the Euro (EUR) is true? *

Mark only one oval.

- The EUR has a €2 coin, whereas the USD has no 2$ bills in circulation.
- The EUR has no cents/pennies.
- The EUR has no quarters (25 cents) as denomination.
- None of the above are true
- I do not know

23. Which of the four methods below is the cheapest way to pay when you are abroad? (Assume that all methods are at your disposal at your destination of choice) *

Mark only one oval.

- Paying in a stronger currency (e.g. paying in USD on an Caribbean island)
- Exchange your currency for the destination currency.
- Withdraw cash from the ATM abroad.
- Pay by card at your destination abroad.
- I do not know

Computer Science

This section contains questions in the field of finance. These questions do not represent the whole field of finance. These questions were selected with the intent to be understood by people who possess basic computer science knowledge. If you do not know an answer, please fill in 'I do not know' or 'I have no clue what happens', rather than guessing an answer. As mentioned earlier: your payment does not depend on the number of correct answers given, so only answer the questions using your own knowledge.

24. Which of the following topics is unrelated to computer science? *

Mark only one oval.

- Designing Systems
- Technical Support
- Blockchain
- Big Endian - Little Endian
- I do not know

25. Which programming paradigm attempts to avoid side-effects as much as possible? *

Mark only one oval.

- Object-Oriented Programming
- Functional Programming
- Procedural Programming
- Declarative Programming
- I do not know

26. If you can verify whether a solution to a puzzle is correct in reasonable (polynomial) time, then __ *

Mark only one oval.

- ◯ You will be nominated for the Turing award.
- ◯ The speed in which the puzzle can be solved does not depend on how long it takes to verify it.
- ◯ The puzzle can be solved as slow as or slower than verifying the solution.
- ◯ The puzzle can be solved as fast as or faster than verifying the solution.
- ◯ I have no clue what happens.

27. Which of the following types of loops allows for variables to be declared outside its scope? *

Mark only one oval.

- ◯ For-loop
- ◯ While-loop
- ◯ Do...While-loop
- ◯ All of the above
- ◯ I do not know

28. Which of the following research fields is a sub-branch of Artificial Intelligence (AI)? *

Mark only one oval.

- ◯ Machine Learning
- ◯ Cognitive Intelligence
- ◯ Human Intelligence
- ◯ Robotics
- ◯ I do not know

29. Suppose Bob wants to sort quite a large list of numbers of size n. Bob's computer has little memory and cannot store the provided list. Which of these algorithms is the fastest (i.e. average run-time complexity) given the aforementioned constraint? *

Mark only one oval.

- ◯ Bubblesort
- ◯ Quicksort
- ◯ Bucket sort
- ◯ Block sort
- ◯ I do not know

30. Which of the four testing levels has the biggest scope? *

Mark only one oval.

- ◯ Unit testing
- ◯ Integration testing
- ◯ System testing
- ◯ Acceptance testing
- ◯ I do not know

31. Which of the four names does not belong to a software design pattern? *

Mark only one oval.

- ◯ Factory
- ◯ Operator
- ◯ Decorator
- ◯ Adapter
- ◯ I do not know

32. Provided an arbitrary program with a random set of instructions that we cannot look into (black-box), can we predict whether the program will stop running? *

Mark only one oval.

- ◯ Yes, because every action must have an end.
- ◯ No, because that would require the halting problem to be decidable.
- ◯ Yes, because we can stop it at any time.
- ◯ No, because the decision trees of programs are unpredictable.
- ◯ I do not know

33. Which type of network makes torrenting, browsing the dark web and blockchain possible? *

Mark only one oval.

- ◯ Peer-to-Peer network
- ◯ Neural network
- ◯ Search network
- ◯ Ring network
- ◯ I do not know

End of the Knowledge Trivia

That's all for the trivia questions. Your feedback is much appreciated.

34. Feel free to write down your feedback:

Interacting With a Conversational Search Agent

This is part II where you will get to ask questions to a chatbot. The instructions are provided in this pdf:

1) https://drive.google.com/open?id=1aliRxhnuW8PsCv7EXhmMMNdmS6_vVNRA

2) Open web.telegram.org or your own Telegram client of choice (Mobile app/Desktop app/etc.)

Make sure to create a Telegram account first, if you have not already done so. If you have any questions, contact us by typing: @conversational_search_researcher in the search bar in Telegram or send an email to us over Prolific.

35. I confirm that ___ *

Tick all that apply.

- [ ] I own/created a Telegram account
- [ ] I opened Telegram
- [ ] I opened the document with instructions
- [ ] I understand the procedure to move on to the second part of the study

Refusal to Provide Consent

You unfortunately cannot participate in this study, because you refused to provide consent. Submitting this form will bring you back to Prolific and your submission will be dismissed.