University of Zurich[UZH]

Conversational Crowdsourcing for Hypotheses Generation in Data Science

Thesis December 11, 2022

Emanuel Graf
of Seengen, Switzerland

Student-ID: 17-534-199
emanuel.graf@uzh.ch

Advisor: **Rosni Kottekulam Vasu**

Prof. Abraham Bernstein, PhD
Institut für Informatik
Universität Zürich
https://www.ifi.uzh.ch/ddis

Acknowledgements

First and foremost, I would like to thank Dr. Prof. Abraham Bernstein for providing me the opportunity to write the master's thesis on conversational crowdsourcing.

I would like to thank my supervisor and mentor, Rosni Kottekulam Vasu, for bringing the weight of her considerable experience and knowledge to this thesis. Her high standards have made me better at what I do. She provided invaluable guidance and support during the development of this thesis.

Thanks also to my parents, Matthias and Therese who provided me all their support throughout this time.

I would also like to thank all the people who have contributed and supported my experiment and thus provided valuable data for my results.

Finally, I wish to give my thanks to the University of Zurich and the DDIS research group for providing me the opportunity to complete this work and pursuing my master's degree.

Zusammenfassung

Die Automatisierung von Data Science ist ein wachsendes Feld, das darauf abzielt, die Anwendung von Data Science Techniken effizienter, genauer und zugänglicher zu machen. Eine der primären Aufgaben in Data Science ist die Entwicklung relevanter Hypothesen. Menschen besitzen die Kreativität und den notwendigen Sinnesfindungsprozess, um Hypothesen zu entwickeln. Dies kann in einer Crowdsourcing-Umgebung verwendet werden, um Hypothesen über einen Datensatz zu generieren.

Qiu et al. (2020c) schlug das Crowdsourcing mittels Konversation "Conversational Crowdsourcing" vor, das den Crowdsourcing-Prozess intuitiver und benutzerfreundlicher machen kann. Darüber hinaus kann dies dazu beitragen, die Teilnahme und das Engagement der Crowdworker zu erhöhen. Inspiriert von dieser Arbeit übernimmt diese Thesis das Konzept des Conversational Crowdsourcing, um Hypothesen mittels dieser Art von Crowdsourcing zu generieren.

Diese Thesis untersucht den Einfluss verschiedener Gesprächsstile, die verwendet werden, um mit Crowdworker zu kommunizieren. Zwei unterschiedliche Stile – "machine-like" und "mixed" – wurden entwickelt und als Konversationsstile verwendet. Darüber hinaus untersucht die Thesis den Einfluss von Informationselementen, wie Text und Visualisierung, welche den Crowdworker präsentiert werden, und ob diese die Qualität der von diesen Crowdworker generierten Hypothesen beeinflussen. Die Thesis betrachtete auch, wie die kognitive Belastung der Crowdworker durch Gesprächsstile und Informationselemente beeinflusst wird. Dazu wurde ein Experiment durchgeführt. In dem Experiment generierten 40 Crowdworker der Amazon MTurk-Plattform 164 Hypothesen in einer Chat-basierten Umgebung. Die generierten Hypothesen wurden von Domänenexperten auf ihre Qualität hin bewertet.

Die Analyse zeigt, dass es komplexe Abhängikeiten zwischen den Versuchsbedingungen gibt. Die Ergebnisse deuten darauf hin, dass textbasierte Informationselemente und ein gemischter Gesprächsstil die Crowdworker weniger kognitiv belasten. Darüber hinaus zeigen die Ergebnisse, dass eine bestimmte Kombination aus Gesprächsstil und Informationselementen die Qualität der Hypothese beeinflusst. Insbesondere die textbasierten Informationselemente und der Gesprächsstil "machine-like" erzeugen Hypothesen von höherer Qualität als andere Kombinationen. Die in dieser Arbeit präsentierten Ergebnisse zeigen jedoch keine statistische Relevanz. Weitere Forschung ist erforderlich, um die in dieser Arbeit durchgeführte Analyse zu stärken.

Abstract

Automated data science is a growing field that aims to make the process of applying data science techniques more efficient, accurate, and accessible. One of the early and primary tasks in data science process is the development of relevant hypotheses. Humans posess the creativity and necessary sensemaking process to come up with hypotheses. This can be used in a crowdsourcing environment to generate hypotheses about a dataset.

Meanwhile, Qiu et al. (2020c) proposed the conversational crowdsourcing which can make the crowdsourcing process more intuitive and user-friendly. Moreover, this can help to increase participation and user engagement. Inspired by this work, this thesis adopts the concept of conversational crowdsourcing to generate hypotheses by a non-expert crowd.

This thesis investigates the impact of various conversational styles used to communicate with the crowdworker. Two distinct styles–"machine-like" and "mixed"–were developed and used as conversational styles. Moreover, the thesis examines the influence of information elements, such as text and visualization presented to the crowdworker and whether these affect the quality of hypotheses generated by these crowdworkers. The thesis also considered how the cognitive loads of the crowd are impacted by conversational styles and informational elements. For this, an experiment was conducted. In the experiment, 40 workers from the Amazon MTurk platform generated 164 hypotheses in a chat-based environment. The generated hypotheses were rated on their quality by domain experts.

The analysis shows that there are complex interdependencies across the experiment conditions. The results indicate that text-based information elements and a mixed conversational style put less cognitive load on the worker. Furthermore, the results show that a specific combination of conversational style and information elements influences the quality of the hypothesis. In particular, the text-based information elements and machine-like conversational style generate hypotheses of higher quality than other combinations. However, the results presented in this thesis do not show statistical relevance. Further research is required to strengthen the analysis done in this work.

Contents

1

Introduction

Automation of data analysis has advanced. Data scientists nowadays have many possibilities for the automation of data analysis. A choice they have is whether they want humans involved in the automation task. Data analysis can profit from the creativity of humans and their proper sensemaking process.

A primary data analysis task and a task that humans can be involve in, is the definition and generation of relevant and useful hypotheses. In the bio-medical field, generating hypotheses is a well-studied research topic. Numerous computer algorithms to do so have performed well in applications like drug development (Schneider et al., 2019). However, the majority of these investigations rely on literature mining and fail to consider alternative data sources' accessibility or the benefits of strategies that include humans in the process (Thilakaratne et al., 2019). Humans are able to offer theories and insights to make sense of the data if they are given the raw data as well as other information such as tables, charts, or graphs in addition to text.

In order to include humans in the data analysis and hypothesis generation process, it requires the search for adequate subjects. This is where crowdsourcing offers promising services, providing a cheap and reliable pool of workers capable of delivering results on demand. Crowdsourcing allows for a variety of complicated activities to be automated (Kittur et al., 2011). Examples are available in finding dimensions for categorizing thoughts or generating values for such dimensions (Huang et al., 2021), making sense of massive datasets (Willett et al., 2012), and in many more areas.

The field of using crowdsourcing for generating hypotheses and for automating the early stages of a data science project has been less studied. Automating the process of creating hypotheses is challenging since it requires human creativity and their participation. Willett et al. (2012) provided methods for enhancing the variety of the hypotheses. They also provide demonstrations of the utility of crowdsourcing in social data analysis.

In this work, the development of hypotheses by crowdworkers using a conversational interface will be investigated. Focus will be laid on which factors influence the cognitive load of the crowdworker and the hypothesis quality generated by the worker. By adapting the work of Qiu et al. (2020c) and creating a crowdsourced experiment where hypotheses are generated and rated on their quality, the hypotheses presented in the next section are addressed.

To the best of our knowledge, no research has been done, especially on the topic of generating hypotheses for data analysis via conversational crowdsourcing. A conversational

interface will be used in the framework of this thesis to answer the proposed hypotheses. It will be looked at the efficacy of this conversational interface for crowdsourcing the creation of hypotheses.

1.1 Research Questions

The main focus of this work is how different conversational styles and information elements for presenting datasets, such as textual description of data, data visualisation and data tables will influence the cognitive load of the crowdworker and the hypothesis quality generated by them. There are two research questions and four hypotheses that will be answered in this thesis. With the help of an experiment which includes two different conversational styles, two different representations of information, and two unique datasets, the following will be tested:

RQ_1**: Is there a difference in hypothesis quality and cognitive load of the crowd worker when using different information elements in a chat based interface?**

To answer this research question, the hypothesis quality generated by workers and the cognitive load of workers will be tested while having two different variants of information elements.

- $H_{1.1}$: Conversations using a combination of data visualisations, tables and text to convey information improve the quality of hypotheses generated in a chat-based interface, compared to conversations without data visualisations.
- $H_{1.2}$: Conversations using a combination of data visualisations, tables and text to convey information leads to a lower cognitive load of the crowd worker, compared to conversations without data visualisations.

RQ_2**: Is there a difference in hypothesis quality and cognitive load of the crowd worker when using different conversational styles in a chat based interface?**

In order to answer this research question, the quality of crowd-generated hypotheses and the cognitive load of workers will be tested under two different variants of conversational style.

- $H_{2.1}$: Conversations using a humanlike conversational style for non-informational discussion and machine-like conversational style for presenting information improve the quality of hypotheses generated in a chat-based interface, compared to conversations with only machine-like conversational style.
- $H_{2.2}$: Conversations using a humanlike conversational style for non-informational discussion and machine-like conversational style for presenting information in a chat interface lead to a lower cognitive load of the crowd worker, compared to conversations with only machine-like conversational style.

1.2 Organization

The further parts of this thesis use the following structure to present the work that was done. Chapter 2 covers the available and relevant literature in the area of research for this thesis. Chapter 3 presents the content creation that was done for the experiment carried out in this work. This chapter covers the development of the different conversational styles and the creation of the information elements from the identified datasets. How the experiment was set up in detail is discussed in chapter 4. The experiment environment for the workers with the different conditions will be discussed. Furthermore it will be looked at how the assignment of quality ratings to hypotheses is accomplished and how the analysis part of the results is developed. In the results in chapter 5, the analysis of the data is presented, described and explanined. The discussion in chapter 6 is used to dive deeper into the meaning of the previously presented results. Chapter 7 lists the limitations of this thesis and the work that was done in here. Some useful ideas for future work on this topic are presented in chapter 8. At last, everything is wrapped up in the conclusion that can be found int chapter 9. Some final remarks together with the key points of this thesis are presented in there. The URL to the Gitlab repository, where all the content of the experiment together with the raw data are stored is in the appendix, section A.3.

2

Background and Related Work

In this chapter, the available work around the topic of this thesis will be reviewed. It is split into three sections in total. The first and most important section is conversational crowdsourcing, followed by the data science pipeline and then the area of hypothesis generation.

2.1 Conversational Crowdsourcing

Previous research about conversational crowdsourcing covers the split up and distribution of work to many workers, also known as crowdsourcing. This crowdsourcing part is combined with a conversational aspect for a more communication-focused approach.

Retelny et al. (2017) look at why tasks that contain complex goals are difficult to solve with the help of crowdsourcing approaches. The paper concludes that the traditional static way crowdsourcing works nowadays can prohibit the achievement of complex tasks through it. The paper indicates that adaptation of this traditional setting is required to progress in this field. This work is a substantial entry point of the area in which this master thesis will work. Following up on the previously mentioned work about crowdsourcing and going into the area of conversational microtask crowdsourcing, Qiu et al. (2020b) worked on the question of how effective conversational interfaces in microtask crowdsourcing can increase worker engagement. To achieve that, they built a conversational agent that tests different conversational styles in an experiment with workers on Amazon Mechanical Turk. They found that conversational interfaces and suitable conversational styles can be effective in improving workers' engagement. But how exactly is it possible to estimate the style of the conversation in conversational microtask crowdsourcing? Again, Qiu et al. (2020a) have worked on this topic. They have now developed methods to estimate the individual conversational styles of workers and find that certain styles, especially *involvement conversational styles* of workers resulted in a significantly higher quality of output, followed by higher user engagement, and the user would perceive less cognitive task load too. A follow-up question to the above would be, how can user engagement be measured? Around this topic, O'Brien and Toms (2010) have identified six attributes of engagement. These attributes are the following: Focused Attention, Aesthetics, Perceived Usability, Novelty, Felt Involvement, and Endurability

(O'Brien and Toms, 2010). The result of their research is a tool that is able to measure the engagement of a user. Furthermore, their findings also show that usability has an important role in the interplay with all other identified attributes. Staying in the field of engaging the user, Baee et al. (2020) have developed a framework to help with the development and design of virtual coaching systems. They focused on four topics: reliability, fairness, engagement, and ethics. Especially interesting for this thesis is the engagement part. Baee et al. (2020) argue that such a system needs to have its core built on a human-centered design approach. An engaging system should hold the attention of the user and should be able to provide value to them not only for the short term but also for the long term. Furthermore, the modeling of engagement relies on metrics like for example click count, frequency of application usage, and the time that is spent on a certain task. What features appropriately represent engagement? Ultimately, they will differ from this virtual coach framework to other domains, such as conversational crowdsourcing. However, the key takeaways of this paper can always be considered and tested to see whether this would be transferable and applicable in other domains, such as conversational crowdsourcing. Spitale and Garzotto (2020) even go a step beyond the engagement of the user through the conversational agent and developed a framework that provides a tool for evaluating and designing conversational agents that are empathic. Their paper adds value to the research in conversational interaction. Specifically to understand the role of empathy in these interaction, which can be combined with the information that origins from the other papers about engagement here.

If conversational agents can detect user engagement and decide, which user engagement style is most productive or efficient for the task, it might also be possible to make the agent even smarter and tailor the conversation to the knowledge a worker brings to the table. Work in this area has been done by An et al. (2021). They find that it can be useful to a conversation if it is known whether a user already has knowledge about a topic. This knowledge reduces the amount of definition requests and paraphrasing. An et al. (2021) successfully test and implement three methods to detect knowledge in the field of Conversation Analysis, namely prior difficulty in understanding, prior exposure to a reference, and self-reports of knowledge. This technique could be useful when thinking about the engagement or mental workload of the workers.

Going back from the area of conversations again to crowdsourcing, an earlier approach to tackling the problem of crowdsourcing complex work is presented by Kittur et al. (2011). In 2011, they had already developed an interesting prototype that helps to split up a complex task and set up the crowdsourcing solution approach to it. They have conducted case studies on article writing, science journalism, and decision-making with this prototype. Their work shows that such crowdsourced articles can reach a similar level of quality as those written by individuals. It also shows that there are difficulties if the end goal is not clearly stated at the beginning of the task - a situation that can usually happen with clients and their desires. Kuttal et al. (2020) worked on a conversational agent to replace a human in pair programming (because finding a good partner and scheduling sessions is difficult). Strategies for creative problem-solving were found, together with conversational style differences when creating such an agent. Furthermore, they analyzed how strategies between human-human and human-agent

collaboration can be transferred and the effects this will bring.

How to convey a complex task to crowdsourcing workers was one of the main challenges that Huang et al. (2021) faced when they completed their design challenge about generating Dimension/Values for categorizing ideas. The main part of successfully transmitting the requirements to the workers was to decompose the cognitive process together with the validation of the completion of each cognitive subprocess. Upon this understanding, Huang et al. (2021) work along a self-defined, task-related 5-step strategy towards this decomposition.

2.2 Data Science Pipeline

In this section, the related work around pipelines in data science and hypothesis generation will be presented. First, in the area of data science pipelines, the work of Wang et al. (2021) looks at auto-generating textual representations of datasets. They have developed a framework for exactly this task, showing that machine-generated data stories are of comparable quality to data stories written by humans. With the work by Wang et al. (2021), it can also be shown how an automatic information analysis from a given dataset can be conducted.

Müller et al. (2012) worked on a project to develop a tool that can support creativity in the biomedical area called "DataCreativityTools" (Müller et al., 2012) to primarily support scientists looking at the data with different approaches. They state that transferring their tool to other domains would depend "on the availability of specific information sources and their stage of development of semantic information infrastructure" (Müller et al., 2012). However, the strategic key takeaways can be used here in the area of this thesis to work with automated data analysis visualisation.

2.3 Hypothesis Quality Evaluation

Willett et al. (2012) provide seven strategies which can improve the quality and the diversity of hypotheses generated by workers in a crowdsourcing environment. Their research shows that if the experiments include feature-oriented formulated questions, good explanations with examples and reference gathering, annotation subtasks (of features in the chart/dataset) and more will result in higher-quality answers from the workers. Furthermore, they have defined a basic formula to measure the quality of hypotheses. A numerical quality score is the result of this formula using a scale from zero to five, where zero marks the worst quality and five the best quality. The score is calculated by adding together the clarity of the hypothesis ("how easy it is to interpret") (Willett et al., 2012)) and the plausibility ("how likely it is to be true") (Willett et al., 2012)), which are both numerical values on a scale from one to five. The sum of these two values is then divided by two to get to a one to five scale again. This overall score is then multiplied by the binary value relevance score, which is based on "whether it explains the desired feature [...]" (Willett et al., 2012). The equation (2.1) shows the formula by Willett et al. (2012).

$$quality = ((clarity + plausibility)/2)(relevance) \tag{2.1}$$

But Willet et al. were not the only ones looking into quality requirements for hypotheses. Quinn and George (1975) formulated criteria on which a hypothesis could be measured for its quality. They formulated an acceptable hypothesis as the following:

> According to Quinn and George (1975), a statement had to satisfy at least one of the following criteria to be an acceptable hypothesis:
>
> - (1) it makes sense
> - (2) it is empirically based
> - (3) it is adequate
> - (4) it is precise
> - or (5) it states a test

These are the five criteria. Moving forward, it's also visible from their work that the more of the criteria mentioned in their work that a hypothesis fulfills, the higher the quality of the particular hypothesis.

2.4 Datasets

Three datasets from Kaggle have been identified which will be used in the creation of information elements for this thesis. The first one is the "Mental Health in Tech Survey" (Men, 2014) from a "2014 survey that measures attitudes towards mental health and the frequency of mental health disorders in the tech workplace" (Men, 2014). It includes various information about the individual subjects and answers to the survey. The second dataset is the "World Health Statistics 2020" (Wor, 2020). It is a dataset to cover the health statistics of the world. It contains country names along with their health data such as road traffic injuries, mortality from environmental pollution, life expectancy, and healthy life expectancy and many more. The third dataset "Google Play Store Apps" (Goo, 2019) is settled in the domain of apps from the Google Play Store. It contains "Web scraped data of 10k Play Store apps for analysing the Android market" (Goo, 2019)

3

Experiment Content

In this chapter, the data acquisition, together with the generation of information elements for the experiments from the identified datasets and the conversational styles will be discussed. First, the identified datasets are presented. After that, it is looked at the development of the two conversational styles that will be used in the experiment. Finally, the information elements that are generated using the identified datasets together with their creation process will be presented.

3.1 Datasets

Kaggle (Kag, 2022) was used to find appropriate datasets for this work. In total, three datasets were identified. As the author of this thesis was interested in health topics, particularly mental health, one dataset was chosen specifically about mental health and another about health facilities in Ghana. The third one is in a completely different domain, containing statistics about the Google Play Store. This third one was included to test the interdisciplinary application of the work presented here. The first dataset has the title "Mental Health in Tech Survey" (Men, 2014) and contains around 1200 data points from a survey of people who work in a tech-related company. It captures 26 information fields about the participants, of which 22 are questions specifically related to the survey topic, for example, "Do you have a family history of mental illness?" (Men, 2014) or "Do you think that discussing a mental health issue with your employer would have negative consequences?" (Men, 2014).

The second dataset "Google Play Store Apps" is - as the name suggests - settled in the domain of apps from the Google Play Store. It contains "Web scraped data of 10k Play Store apps for analysing the Android market" (Goo, 2019). Reviews, Ratings, together with the App names, their last update date, and more technical details are available in this dataset.

"Ghana Health Facilities" (Gha, 2018) is the name of the third dataset and provides a listing of health facilities in Ghana, together with data about the ownership and type of health facility, organised by geological data such as region and districts. This dataset was used solely for the tutorial which is provided to the worker within the experiments, for the workers to learn how to perform the task.

3.2 Conversational Styles

To answer $H_{2.1}$ and $H_{2.2}$, different conversational styles needed to be created. The work from Stan (2020) introduces two conversational styles: humanlike style and machine-like style. In this section, both styles will be presented with their application to this paper. Next to that, a mixed style, born out of the previously mentioned two styles, will be introduced as a potential approach to tackle the possible problem of distraction by a certain style in different parts of the workers' workflow through the task.

3.2.1 Humanlike Conversational Style

Humanlike conversational voice/style in chats contains personal, informal, and playful messages according to Stan (2020). For the chatbot, human-like linguistic elements were introduced into the script as presented by the work of Stan. However, this style in its pure form will not be part of the experiment conditions, in favour of the later discussed mixed style, because of the fact that the mixed style can represent humanlike conversational style as well, while maintaining a better level of accurate information transmission (Paula Chaves et al., 2019). The following elements from their work were used in the scripts for the chatbot.

Message Personalization: Greeting the stakeholder/worker personally is an effective way to create a humanlike voice. Since the name of a certain worker is unknown at the point of the experiment, it was tried in the script to work around this issue by greeting them in a friendly and still personal way: "Hi there!". Another type of message personalisation used in the scripts was to address the stakeholder/worker directly. Two examples of this would be the following. First in the beginning right after the greeting, the worker was asked how they were doing today. At another point in the scripts, they were told that the requester needs help from them to generate a hypothesis. In this question, personal pronouns were used. Next, to address every stakeholder that is present in the conversation , the chatbot itself was addressed as well. This is also considered a message personalisation linguistic element which was adapted from the list by Stan (2020).

Informal Speech: Informal speech means casual everyday language and the use of certain special elements. Non-verbal cues are one of them. These were used in the scripts by inserting emoticons now and then to produce a more friendly and welcoming environment. Through that it should stimulate the worker to generate higher quality work and lower their cognitive load. There were other types of informal speech in the work of Stan (2020), such as abbreviations (lol, pls) or interjections (haha, oh), but these were not used since it was thought that these would create a too playful and distracting environment.

Invitational Rhetoric: Invitational rhetoric as Stan (2020) describes it should stimulate the crowdworker to engage in the conversation and support a mutual understanding between both parties. In the scripts, two elements were used for this. First, by acknowledging the worker's work by thanking them for a message or pointing out the importance of something they just texted. Second, by trying to stimulate dialogue by asking the

worker what they think and whether they want to say something about a certain subject. This element was also used to get the workers thinking about a certain topic in a certain way so they would be able to produce hypotheses in the upcoming steps of the scripts.

3.2.2 Machine-like Conversational Style

This style (which is also referred to as Robot Voice or style and is abbreviated with "mach") does not use any particular elements that would make the messages from the chatbot to the worker personal, informal, or playful in any way. It stays neutral and focused on information transmission without any interference. Of course, the chatbot still needs to be polite, as the paper by Paula Chaves et al. (2019) points out. However, this paper also presents the importance of using unambiguous language, as mentioned earlier, when the goal of communication is information transmission. And with this information in mind, the third category, "mixed conversational style", was created.

3.2.3 Mixed Conversational Style

A short overview of the idea of this style/voice (abbreviation for this style: "mix"): In the first sections of the scripts for the hypothesis generation task, the conversational style will be humanlike style. In the sections of the scripts that are about hypothesis generation this changes.In these sections, where the workers are asked to produce a hypothesis about a given dataset, a machine-like conversational style will be used as the conversational style.

When it comes to information transmission, communication should be clear, straightforward, and completely unambiguous to avoid any interference that could potentially deform the information to be transmitted (Paula Chaves et al., 2019). However, not all sections of the scripts are about down-to-the-point data and information transmission. Rather it also focuses on the worker themselves and what they are all about. Especially in those parts of the script that are conducted before the main hypothesis generation part:

- where the worker introduces themselves
- where the worker lists their knowledge about the domain of the dataset
- where the general task and the example task are introduced.

In these sections, important information is exchanged as well. However, compared to the other sections, it is also about forming a space for the worker where they feel welcomed and personally valued. To form a space for the worker as just mentioned is the first reason for the introduction of mixed conversational style. The second reason is the idea that this approach could lower the cognitive load while improving the quality of the hypotheses. That is because the information transmission in the important areas is kept at the most accurate level possible.

3.3 Information Elements

An important aspect of this thesis is to convey information about a given dataset. This is important for the crowdworkers, in order for them to be able to generate hypotheses. To test whether different types of information elements influence the quality of the hypotheses generated by the workers ($H_{1.1}$) and the cognitive load of the workers while doing the task ($H_{1.2}$), two types of information elements were chosen. First, textual elements present information through words and sentences. Second, visual elements present information by showing charts, diagrams, and other visual figures that have statistical meaning.

3.3.1 Textual Elements

Abbreviation for this term: "text". Alternative name: "Textual Information Elements". The most frequently occurring textual element in the scripts conveys information about correlations between dataset features. For this, the following structure was used:

'{value_1}' and '{value_2}' have a strong connection to each other. Can you think of any other feature that might have some correlations to any of the mentioned values? List the features below, separate multiple by semicolon ";" :

{value_1} and {value_2} represent two different dataset features that are somehow (positively/negatively) correlated to each other. It was tried to give as many hints about the data as just explained as required for the crowdworkers to think scientifically about the given dataset and hypothesis generation. At the same time it was tried to reduce the amount of hints given. The reason for this is to not lead the crowdworker into a specific biased direction, or limiting their field of view to just the variables that were provided.

3.3.2 Visual Elements

Abbreviation for this term: "viz". Alternative name: "Visual Information Elements". Throughout the work of creating meaningful visualisations, there were different approaches of developing them. In the beginning, there was uncertainty about the exact scope and through it what the requirement for the chatbot tool would be. One thought that came up was to dynamically generate the data visualisations in the browser based on the input dataset given by the scientist. However, it was quickly realised that such an approach would most certainly be out of scope for the project and could potentially be harmful to the answer to the initial hypotheses by shifting the focus of the work. Therefore, it was realised that switching to a data scientist-oriented environment such as R or Python would be more productive. As Python was used in other parts of the project, it was decided to use Jupyter notebooks to generate visualisations. Most of the time, the Altair Library was used to create visualisations (Veg, 2020) as well as the Plotly library (Plo, 2022). Both libraries were tested, but to stay consistent across all datasets and visualisations, it was decided to only use visualisations created with the Altair library.

Existing notebooks from Kaggle were used to explore the datasets. In these notebooks, To get ideas of how good and creative visualisations for the identified datasets would look like, notebooks generated with the datasets by users presented on Kaggle were used to start this process of creativity. All these notebooks are publicly available and linked on the respective dataset page on Kaggle.

3.3.3 Special Case Tables

Tables are somewhat in between the two elements discussed above. Having some non-textual elements, such as grids and colors, means they could be going in the direction of visual elements. But since they are heavily reliant on the text within them together with the fact that it's way easier on the eye to represent sample data in a table rather than in full text or even worse in an enlisting it will be considered to be a textual element, which allows it to be inside scripts that are text only. This argument is the main reason for not having a table-only condition in the experiment (see section 4.2).

3.4 Informational Elements within the Experiments:

For all eight experiment conditions visible in section 4.2, a consistent representation of data had to be reached to reduce the external influential factor that could have a potential impact on the result of the quality of the hypotheses generated by the workers. Therefore, the following elements were created and used within the conditions:

Tutorial Section across all Conditions: Short introduction of the dataset (Ghana Dataset) together with a text that presents some statistical information:

> "In Ghana, there are many types of health facilities. In the dataset, there are for example clinics and training institutions. The region with the largest amount of clinics is the Greater Accra region with 283 clinics. This region also holds the record for the most training institutions. The region with the lowest amount of clinics is Upper West. It has only 10 clinics, and also has the least number of Training facilities, counting 4. All regions in between show a gradual slope between training institutions and clinics."

Tutorial Section across the Visual Elements Conditions: For the introductions of all the experiment conditions that include visualisations for information transmission, a heatmap of the health facilities and the regions where the health facilities are located in Ghana is shown in figure 3.1.

Introduction to the dataset in the main task: For the introduction of the dataset that was used by the MTurk workers to generate hypotheses, a short informational text was presented to the workers, including how the data was obtained and some insights into the dataset. This text in its structure is similar to the one in the tutorial, with its content being adapted to the dataset. Furthermore, a table was included in the introduction part that presents an overview of all the captured features of the dataset in question. For one dataset, this table also included some random sample data for the

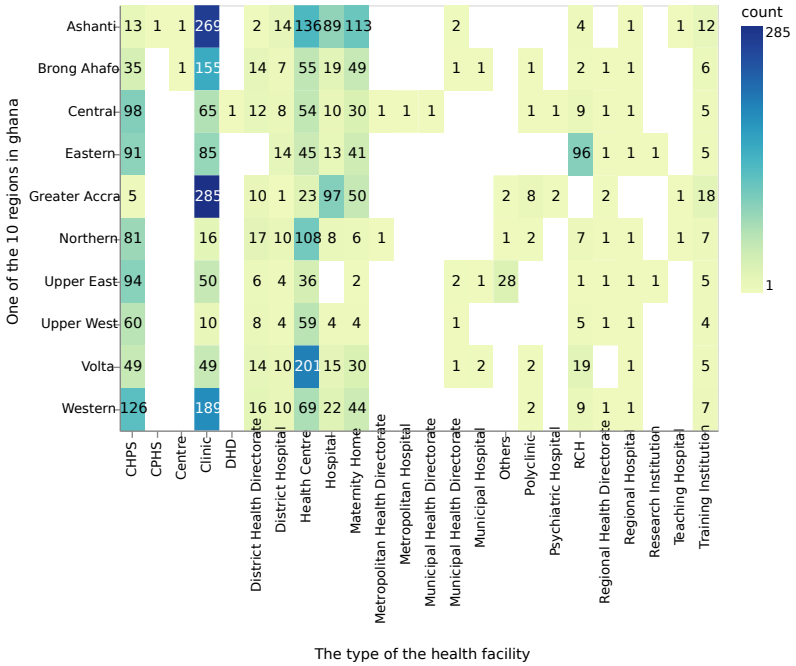


Figure 3.1: Visualisation for tutorial section of all visualisation experiment conditions

workers to understand what possible data points for this feature could look like. For the other one, all data points were yes/no answers, therefore it was not necessary to provide sample data and it was decided to provide the explanation of the features in full sentences rather than sample data.

Visualisations in the main task: For the experiment conditions that included visualisations to convey information about the dataset to the workers, three types of charts were chosen. There were more types in the initial phase of generating visualisations, but eventually, it was decided to use bar charts, stacked bar charts, and heat maps. In the appendix section A.1 all of the used visualisations are shown.

Text in the main task: For the textual information elements, the workers were confronted with similar sentences as for the visualisations, but the workers had no visualisations, they had to look up the necessary information in the dataset overview. However they had the same possibility to look information up in the dataset overview in the visual elements conditions. An example sentence for such text can be seen in the previously discussed “Textual Elements” in subsection 3.3.1.

4

Experiment Setup

This chapter is devoted to the setup of the crowdsourcing experiment. As mentioned in Ramírez et al. (2021), it is a key factor to the reproducibility of an experiment to report the setup of the experiment in a detailed manner. To accomplish this in the current setting, the checklist that was defined by Ramírez et al. (2021) was used. Some of the content of this checklist is already covered in previous or the upcoming sections of the thesis and will not be repeated in this chapter. The filled out checklist is visible in the appendix, section B. The URL to the Gitlab repository, where all the content of the experiment together with the raw data are stored is in the appendix, section A.3.

4.1 Workflow of Experiment for Workers

In this section, the workflow of the experiment will be discussed. Within the task design it will be looked at how the workers were assigned to the task, how the task interface looks like and how the script tries to lead the workers to think in the bigger picture of a dataset.

4.1.1 Task Design

For the task design of how a worker will step through the experiment, the work from Qiu et al. (2020a) was used to establish a solid fundament. Some changes have been made to their Conversational Task Design to match the goals set for this thesis. The Conversational Task Design is visible in the figure 4.1. Its content is discussed in the respective chapters.

4.1.2 Worker Assignment to Task

The workers are selected randomly and as a pre-screening measurement, they are required to pass a qualification test and have some further qualifications, see section 4.3.2. The allocation to a certain experiment condition is arbitrary and is taken care of by Amazon. Amazon assigns a worker that has passed all the quality criteria to any of the eight conditions. Each worker is only able to participate in one experiment condition in total, meaning that the experiment will be according to a between-subjects study

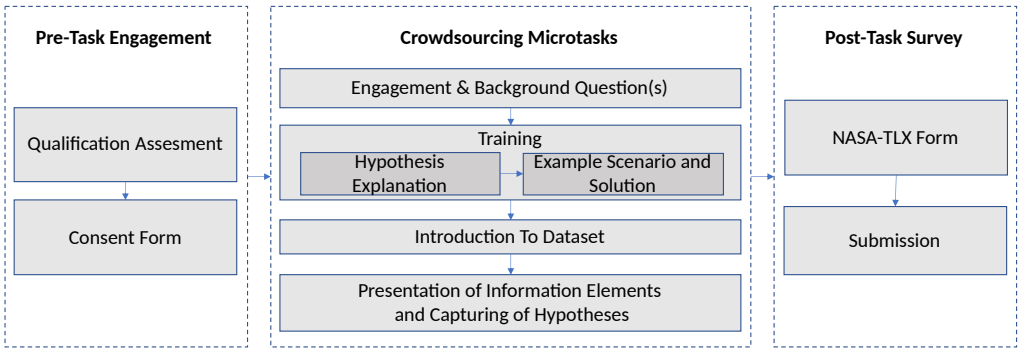


Figure 4.1: Customized Conversational Design, originally from Qiu et al. (2020a)

design. This decision was made based on the reasoning that there might be a learning curve in "writing good hypotheses" through experiments that will affect further experimental outcomes and that the effect of the different conditions might be diminished by letting workers participate in multiple conditions. The fact that the assignment of crowdworkers to a task is taken care of by Amazon also helped with dropout rates. If a crowdworker finishes without completing the task, the MTurk platform automatically assigns this task to another crowdworker.

4.1.3 Task Interface

The task interface is embedded in the MTurk environment (Ama, 2018a). The task interface greeted a worker with the consent popup, which contained a scrollable text element and an accept button. This is further explained in section 4.1.4. Once accepted, the worker was presented with a chat interface. Chat bubbles popped up when the chatbot said something, or when the worker replied to the chatbot. This chatbot is an altered version of the work by Qiu et al. (2020c). The main things that have changed for the worker in contrast to the tool that Qiu et al. (2020c) provided were that this chatbot can present visualisations (iframes) as shown in figure 4.2 and sliders as in figure 4.3. Apart from these, the chat interface is the same as the one from Qiu et al. (2020c). Dr. Qiu has multiple papers with multiple altered versions of his chatbot. He as well has papers with chatbots that look close like the TickTalkTurk, but use another technology. Therefore it is hard to point out the exact small differences to the chatbot in this thesis. This is why only the main changes to the chatbot used here are presented.

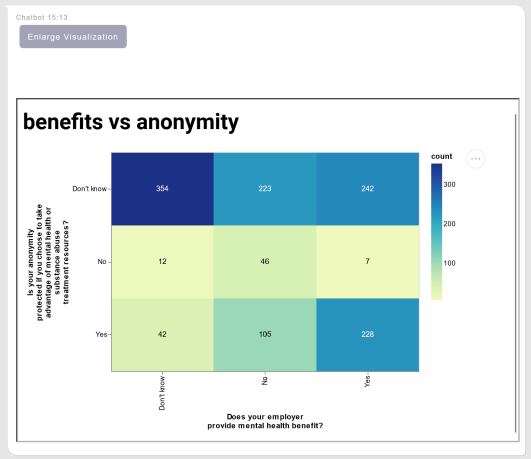


Figure 4.2: Visualisation element in the chatbot

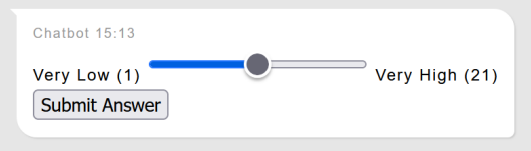


Figure 4.3: Slider element in the chatbot

4.1.4 Introduction Popup

The introduction popup covers many areas of the requester-worker interaction. Moreover, it also includes the privacy statement and data treatment, informed consent, and the participation awareness text. At the beginning of the popup, it is mentioned that this task is part of a research study from the University of Zurich. The workers from Amazon MTurk work under the privacy and data agreement which Amazon provides (Ama, 2018b). For the experiment, the workers received the information that the only personal information that will be available to the researchers is what is publicly available on their MTurk profile and any information that they choose to provide during the study. They also had to approve that they fullfil all the requirements that the researcher requested, to ensure a minimum level of quality in the responses. These requirements were that the worker is fluent in English, that the worker did not do the task and some more. By clicking "Accept", the workers acknowledge that they have read the rules and privacy policy, that they certify to be 18 years of age or older, and agree that their participation is voluntary.

4.1.5 Force Workers to keep the bigger picture

For the tasks in the experiment to result in qualitative hypotheses, it is important that the workers always understand and work with the bigger picture of a dataset when

presented with specific graphics or textual information. Otherwise, it could result in limited hypotheses that focus only on the latest information element presented to the worker. The question that needs to be answered to achieve this is: How to force the worker to think in the bigger picture?

One possible answer to this is the introduction of subtasks, which will eventually lead to the case that a worker will have to educate themselves with the information provided outside of the current information element but still within the experiment.

Subtasks are questions to the workers that ask them to provide information about relationships between the provided information elements. If the subtask answers make sense and are qualitatively acceptable, it may be valid to say that the worker could successfully gather the information to be able to think within the whole information sphere provided to them within the task. Furthermore, the idea of the subtask picks up the findings of Huang et al. (2021) from their work "Task Decomposition". There, they take a complex task and break it into multiple small steps. These are more comprehensible for the workers. Another factor that was considered when creating these subtasks was to get the most out of the answers of the workers and thus formulate the questions accordingly. The paper by Willett et al. (2012) focussed not only on quality formulas but also presented strategies for formulating questions to produce high-quality answers. In the introduction of the overall tasks, good examples of answers were used, and in the subtasks, feature-oriented prompts were used, such as "please explain feature "xy" in one sentence or less". This directly asks the worker to deal with the details of the dataset and get familiar with it to eventually generate valuable hypotheses.

A sub-task might look as follows:

- Chatbot: Here is another visualisation.
- Chatbot: *Sends Visualisation*
- Chatbot: Describe one feature of your choice in this visualisation (feat_X/feat_Y) in no more than one sentence.
- — Wait for Workers' input —
- Chatbot: The features in this visualisation have a strong connection to each other. Are there any other features that might have some correlations to any of the mentioned values?
- — Wait for Workers' input —
- Chatbot: Please write down the hypotheses about the relationship between the features you found. Write complete sentences [...]
- — Wait for Workers' input —

4.2 Experiment Combinations

As discussed in the section 3.2 about the possible conversational styles, a total of two different styles will be tested. These two styles can be applied to the identified datasets. Two of the three identified datasets were used for the main task, and one dataset for the introduction/tutorial part of the task of every condition. Thus, the conditions use either of the two datasets for the main task. For the information elements, there are also two possible choices as well: visualisations and textual representation of information. This in total results in the final eight experiment conditions that will be tested. All combinations are visible in the table 4.1.

These eight conditions are used to test the quality of the generated hypotheses and the cognitive load of the worker when faced with different information elements and different conversational styles.

No.	Experiment Name	Dataset	Conversational Style	Informational Elements
1	ds1-csmach-text	Mental Health	Machine-like	Tables + Textual Explanations
2	ds1-csmach-viz	Mental Health	Machine-like	Tables + Data Viz
3	ds1-csmix-text	Mental Health	Human- & Machine-like	Tables + Textual Explanations
4	ds1-csmix-viz	Mental Health	Human- & Machine-like	Tables + Data Viz
5	ds2-csmach-text	Google Play Store	Machine-like	Tables + Textual Explanations
6	ds2-csmach-viz	Google Play Store	Machine-like	Tables + Data Viz
7	ds2-csmix-text	Google Play Store	Human- & Machine-like	Tables + Textual Explanations
8	ds2-csmix-viz	Google Play Store	Human- & Machine-like	Tables + Data Viz

Table 4.1: All experiment conditions

Total number of expected submissions: There are 8 conditions. For each condition, 5 workers will complete the task which results in a total of 40 submissions.

4.3 Experiment Environment

In this section, the crowd (Workers) is presented, which works on the tasks. Furthermore, the rewards for the tasks are explained, as is how quality control is carried out.

4.3.1 Workers

Amazon MTurk (Ama, 2018a) offers an accessible and reliable source of workers who can carry out assigned tasks online. It was decided to acquire the help of these workers for the experiments in this thesis to answer the stated hypotheses. No specific demographics were required for the tasks provided, but some minimal quality requirements needed to be fulfilled by the workers to participate in the experiments. These will be discussed in the upcoming section.

4.3.2 Quality Control

Amazon MTurk provides some inbuilt quality control features to ensure a minimal quality of the workers performing the tasks. To obtain workers that fulfill a general level of quality, MTurk provides some generic settings that can be adjusted. Two of them used in the experiments in this thesis are *"Worker_NumberHITsApproved"* which indicates how many HITs submitted overall by a Worker have been accepted. A zero or positive integer represents the value, and secondly, *"Worker_PercentAssignmentsApproved"*, which is the portion of the Worker's submissions that the Requester ultimately authorised as compared to all of the Worker's submissions. An integer between 0 and 100 defines this value (Qua, 2022). For the experiments, it was decided to only accept workers with at least 500 approved hits and an overall score of approved assignment of higher or equal to 95 percent. It was also chosen to implement a short qualification survey as a pre-selection process. In this survey, general questions about hypotheses and the worker's understanding of them are asked. The worker has to score at least 80% of the answers correct to gain the qualification to participate in the real experiment.

Furthermore, it was a requirement that one worker may only complete one condition of the experiment and shall be excluded from completing any of the other seven available conditions. The idea behind this is to reduce a bias or possible order effect on the worker that might occur if they complete one experiment condition and are affected by it and then go to an arbitrary other one and pose answers in a biased manner. To do so, it was required to assign a qualification to all the workers that have started an experiment condition task. There are procedures to assign qualifications to workers to exclude them from a certain task. It can be done in a manual way, with Excel or CSV files for example. However, because these options are done in manual work, these options to block a worker are asynchronous. That means that the worker will be blocked some time after they have completed the task on MTurk. But based on the reason that all 8 experiments are released simultaneously, it's a necessity to immediately be able to block a worker, because once they finish a task, they are redirected straight to the next task, which is likely to be another experiment condition. Therefore, an automated solution that could handle the blocking process immediately was required.

Because the experiment conditions are loaded into MTurk via the API in a static manner, only JavaScript code could be executed and no server-side code execution such as PHP was possible. There is no implementation for the AWS API for client-side JavaScript and therefore, the assignment of the exclusion qualifications to the worker

was required to originate from a server-side call. Requests from the Amazon MTurk platform to any outside servers are restricted by CORS-Policies (Cro, 2022). However, a workaround that was found after some digging was that an image from the external server could be loaded that would trigger the PHP script on the external server to successfully block the worker by assigning them the qualification. This idea originated from personal observations of many large website tracking and analytics provider, where they use this tactic to track website visitors via a one pixel image, presumably for users with disabled JavaScript or similar.

Demographics of the participants: There were no demographic restriction in place for the experiment. Amazon MTurk Crowdworkers with any demographic were able to join the experiment. Because of this, together with the information that the MTurk results provided and the question that were asked in the experiment, no data about the demographics of the crowdworkers were collected.

Rejection Criteria: Not all workers put in the same effort in generating hypotheses and the researcher reserved the right to reject certain results. This was especially the case when no visible effort was put in to generate hypotheses and therefore the results of this submission could not be used. Only one example occurred during the experiment, where one worker just repeatedly replied "Great, Great" and "Yes" to all of the questions. This submission was the only one to get rejected. There was also the case where workers did not generate hypotheses in all three of the available prompts during the task. However, based on the fact that in all of these cases they wrote a hypothesis in two of the three prompts, these submissions were not rejected. These rejection criteria also sum up the post-task checks: all hypotheses were quickly scanned manually if they include normal English language and full sentences.

Ethical Approval: The experiment received ethical approval from the IRB from the University of Zurich on the November 23, 2022.

4.3.3 Rewards

To ensure that workers receive the minimum hourly wage of 7.50 USD, the average maximum time a worker would take to complete the task was approximated by giving the task to family and friends. It was found that on average, a worker would not need more than 30 minutes to complete the task. This approximation was to be found correct, as only three of the 40 workers took longer than 30 minutes. Therefore, the reward in Amazon MTurk was set to 3.75 USD to match the minimum hourly wage, and the workers were told that they get paid for a 30 minutes task.

4.3.4 Execution of the Experiment

The part of the experiment which was executed in the Amazon MTurk took around 24 - 36 hours to complete. To capture all the ratings for each of the hypotheses, between one to two weeks were used to send out inquiries to people asking to rate and wait for their work to be completed.

As part of the pilot studies, friends and family were asked first to participate in the task in the setting of think-aloud sessions. This has helped to refine the text that was shown to the worker through the chat interface. Furthermore, before the main task was launched on MTurk, two experiment conditions were launched on MTurk with one worker per task each to see, whether everything goes according to plan. In this initial pre-test phase on MTurk, the first result returned by a worker contained bad work, this worker only answered every question with "Great, Great", "Yes" and similar short wordings. This result was rejected, and one of the two assignments got reassigned. After that, both of the tasks returned acceptable quality answers, with proper English and full sentences.

After all of the tests were done, all eight experiment conditions were launched and shortly thereafter, all of the results were available. During the time the task was live on MTurk, it was checked frequently, how the results looked. This was done to reject possible gibberish results and to act quickly if some reassignments of the task would be necessary. Due to the fact, that an astonishing 100% of the tasks published on MTurk returned results that were useable, no rejections were needed (except for the pilot execution).

4.3.5 Apply a quality score to the Hypotheses

Upon gathering all of the hypotheses from the task submissions on MTurk, the hypotheses needed a quality score. In order to determine the hypothesis quality, each hypothesis from the submission needed a rating. To achieve this, a tool to rate hypotheses based on PHP, HTML, JavaScript, and MySQL was built. It includes many features. This tool shows the people who rate the hypotheses a simple HTML form. In this HTML form is the hypothesis to rate and some background information about the database, upon which the specific hypothesis was generated. Furthermore, the seven quality criteria questions, which are discussed in detail in section 4.4 are in the tool. These quality questions were structured as Yes/No radio buttons for the binary questions and a 5-star rating scale for the questions that require such a scale rating. For each question, the explanation of this quality criteria could be shown by clicking on a "help" button in the respective question area. To capture an accurate rating of one hypothesis according to the quality criteria, it was decided that multiple ratings by different people would improve this accuracy, rather than just relying on one person's rating of one hypothesis, which could potentially be biased on the person's perception of the circumstances (their perception of the dataset, certain ways a hypothesis was formulated, etc), their own knowledge and further aspects. Therefore a minimal rating count of 3 ratings per hypothesis from different raters was set. For an even distribution of the ratings across all hypotheses, the following steps have been taken. The hypotheses were shown to the raters sorted first by the current rating count of the respective hypothesis and second by the dataset. The thought behind sorting by dataset was, that raters would not constantly need to switch between datasets and their background information, which could potentially lead to confusion. with the current system in place, it could happen that one hypothesis received multiple ratings. This happens when two or more raters are rating

at the same time. It could have been prevented by storing a "hypothesis reservation" in a separate database table. This reservation would make sure that once a hypothesis was served to a rater, no other rater shall receive the hypothesis for a certain amount of time. This approach was not implemented due to the scope of this thesis and because the occasion that two raters would rate at the same time did not happen often. The results show that less than 1% of all ratings resulted in a higher overall rating count per hypothesis. This is further discussed in the results section 5.2.

Recruiting Raters: The people who would rate the hypotheses needed to fulfill certain criteria to be a rating person. The requirements were that the person would have to have at least a Bachelor's degree or similar, they need to be fluent in English and they must have some experience in Data Science or an equivalent skillset with analytical tasks that could make up for it. People from the University were recruited such as students. In the University context, no pre-screening took place, as it was to be expected that these people fulfill the criteria. The people who expressed the interest in rating were sent the link to the rating tool directly. There was also a recruitment process outside of the University context, namely, on the internet, forums such as Reddit (/r/, 2022), and other places. For those people that wanted to rate hypotheses and came from such an external source, a qualification questionnaire took place. A Google form was used as a tool for this. Interested applicants were checked whether they fulfil the criteria by asking them questions such as "What is the highest educational qualification you hold or are currently pursuing?". However, due to the fact that there was no monetary reward for people who would rate hypotheses, it was to be expected that this part of the experiment would take up more time and more incentives would have to be produced in order to recruit people and motivate them to rate hypotheses. A general idea was to implement some common marketing techniques such as to generate a certain level of sense of urgency (without causing stress) and to implement some gamification elements into the rating process to motivate and bind the raters to the task to make them rate as many hypotheses as possible. This was done while carefully making sure that these elements would not have any influence on the conscientiousness and quality of the work that the rater provide. One step of precaution that was taken to not distract raters from the rating process is that on the main rating page, where the raters do the actual rating, no gamification features are shown at all. The focus on this page was clearly set on the rating. Further measures to not influence, bias, or distract the raters were taken in the respective elements themselves. The gamification and motivational elements will be explained further in the following section.

Gamification and motivational Elements in Quality Tool: When a potential rater opens up the rating tool introduction page, they are greeted with a short text about what this tool is all about and some background information about the experiment. After that, the first element that should motivate people to rate more is presented to the potential rater. It shows a progress bar that updates according to the current count of ratings. Together with this progress bar, a text informs the potential rater about the average rating amount that is needed per rater in order for the rating count to reach the desired and required amount of ratings. Figure 4.4 shows this element at 232 ratings out of 640. The maximum amount of ratings per hypothesis was regarded as flexible,

whereas the minimal amount of required ratings per hypothesis was set to be three. In the progress bar, the average amount per hypothesis is set to be around 4.

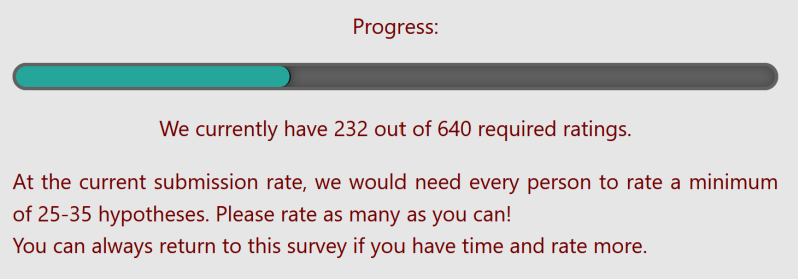


Figure 4.4: Create a sense of urgency

Figure 4.5 shows the button positioned at the bottom of the introduction page. It redirects to the rating part of the tool. The button has a pulsing shadow that should make sure that the potential raters do not overlook it. Also visible in this figure is the input form for the username. This is a completely optional field that lets the rater compare themself to other raters by appearing on the leaderboard. The leaderboard, of which parts are visible in Figure 4.6 is there to incentivise the raters to beat other raters and climb to the top spots of the list, eventually rating more. All ratings are in there. For people who did not provide a username, a random name shows up. Examples are "Unknown Rater" or "Unknown Hero".

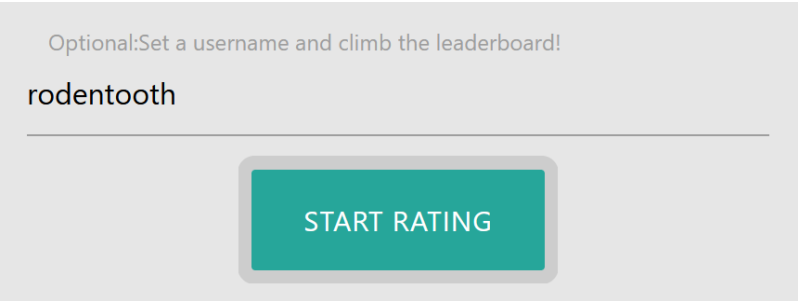
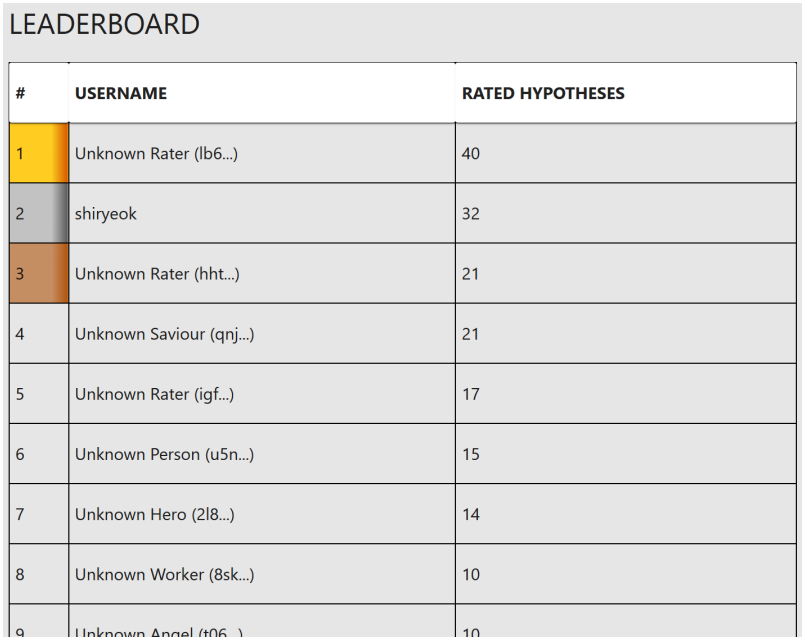


Figure 4.5: "Start Rating" button together with username input

The next motivational feature is the thank you message on the submission confirmation screen. In figure 4.7 the message is visible that shows up to over 20 different motivational messages, which are also dependent on the score a rater already has. An example of that is: For the first three ratings, the rater receives another incentive to rate further hypotheses. The text shows: "Ready for the next hypothesis to rate?". The last feature is connected to these messages. For every 10 ratings, the rater receives "XX Ratings! Your kindness level increased +1" as a message. The X stands for the personal rating count of the rater. This connects to the "personal stats" feature.

The "personal stats" feature are achievements that can be gained by raters for every ten hypotheses they rate. Visible in figure 4.8 it shows that there are four different



LEADERBOARD

#	USERNAME	RATED HYPOTHESES
1	Unknown Rater (lb6...)	40
2	shiryeok	32
3	Unknown Rater (hht...)	21
4	Unknown Saviour (qnj...)	21
5	Unknown Rater (igf...)	17
6	Unknown Person (u5n...)	15
7	Unknown Hero (2l8...)	14
8	Unknown Worker (8sk...)	10
9	Unknown Angel (t06...)	10

Figure 4.6: Upper part of the leaderboard at an arbitrary point during the experiment

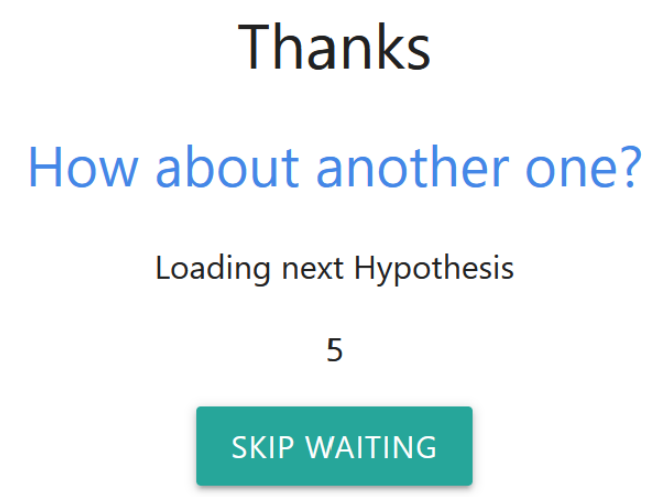


Figure 4.7: Thank you message

categories in which a rater can reach four levels each. The feature has no meaning for the rating itself and the wording was chosen as such so that the feature has no connection to the rating task in any way. It should be a fun addition to the rating process and incentivize raters to achieve new levels in their stats by rating the hypotheses.

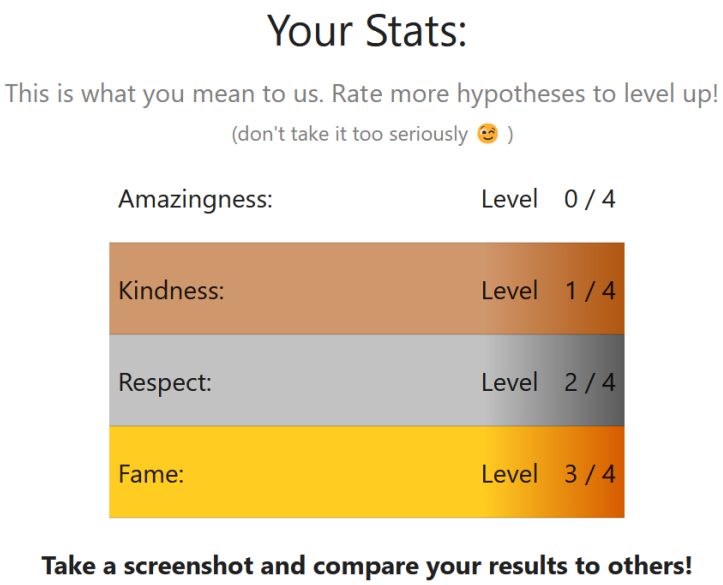


Figure 4.8: Personal Achievements together with the colour changing feature

4.4 Hypothesis Quality Measurement

The formula visible in the equation 2.1 given by Willett et al. (2012) is used for the purpose of this work. The formula is adapted to fit the requirements more closely for the criteria presented in this work. It is used together with a second formula that is based on the work by Quinn and George (1975). The mean of both formulas are used as the final quality measurement for a hypothesis.

Four of the five criteria introduced by Quinn and George (1975) are used in an additional formula. Criterion number five, "it states a test, an explicit statement of a test" (Quinn and George, 1975) was dropped from the inclusion into the formula presented based on the reason that tests as defined in the work of Quinn and George are not included inside the activities of the tasks presented to the workers. An example of a test would be "I could try out my idea (hypothesis) by putting several little bottles with differing amounts of water in them in a tub and then seeing which ones would sink." (Quinn and George, 1975). This shows the connection of this "Test"-criterion to the niche domain in which the paper by Quinn and George was written. In that niche, this criterion had its perfect purpose, especially when teaching children how to write

hypotheses. However, as already mentioned, the other criteria can be adapted to this work, whilst the "test"-criterion fails with its adaptation capabilities for its use here. Therefore, the following four criteria by Quinn and George were used:

- (sense) Does this Hypothesis make sense? Answer in binary 1/0
- (empirical) Does this Hypothesis include empirical observations? Answer in binary 1/0
- (precise) How precisely is this hypothesis written? Answer on a scale of 1-5
- (variables) Does the hypothesis include at least 2 variables? Answer in binary 1/0

These criteria are combined to the following formula:

$$QualityScore_{Q\&G} = \tag{4.1}$$
$$([variables * 5 + sense * 5 + empirical * 5 + precision]/4)* \tag{4.2}$$
$$(\lceil [variables + sense + empirical + \lfloor precision/5 \rceil]/4 \rceil) \tag{4.3}$$

In the equation line 4.2, the quality score from 1 - 5 is evaluated as Quinn and George state in their paper that for each criterion fulfilled, another point is added to the quality score. To finally get a rating from 1-5, in the created formula here, the four available criteria are scaled to 5 and then divided by 4. Quinn and George also say in their paper, that at least one of those introduced criteria has to be fulfilled, in order to be an acceptable hypothesis. That is the reason for the equation line 4.3 in the quality formula. In this line, it is tested whether one of the four quality formulas is true. If not, the whole quality score of a given hypothesis is zero. If one criterion is true, the score ranges from one to five, dependent on the equation line 4.2. With this approach, it was tried to match the criteria and the formula given by Quinn and George (1975) as closely as possible.

For the other part of the final quality score for the hypothesis, the criteria that already existed in the formula by Willett et al. (2012) were used, which are the following three:

- (clarity) How easy is it to interpret? Answer on a scale of 1-5
- (plausibility) How likely is it to be true? Answer on a scale of 1-5
- (relevance) Does it explain the desired feature of the chart? Answer in binary 1/0

The criteria "relevance" was reformulated to "Does it make use of features related to the problem scenario?" to match the use case in this thesis more closely, as it might happen that either A) there is no chart but rather just textual representation of text or B) the workers state a hypothesis using variables that might be in the problem domain but are not included in the dataset as a feature. Such hypotheses are still interesting and should not be considered of lower quality solely based on that reason.

The initial formula (2.1) combined with the formula of Quinn and George is visible in equation 4.4

$$Quality = (QualityScore_{Q\&G} + QualityScore_{Willet})/2 \tag{4.4}$$

As some of these quality criteria might be subjective and rated from one rater to another differently, it is tried to reach a higher accuracy by having ratings from more than one person for one hypothesis. All the available ratings per hypothesis are converted to the quality score and then aggregated by their overall mean score.

4.5 Pipelines and Data Preprocessing

To make the best use of the knowledge the author had on PHP and MySQL, it was decided that for the pipelines and data preprocessing PHP and MySQL would be used. Moreover, the rating tool is already written in PHP, which lowers the barriers for data transfer. Therefore, this decision saves time and the available resources can thus be spent where they would be most effective, such as the creation of the experiment conditions, the creation of the quality rating tool and the analysis.

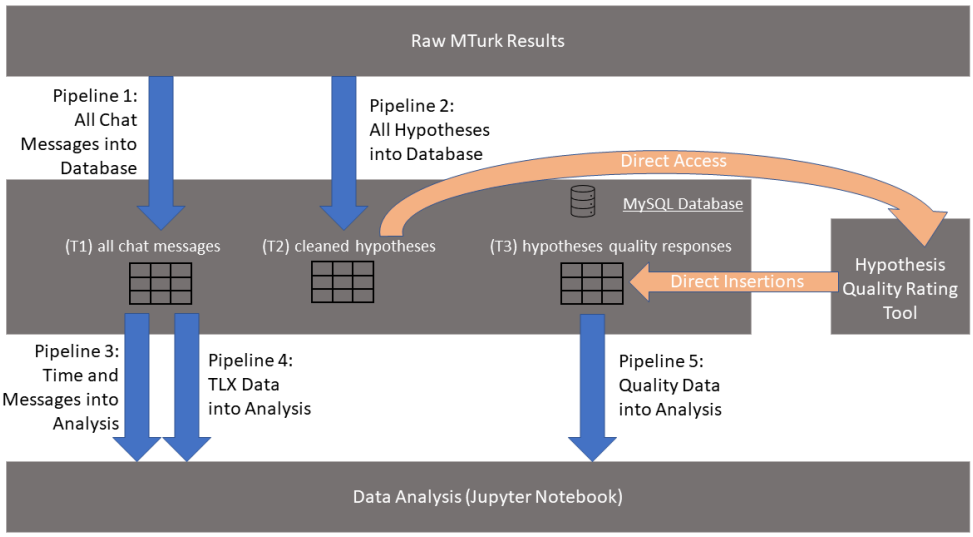


Figure 4.9: Explanatory Overview of Pipelines used to streamline the experiment and analysis process

In figure 4.9 an overview of the developed data handling and the preprocessing system can be seen. The raw MTurk results were available as a JSON string and included all the data described in 4.6.1, except for the quality rating data, which would be created in the quality rating tool. A MySQL database is in place to make it easier to create

subsets of the data and make the selection and updating of partial data easier. Three tables are inside the database. First, a table (T1 in figure 4.9) that contains all raw chat messages with their timestamps and message number, and some meta information such as experiment condition and MTurk worker id. Then, there is the table (T2 in figure 4.9) with all the 164 cleaned hypotheses, which furthermore contains the same metadata on all the hypotheses as the messages in the table "all chat messages" (T1). Finally, there is the table (T3 in figure 4.9) that contains all the responses from the quality tool. Each row contains all the rating information plus the id of the hypothesis from the hypotheses table, together with the user session hash of the rater, to distinguish between multiple raters. In this table, there is also the username of the rater stored, for the leaderboard gamification feature.

Pipeline 1 reads out the whole raw JSON and puts its content without any altering into the table "all chat messages".

Pipeline 3 takes all the messages, counts the time in between the messages, and stores this information in the messages table, which is then converted to a CSV file, which can be directly used inside the analysis notebook.

Pipeline 4 filters all the messages to match the ones that include the answers to the NASA TLX survey. It cleans up the message number field for all survey question numbers to range from one to six and puts this information in a CSV file, which can be directly used inside the analysis notebook.

Pipeline 2 collects only the hypotheses from the raw JSON by looking at the message numbers which are set to be the hypotheses responses from the workers. The message numbers are looked up manually according to the developed chat messages. The hypotheses get cleaned manually inside the table, by going through the table and deleting the hypotheses that do not match the filter set.

The data from the "cleaned hypotheses" table can be directly read by the hypothesis quality rating tool, which displays all the hypotheses to the raters. Upon a finished rating, the tool stores the rating by directly inserting it into the table "hypotheses quality responses".

Pipeline 5 creates a join between the hypotheses, their experiment conditions, and the rating they have received. It creates a CSV from this data, which can be directly used inside the analysis notebook.

4.6 Workflow of Analysis

In this section, the workflow of the data analysis will be discussed. It will be looked at how the data will be analysed and what steps need to be taken to be able to analyse the data in the desired way.

4.6.1 Available Data

The expected data from the experiment is enlisted below. In front of each element is the datatype in which the data will be available.

- (String) The hypotheses generated by the workers
- (Timestamp) The timestamps from the messages sent by the workers
- (int) The answers to the NASA TLX survey
- (String) The knowledge of the worker on the given dataset domain
- (int/bool) The seven criteria for the quality scores for each hypothesis generated by the raters (see section 4.4)

4.6.2 Workflow of Hypothesis Quality Evaluation

The workflow to evaluate the quality of each of the hypotheses is based on the quality evaluation tool, the scientific workers, who have experience with hypotheses and the code part, where the results are analyzed. In figure 4.10 the workflow of the evaluation for the quality of the hypotheses is visible. After all the experiments have been submitted, the results are exported as a string in JSON format from the jupyter notebook which handles the MTurk experiments. The messages are filtered in one of the five PHP data processing pipelines, and the hypotheses are extracted from the submissions. These hypotheses are then imported into the hypothesis quality evaluation tool. The hypotheses are then presented to scientific hypothesis experts. These will rate the hypotheses on the defined rating criteria, discussed in section 4.4. Each participant is free to rate as many hypotheses as they like.

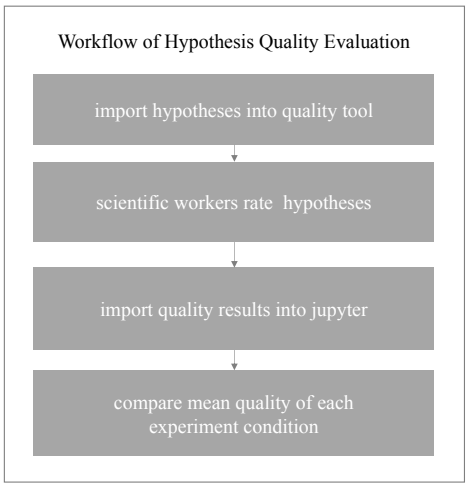


Figure 4.10: quality measurement workflow

4.6.3 Workers Cognitive Load Measurement

To measure the cognitive load of the workers across the experiment conditions, the NASA TLX score (TLX, 2020) was used. Similar to the work of Qiu et al. (2020c), at the end of the task, every worker was asked all of the six criteria of the NASA task load index on a scale from 1 to 21. These criteria are then scaled to 100 and the mean score which is ranging from zero to 100 is taken to compare the different experiment conditions to each other. From the available data, it is also possible to look at the time it took the worker to answer certain parts of the task. There is the assumption that if a worker spends more time replying to the chatbot, then the cognitive load would be higher. However, this assumption is to be made with caution, as it might also be the case that the worker just went away for a cup of coffee during the task or something similar. Such incidents could affect the time that is spent by the worker to reply to the chatbot.

5

Results

In this chapter, the results gathered from the experiment will be analysed. First, the available data will be explained, together with the decision for the appropriate statistical test. After that, the data will be analysed and the results of the experiment compared to each other.

5.1 MTurk Submissions

Description	Value
Received MTurk Submissions	40
Potential Hypotheses in the Submissions	170
Used Hypotheses for further Analysis	164

Table 5.1: Overview of the MTurk submissions

An overview of the available data from the MTurk results can be seen in table 5.1. The 40 received submissions from MTurk were manually checked on their quality and usability in the further analysis. This check was based on the answers provided in the submissions. In the pilot study, it was required to reject one submission because of gibberish answers. Other than in the pilot study, it was found that all of the experiment submissions included a minimum acceptable level of English language. This minimum level means that an answer must be understandable within the context of the conversation. Furthermore, all the answers of the result were in alignment with the questions that they were given. Thus, all the answers were used in the further analysis. From the 40 submissions, 40 NASA TLX form submissions were captured. A total of 170 responses to the request to generate a hypothesis were extracted. From these 170 responses, six had to be removed. That is because they included faulty responses, such as "yes", "sure" or "I think so". It is believed, that this was the result of some different interpretations of some hypothesis generation requests. Especially in the experiment condition with dataset 1, text-based information elements, and mixed conversational style. There, the last question "*Can you please formulate a hypothesis about possible correlations with other features from this information?*" seemed to be understood as a

yes/no question. This had the effect that the answer was based on whether it is possible to generate a hypothesis rather than a request to generate a hypothesis. After these answers were removed, a total of 164 hypotheses with an adequate level of English language were available. These were then put into the quality tool for rating. In figure 5.1 the total amount of available hypotheses available for rating per experiment condition is shown. It is visible that especially for the condition in dataset 2 and mixed conversational style, there are fewer hypotheses than in the other conditions. From this figure, it is also visible that the amount of generated hypotheses is not necessarily dependent on conversational style. This is based on the observation that the differet conversational styles are alternating between the sorted conditions. There some indication, that the amount or quantity of hypotheses generated might be dependent the choice of information elements, as the top three values are all from the conditions with visualisations as information elements. This will be discussed further at a later point in the results, together with more material to analyze.

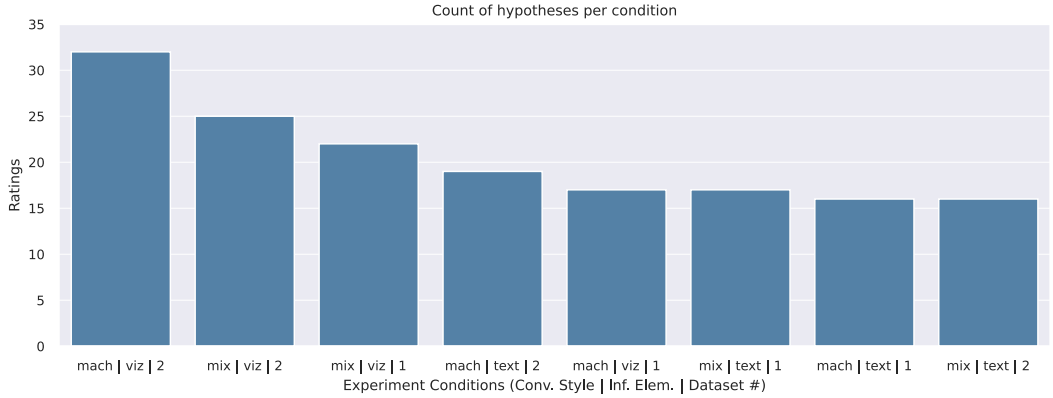


Figure 5.1: Amount of generated hypotheses per experiment condition

The TLX score of the hypotheses follows a normal distribution, according to normality testing. The skewness of the overall TLX scores is 0.1, the average kurtosis is -0.3 (Fisher's Definition) and the average Shapiro-Wilk statistic score (Shapiro and Wilk, 1965) is 0.98. Furthermore, the sample sizes amongst the conditions are equal, as every worker filled out the NASA TLX survey in the MTurk task. As a result, the TLX scores are compared using a nonpaired t-test (Student, 1908) to identify differences among the conditions. In the visualisations used for comparison, the mean value is used as the data follows a normal distribution.

5.2 Quality Tool Submissions

An overview of the available data from the MTurk results can be seen in the appendix table A.1. From the quality tool, 496 hypothesis ratings were captured. 47 people rated all of the available hypotheses. Per hypothesis, 3 total ratings were collected.

Description	Value
Amount of Raters	47
Received Submissions	496
Ratings per Hypothesis	3
Used Ratings for further analysis	492

Table 5.2: Overview of the quality tool submissions

Because of that, the number of available ratings per condition as visible in figure 5.2 is directly correlated to the number of available hypotheses per condition. In figure 5.3 it is visible that the mean of all raters have rated around 7 hypotheses. Four hypotheses had four ratings overall, due to a technical design limitation in the rating tool. One of those ratings for each hypothesis was deleted randomly. After capturing all necessary ratings, the rating tool remained online and accessible. It was tried to reach 4 ratings per hypothesis, which meant 164 additional ratings. This would have brought more diversification in the analysis. These additional ratings could not be collected due to time constraints. Therefore the analysis was done with three ratings per hypothesis. The 10 best and 10 worst hypotheses can be seen in the appendix, section A.2.

The quality ratings of the hypotheses do not follow a normal distribution, according to normality testing. From all of the conditions, the average skewness is -0.58, the average kurtosis is -0.37 (Fisher's Definition) and the average Shapiro-Wilk statistic score (Shapiro and Wilk, 1965) is 0.9. Furthermore, the sample sizes amongst the conditions are not equal, as worker were not restricted in the MTurk task by a hypothesis writing limit. As a result, the quality scores are compared using Mann-Whitney U tests (Mann and Whitney, 1947) to identify differences among the conditions. In the visualisations for the rating analysis, the median value is used for comparisons, as the data does not follow a normal distribution.

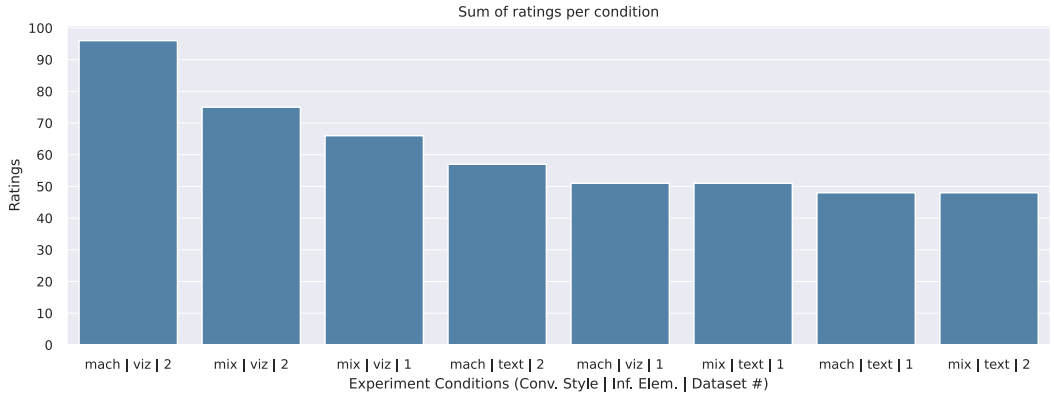


Figure 5.2: Amount of generated ratings per experiment condition

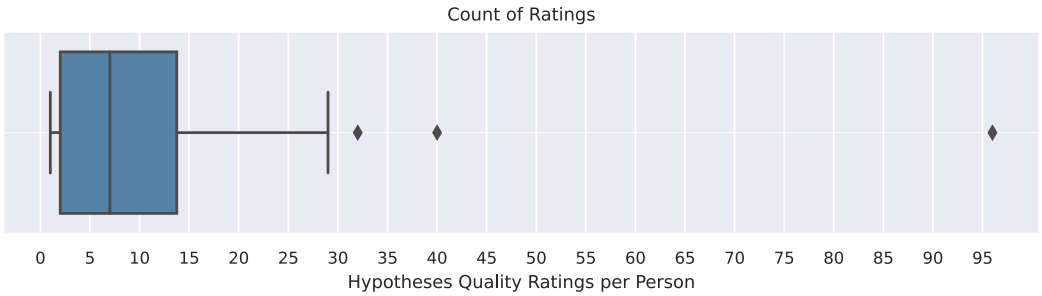


Figure 5.3: Amount of rated hypotheses per person

5.3 Cognitive Load

In this section, the results of the experiment will be presented to compare the cognitive load of the workers across different conditions. The data resulting from the time measurement and the NASA TLX survey results will be used. With this evaluation, $H_{1.2}$ and $H_{2.2}$ are addressed.

As a first step, the statistical significance of the data will be examined. In table 5.3 all conditions are listed where it was possible to reject the null hypothesis of the nonpaired t-test. These conditions have a P-value of <0.05.

Condition Name A	Condition Name B	P-Value
mach-text-1	mix-text-1	0.013
mach-viz-1	mix-text-1	0.015
mach-viz-2	mix-viz-1	0.030
mix-text-1	mix-viz-1	0.005

Table 5.3: Experiment condition comparisons that reject the t-test test null hypothesis in the TLX analysis

If only conversational elements are compared to each other, as well as only information elements and only datasets compared to each other, the P-value of all these comparisons is >0.05.

It was furthermore also not possible to reject the null hypothesis, if only information elements are compared together with conversational styles, with merged datasets. All of the P-values in these comparisons resulted in >0.05.

This means that for the comparisons enlisted in table 5.3 the mean ranks of the two groups are likely to be not equal.

Figure 5.4 shows that for text-based information elements, mixed conversational style results independent of the dataset to lower cognitive load than machine-like conversational style. For conditions that include visual information elements, the opposite is visible from the figure. There, mixed conversational style leads to a higher cognitive load compared to machine-like conversational style, again independent of the dataset.

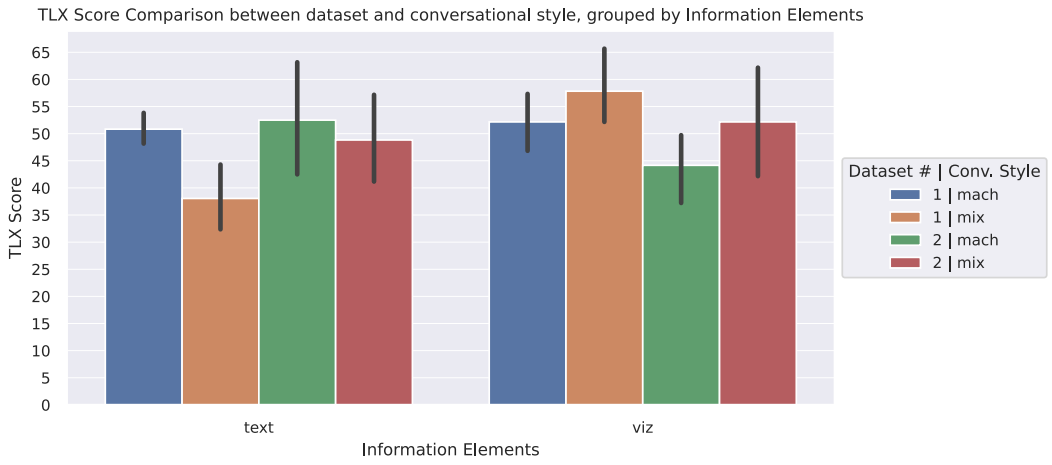


Figure 5.4: The TLX score grouped by information elements, split up by dataset and conversational style

That means that for both of the tested datasets, this occurrence is observable while taking into account the statistical significance discussed previously.

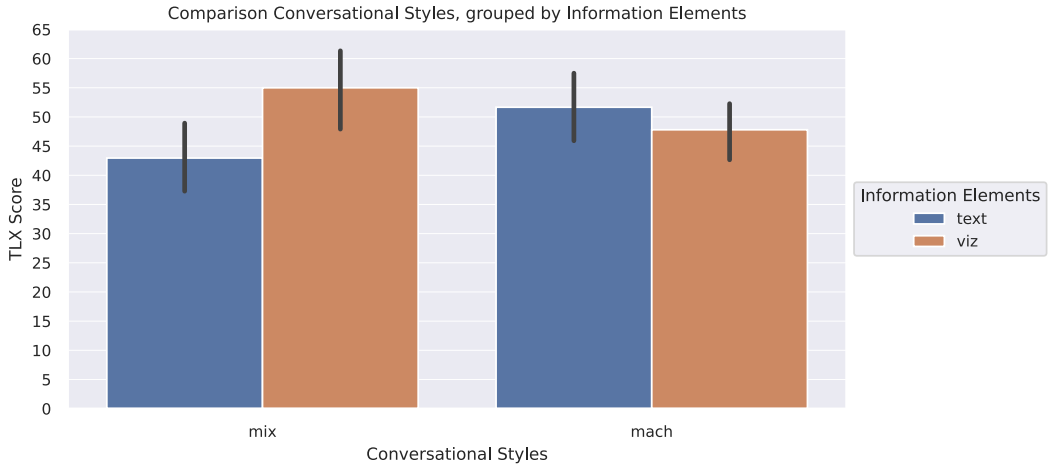


Figure 5.5: The TLX score between conversational styles, grouped by information elements

This effect that was just discussed is better visible in figure 5.5, where the datasets were merged. Conditions containing textual information elements lead to a lower cognitive load on the TLX scale compared to conditions with visual elements, in mixed conversation style, while in machine-like conversational style, the opposite is the case.

When looking at the conversational styles from a time measurement perspective, visible in figures 5.6 and 5.7, it is observable that the conversational styles are, other than the information elements not in an ordered fashion. This means that the mixed and

machine-like conversational styles are alternating, starting with a mixed conversational style condition that holds the longest time used to generate hypotheses and to complete the task. Then, a machine-like conversational style follows, followed by a mixed style and so on. Because of this inconclusive behaviour, it was tried to look at it from a different perspective. When merging the conditions together to compare the time just between the conversational styles such as in figure 5.8, it is visible that the difference is minimal. Furthermore, due to the statistical insignificance discussed earlier, it cannot be told for sure that one or the other conversational style would lead to a lower time used and thus a lower cognitive load.

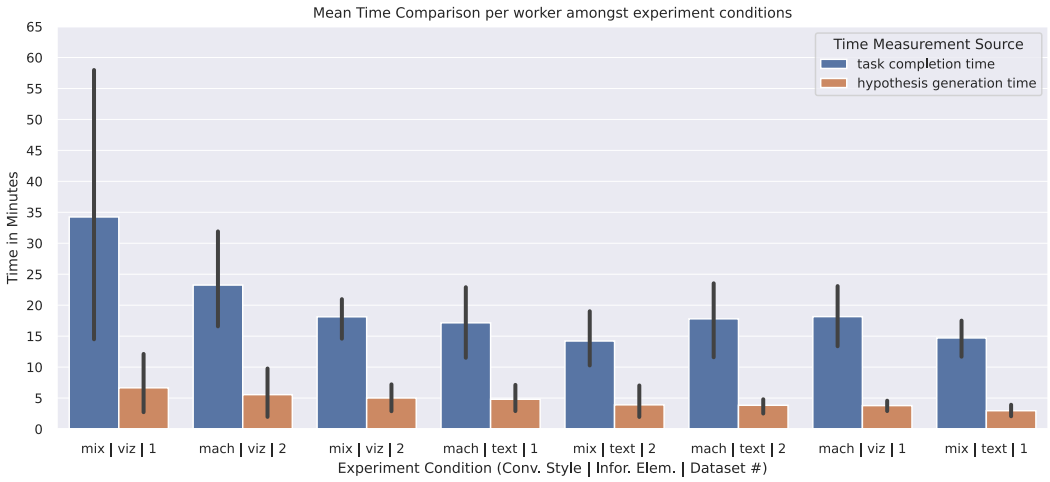


Figure 5.6: The time used per worker to generate hypotheses and to complete the whole task, grouped by experiment condition, ordered by hypothesis generation time

The time comparison did not result in conclusive answers. Moving on, the impact of the two conversational styles on the TLX score is analyzed, which should directly impact the cognitive load that was tested during the experiment. In figure 5.9, the difference between the two conversational styles can be seen. The median of the machine-like conversational style is higher, but the first and third quartile both lies within those of the mixed conversational style. For $H_{2.2}$ "Conversations using a humanlike conversational style for non-informational discussion and machine-like conversational style for presenting information in a chat interface lead to a lower cognitive load of the crowd worker, compared to conversations with only machine-like conversational style." this means that by just looking at the median in the overall comparison as in figure 5.9, the hypothesis can be answered with yes. That is because the median of the TLX score of mixed conversation style is lower. A possible conclusion from the other information based on the comparisons made is as follows: When trying to find out the impacts of conversational styles on the cognitive load, it might also be dependent on the way that the information is being presented. If there are visual information elements the cognitive load is lower with a mixed conversational style. If there is only text-based information given to the

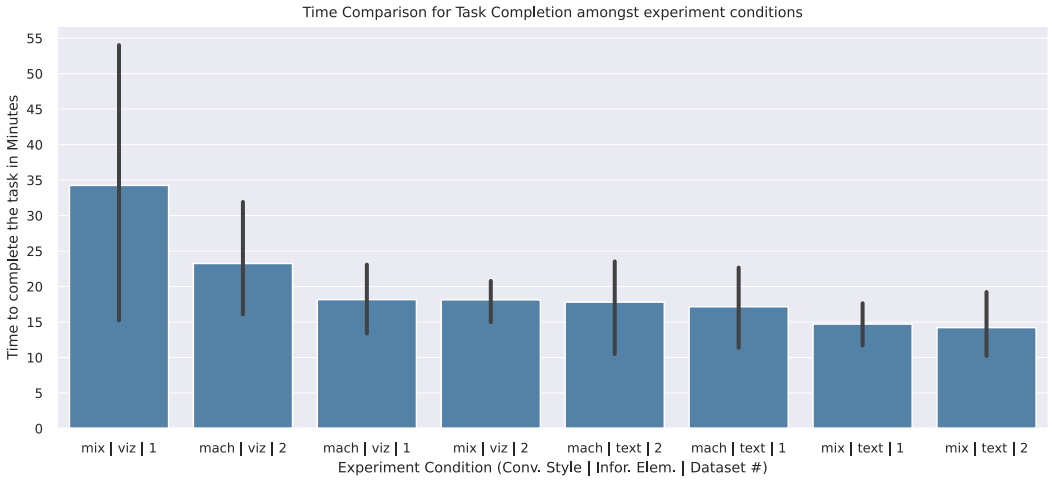


Figure 5.7: The time used per worker to complete the whole task, grouped by experiment condition

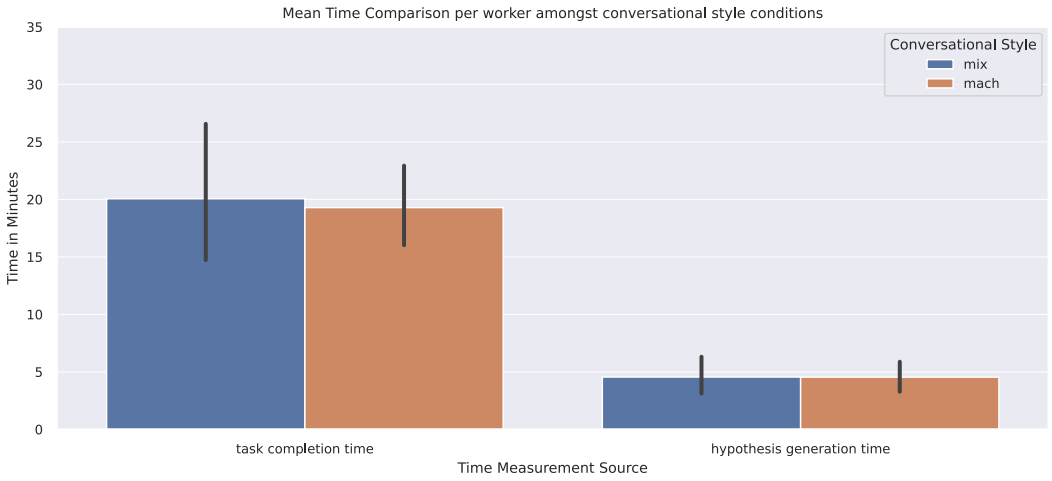


Figure 5.8: The time measurements between conversational Styles

worker, a machine-like conversation style leads to a lower cognitive load.

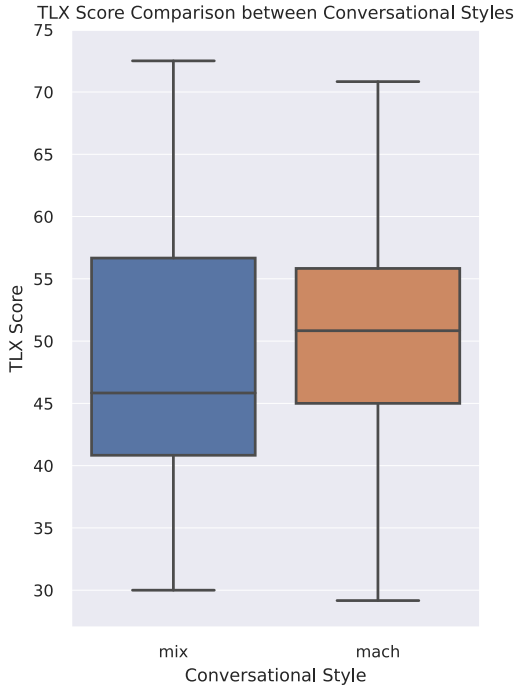


Figure 5.9: The TLX score between conversational styles. All other conditions are merged.

In figure 5.6, the conditions are sorted by the time it took the worker to generate a hypothesis. It is visible that the top three conditions that took more time than the others are all conditions that include visualisations as information elements.

From figure 5.7, in which the bar plots are sorted by the overall time it took the worker to complete the task. There, it is noticeable that all the top four conditions that took the longest for the workers to complete include visualisations as information elements.

This observation is visible as well in figure 5.10 where a clear gap between the time of visual and textual information elements is visible.

With this information, the question arises whether this increase in time also leads to the TLX score to be higher in these conditions, and thus the cognitive load as well. To get more insights on this topic, it is possible to look at each single TLX survey question and compare it between the two information element conditions. In figure 5.11 all questions are enlisted, substituted by the question number. Here again are all the NASA TLX questions (TLX, 2020) enumerated:

1. How mentally demanding was the task?
2. How physically demanding was the task?
3. How hurried or rushed was the pace of the task?

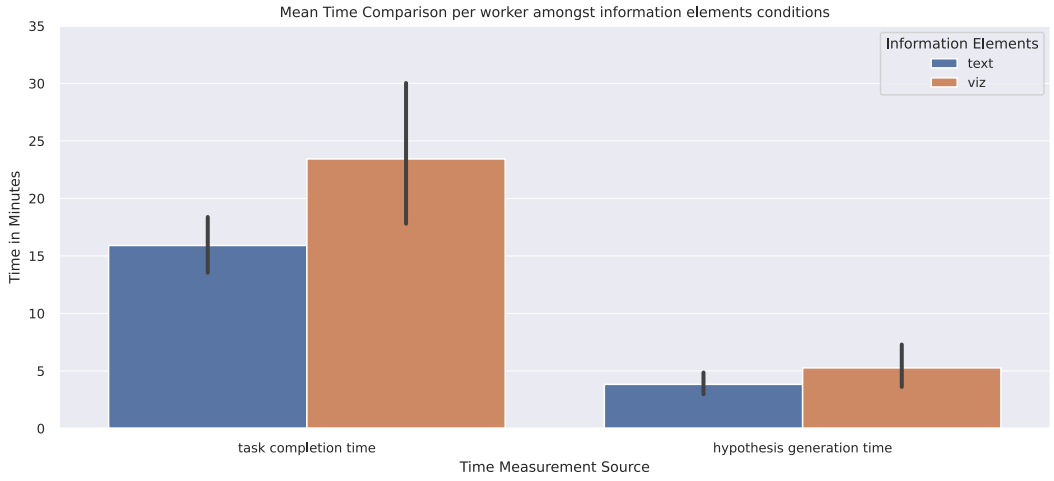


Figure 5.10: The time measurements between information elements

4. How successful were you in accomplishing what you were asked to do?

5. How hard did you have to work to accomplish your level of performance?

6. How insecure, discouraged, irritated, stressed, and annoyed were you?

From this figure (5.11), it can be seen that except for question number three, all TLX scores and thus the cognitive load is higher for visualisations as information elements. Furthermore can be observed that for all TLX scores, the overall pattern is the same for textual and visual information elements. If one is high, the other is high as well, and the same for if they are low. Interestingly is the fact that to all appearances it seems that the physical demand was higher for the worker than the mental demand, although the only physical activity required for the tasks included moving the mouse and typing on the keyboard.

Finally in figure 5.12 the two datasets and the two conversational styles are merged and only the information elements are compared. Here it is also observable that overall, visual information elements generate a slightly higher cognitive load.

Based on these results, it can be concluded for $H_{1.2}$: "Conversations using a combination of data visualisations, tables and text to convey information leads to a lower cognitive load of the crowd worker, compared to conversations without data visualisations" that this does not hold true. Pure text-based information transmission shows to have a lower cognitive load for the workers. Furthermore, it can be said that with the information elements used in the experiment, the worker takes less time to generate a hypothesis and in general to complete the task if there is only text-based information.

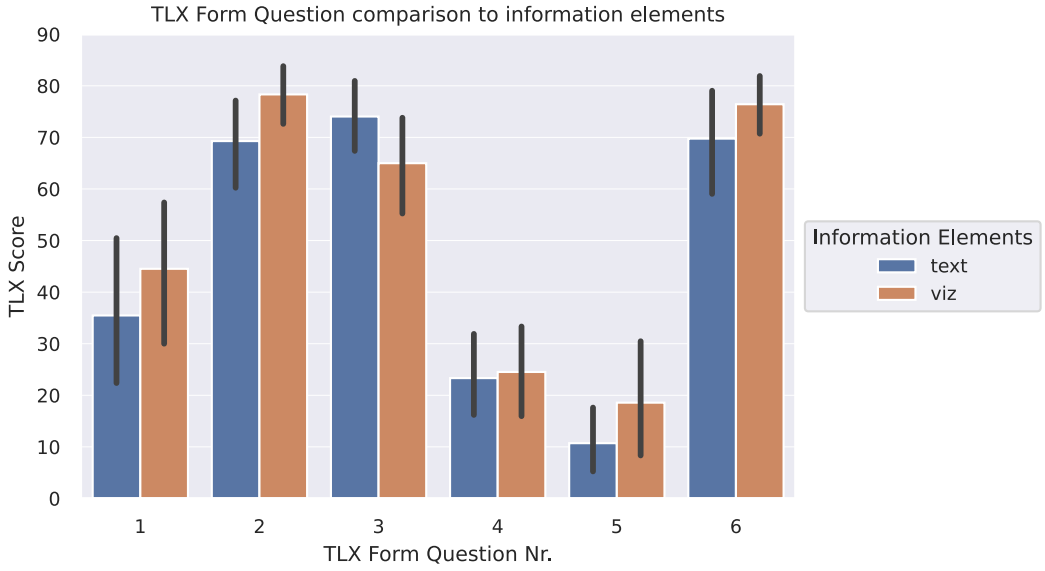


Figure 5.11: Every question of the TLX survey, compared with informational element

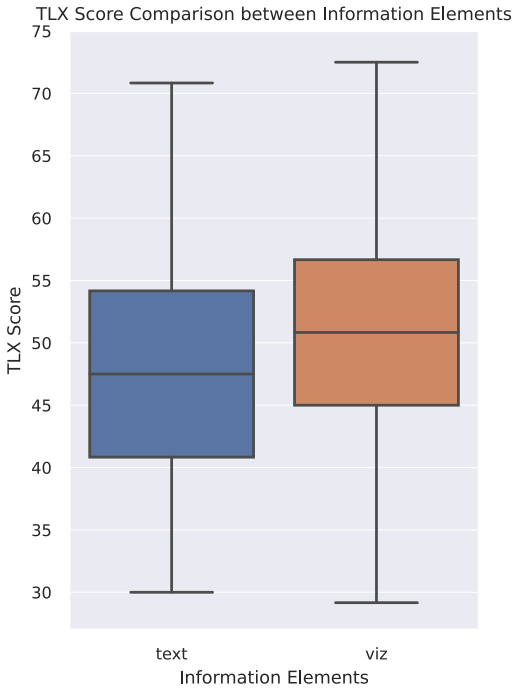


Figure 5.12: TLX score comparison of information elements, merged with all other conditions

5.4 Hypotheses Quality

In this section, the results of the experiment will be presented to compare the quality score of the hypotheses across the different conditions. The data that will be used is taken from the hypothesis rating tool. With this evaluation, $H_{1.1}$ and $H_{2.1}$ are addressed.

First, it will be looked at the statistical relevance of the data. In table 5.4 all conditions are listed where it was possible to reject the null hypothesis of the Mann-Whitney U test. These conditions have a P-value of <0.05. All other comparisons between the conditions have a greater P-value than 0.05 and thus cannot reject the null hypothesis of the Mann-Whitney U test.

Condition Name A	Condition Name B	P-Value
mach-text-1	mix-viz-2	0.049
mach-viz-1	mix-viz-2	0.033
mach-viz-2	mix-viz-2	0.006
mix-text-1	mix-viz-2	0.008
mix-viz-1	mix-viz-2	0.011

Table 5.4: Experiment conditions comparisons that reject the Mann-Whitney U test null hypothesis

If only conversational elements are compared to each other, as well as only information elements and only datasets compared to each other, the P-value of all these comparisons are >0.05. Therefore, these comparisons cannot reject the null hypothesis of the Mann-Whitney U test.

It was furthermore also not possible to reject the null hypothesis, if only information elements are compared together with conversational styles, with merged datasets. All of the P-values in these comparison resulted in >0.05.

This means that only for the comparisons enlisted in table 5.4 the median of the two compared conditions are likely to be not equal with a statistical significance.

In figure 5.13 it can be seen that for textual information elements (left group) machine-like conversational style shows a higher quality score than mixed conversational style. This is true for both datasets. For visual information elements (right group) there is a difference between the datasets visible. For dataset one (blue and yellow), the machine-like conversational style has a higher quality score than the mixed conversational style. For dataset two (green and red) the opposite is the case. In dataset two, mixed conversational style has a higher quality score than machine-like conversational style. The comparison in dataset two between mixed and machine-like conversational style in the visual information elements condition furthermore is able to reject the Mann-Whitney U test null hypothesis. However, it needs to be pointed to the P-values in table 5.4, which tells that the statistical significance of this observation is not provided. Apart from the statistical significance, it can be said for this graphic, the quality score is dependent on the dataset, the information elements, and the conversational style.

The meaning of figure 5.14 is similar to figure 5.13 discussed previously. However, the

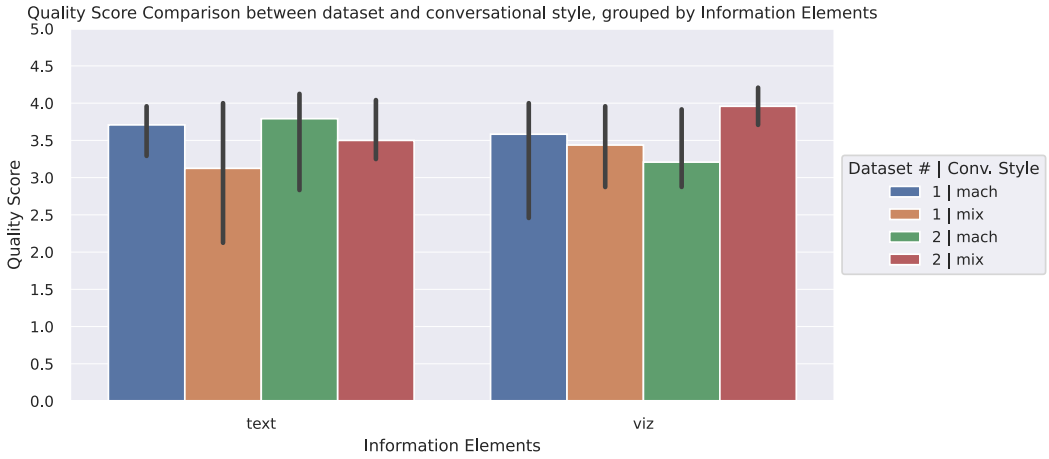


Figure 5.13: The hypothesis quality score grouped by information elements, split up by dataset and conversational style

grouping is switched from information elements to conversational style. It is observable for mixed conversational style (right group) that dataset 1 and visual information elements has a higher quality score than dataset 1 and text. The same for dataset 2, visual information elements have a higher quality score than text. For machine-like conversational style, the exact opposite is observable. For dataset 1 as well as dataset 2, the text condition has a higher quality score than visual information elements. To summarize, it is visible that different information elements affects the quality score. Which choice of information elements leads to a higher quality score is not dependent on the dataset in this figure. Thus it can be said that in contrast to figure 5.13 the observed influence on the quality score are dataset independent. However, it shows that the choice of information elements that leads to a higher quality score also depends on the conversational style.

To get a more refined picture of what is happening between the conditions, further figures are discussed here. In figure 5.15 the conditions of the experiment are shown in an ordered arrangement. Machine-like conversational style combined with textual information elements have the highest quality score. machine-like conversational style combined with visual information elements have the lowest quality score. The comparison between text and visual information elements in machine-like conversation conditions is well visible in figure 5.14 in the grouping on the left. The statements from before are again validated: If the conversational style used is machine-like conversational style, text leads to higher hypotheses quality score compared to visual information elements. However, if the conversational style used is a mixed conversational style, visual information elements scores a higher hypothesis quality compared to text. This is observable in both of the just mentioned figures.

Figure 5.16 illustrates the effects of information elements on the quality score in a similiar way as in figure 5.14. The difference between the two figures is mainly, that the

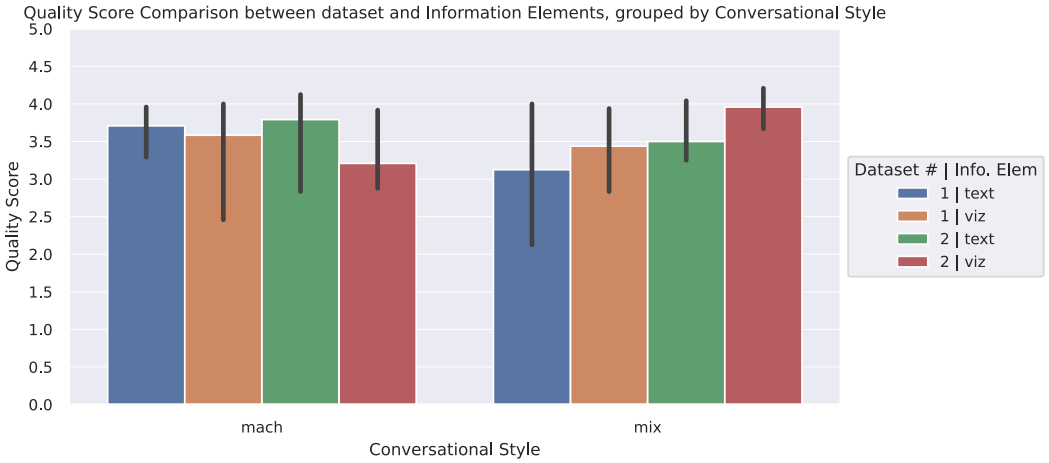


Figure 5.14: The hypothesis quality score grouped by conversational style, split up by dataset and information elements

Figure 5.15: The hypothesis quality score across all experiment conditions, datasets merged

datasets are merged in figure 5.16. By doing so, it becomes visible how the information elements influence the hypothesis quality score differently across the conversational styles. It shows that for machine-like conversational style, text based information elements achieve a higher quality score compared to visual information elements. Mixed conversational style on the other hand achieves a higher quality score when combined with visual information elements.

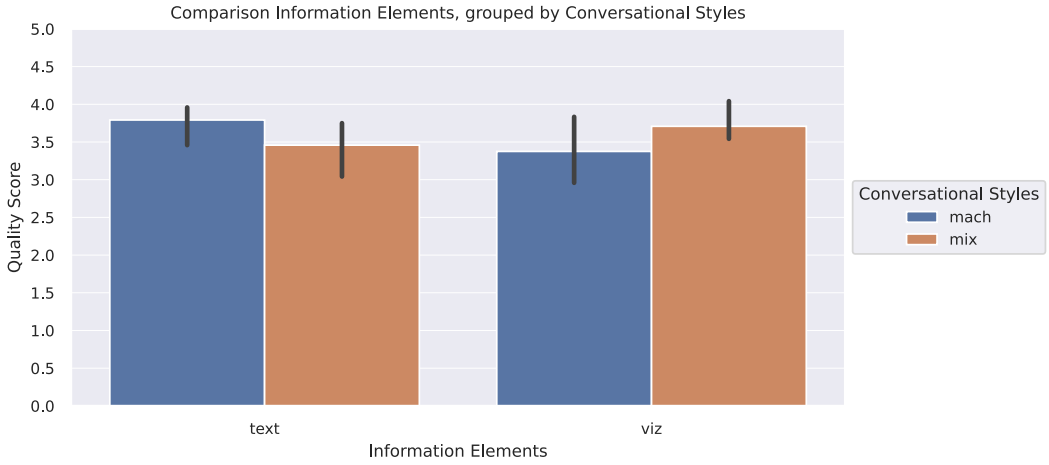


Figure 5.16: The hypothesis quality score across conversational styles, grouped by information elements, datasets merged

In figure 5.17 it is visible that visual information element when combined with mixed conversational style achieves a higher quality score than all other conditions. It is not possible to conclude that this would origin from the visual information elements. Because when looking at the machine-like conversational style, visual information elements score lower. Even more, for visual elements combined with machine-like conversational style the resulting quality score is the lowest overall visible in this figure.

When comparing only conversational styles to each other, as in figure 5.18, the following can be observed. The median quality score for machine-like conversational style is at 3.58 and for mixed conversational style it is a little bit better with 3.63. This is a small difference of just 0.05 quality score points between the two. Combined with the results from the Mann-Whitney U test, it leads to the reasoning that in this overall comparison, there is no significant difference between the conversational styles visible that would affect the hypothesis quality score.

In figure 5.19, the same inconclusive behaviour as previously discussed is visible. This time it is for information elements. Textual information elements reach a median quality score of 3.58. Visual information elements performed slightly better and reaches a median quality score of 3.65. The difference between the two is again negligibly small—just 0.07 quality score points. This supports the claim that the results are inconclusive when taken together with the Mann-Whitney U test results. This indicates that there is no significant difference in the information elements in this overall comparison that would

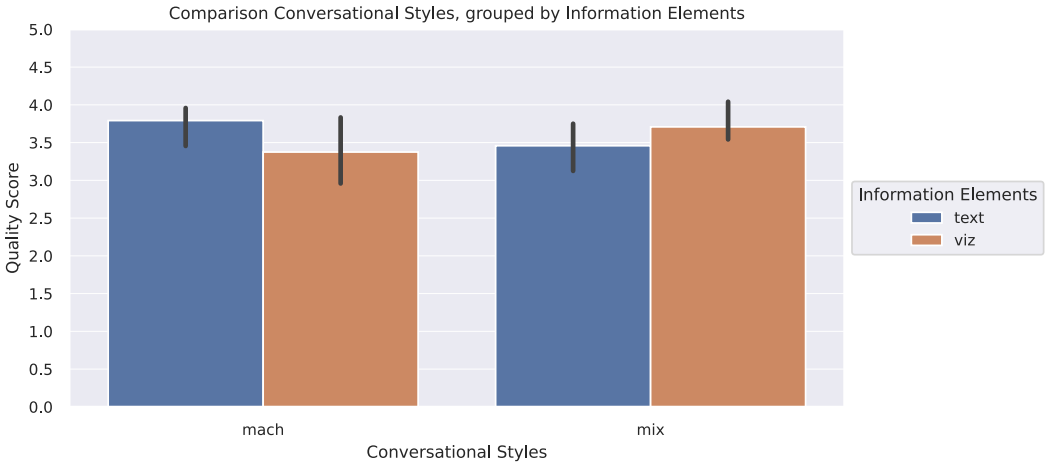


Figure 5.17: The hypothesis quality score across information elements, grouped by conversational styles, datasets merged

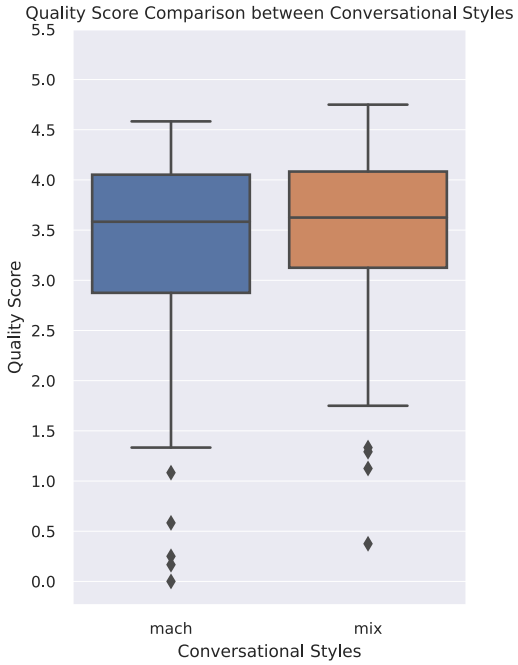


Figure 5.18: The overall hypothesis quality score between conversational styles

have an impact on the score for the quality of the hypothesis.

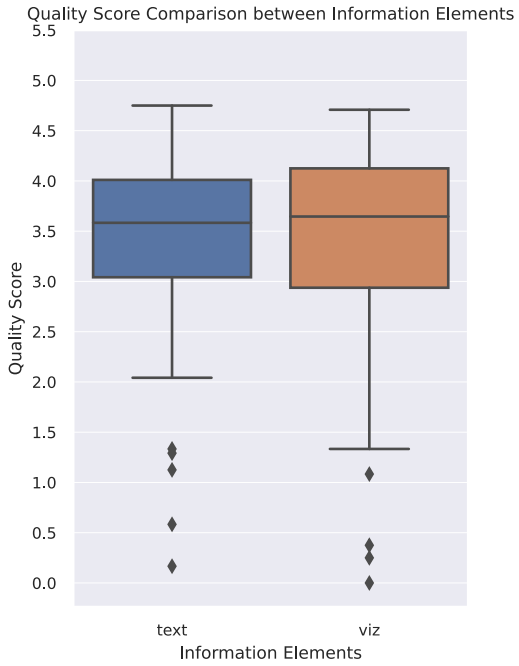


Figure 5.19: The overall hypothesis quality score between information elements

To compare, which dataset would lead to a higher quality score overall, figure 5.20 can be used. Dataset 1 has a median quality score of 3.50. Dataset 2 performed better and has a median score of 3.69. The reason behind this performance increase could be that dataset two contains more numbers as data, while dataset 1 mostly contains binary or categorical answers to survey questions. It might be the case that the workers may have found it easier to generate hypotheses with a dataset that is more based on numbers. Here again, the Mann-Whitney U test results are not able to reject the null hypothesis. Thus, the difference in the median between the two datasets is not significant.

To answer $H_{1.1}$ "Conversations using a combination of data visualisations, tables and text to convey information improve the quality of hypotheses generated in a chat-based interface, compared to conversations without data visualisations" it can be said that solely from the median comparison between the information elements as in figure 5.19, the answer would be yes, visual information elements, tables and text lead to a higher quality score. Together with that answer, there are strong limitations due to the small difference and the failed rejection of the null hypothesis of the Mann-Whitney U test. However, the analysis presented here provides insights beyond this overall comparison. While taking into account the limitations mentioned, it can be concluded that $H_{1.1}$ holds true when using the condition with mixed conversational style. There, visual information elements tend to lead to a higher quality score. Having said that, $H_{1.1}$ must be answered with no when looking at mixed conversational style. There, text based

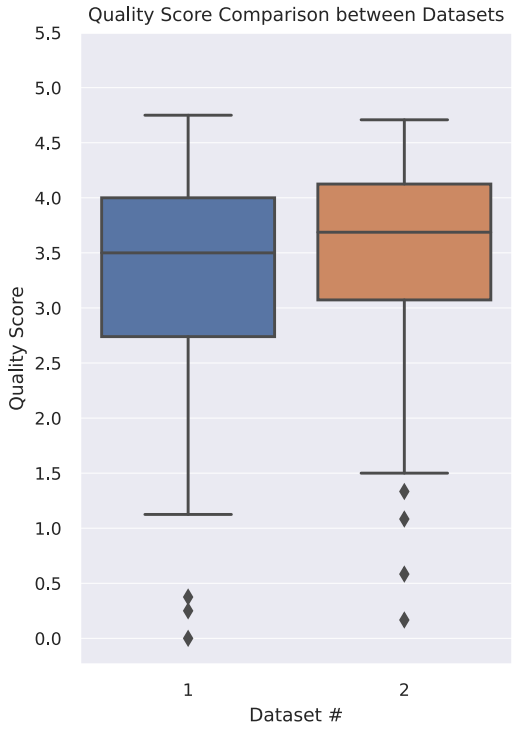


Figure 5.20: The overall hypothesis quality score between datasets

information elements tend to lead to a higher quality score. (see fig. 5.14 and 5.17)

For $H_{2.1}$ "Conversations using a humanlike conversational style for non-informational discussion and machine-like conversational style for presenting information improve the quality of hypotheses generated in a chat-based interface, compared to conversations with only machine-like conversational style", the following conclusion can be made based on the presented results. Solely based on the overall median comparison between the two conversational styles as visible in figure 5.18 the hypothesis can be answered with yes, as mixed conversational style leads to a higher quality score. However, the difference between the two medians is marginal and the Mann-Whitney U test shows that there is no significant difference. When differentiating this overall result by the available conditions, the following can be said. For conditions where text based information elements are used, $H_{2.1}$ does not hold true. That is because it shows that in these conditions machine-like conversational style leads to a higher hypothesis quality score. For conditions that include visual information elements, it depends on the dataset whether $H_{2.1}$ holds true or not. For dataset 1 with visualisations $H_{2.1}$ does not hold true, machine-like conversational style gives higher quality scores. While for dataset 2 with visualisations $H_{2.1}$ holds true, mixed conversational style shows to result in a higher hypothesis quality score (see fig. 5.13 and 5.16).

Figure 5.21: The hypothesis quality score across all quality formulas, grouped by experiment conditions, datasets merged

As a final step in the result section, a comparison between the different formulas used to compute the quality score is presented in figure 5.21. It can be said that the quality score resulting from the formula by Quinn and George is always higher than the quality score by Willet. This happens due to the fact that the formula by Quinn and George contains more binary values (3 in total) that are directly scaled up to a 5 if they are true. The formula by Willet contains scales from 1 to 5 that contribute towards the final score and thus has a higher granularity.

6

Discussion

In this chapter, the findings of the results will be discussed. Three point of views will be taken for this. First, the reflection on information elements, followed by the reflection on the conversational style. As a third point, the implications on cognitive load and hypotheses quality will be discussed.

6.1 Reflection on Information Elements

Through the experiment carried out in this thesis, it was discovered that visualisations do not necessarily lead to a lower cognitive load for crowd workers. It was shown that a text-based representation of information in a chat-based environment can have a lower cognitive load than a representation of information that includes visualisations. The thoughts on this are that it might be the case that a visualisation takes more time to study and fully understand rather than just reading through a text. The cognitive load might be higher based on the extra work a visual element causes.

For the quality score, the experiment has uncovered that visualisations may lead to higher-quality hypotheses. However, this conclusion is dependent on the chosen dataset and the chosen conversational style. Only a small portion of the comparison of the results from the experiment condition can reject the Mann-Whitney U test null hypothesis. Thus, it is necessary to point out that the comparisons in the results presented here may not have enough statistical power to draw effective conclusions to answer the initial hypotheses.

6.2 Reflection on Conversational Style

For the cognitive load part, the results did not yield a conclusive answer on which condition would clearly lead to a lower cognitive load for crowd workers. When comparing the boxplots of both styles of the TLX mean score, the machine-like conversational style is shown to be slightly higher than the mixed conversational style. The first quartile is also higher compared to the mixed style. However, all the other parts, the minimum, maximum, and third quartile, are lower than in the mixed style. Looking into other comparisons, as the result section has shown, is inconclusive. However, some insights

from these in-depth comparisons show that the conversational style, in dependency with the information elements presented, could impact the NASA TLX score and, thus, the cognitive load of the workers.

A mixed conversational style can, like visual information elements, contribute to higher-quality hypotheses. However, upon unfolding the analysis and comparing each of the conditions to each other shows that the mixed conversation style will only lead to higher quality hypotheses when visual information elements and specific datasets are used. If these conditions are not met, a machine-like conversational style is shown to result in higher-quality hypotheses. The null hypothesis of the Mann-Whitney U test could be rejected for most comparisons between the combination *dataset2 - mixed conversational style - visual information elements* and the other experimental conditions, which might give a hint that it is possible to strengthen the analysis of this thesis in the future with a statistically significant experiment.

6.3 Implications on Cognitive Load and Hypotheses Quality

It can be said within the limitations this thesis shows that a textual representation of information can be used to lower the cognitive load of crowd workers in a chat-based interface with tasks like the ones presented during the experiment of this work.

To come back to the initial research questions RQ_1 and RQ_2, for both it is possible to say that there is a difference observable. However the statistical significance of this difference is not given.

To reach an improved hypothesis quality, this thesis suggests using visual information elements, together with tables and text and combined with a mixed conversational style. The dataset used can and should be tested to determine whether it influences the quality score. Dependent on that, it may be switched to information elements containing text only and/or a machine-like conversational style.

7

Limitations

In this chapter, the limitation of this research is presented. Three different viewpoints will be taken for this. The construct validity will look at the overall concept and how the project and experiment were structured. The internal validity will look at the vailidity of the data and the results drawn from them. At last, in external validity it will be looked at how generalizable the findings of this work are.

7.1 Construct Validity

The concept of the thesis was well defined in its proposal. The hypotheses that were defined in this concept were the initial key to structure the experiments. After this initial phase, the hypotheses were adapted and refined to match exactly the tested conditions in the experiment.

To support construct validity further, multiple mesaurements to answer the hypothesis were implemented. For the cognitive load, the time measurements as a hard fact and the NASA TLX survey as proven tool to estimate cognitive load were used. A threat that is possible limiting the validity of the results on hand is the quality definition of a hypothesis. There are sources available that have quite a different view on how the criterias could look like. To countersteer this threat as good as possible, for the hypothesis quality, two different formulas from two peer reviewed papers to measure the quality of hypotheses were used. The criterias that define these two formulas are unique for each formula. The criterias were filled out by qualified raters that were considered experts on the field of rating the quality criterias of hypotheses. To strengthen the validity of this rating, each criterion was rated by multiple individual raters separately.

7.2 Internal Validity

The main threats for internal validity is the sample size, the statistical relevance and the complexity of the tasks in the experiment. The amount of workers from MTurk recruited to generate hypotheses was limited through budget, time and scope of the project. By limiting this sample size, there is an automatic threat to internal validity, as the sample size is not statistically relevant. A further threat is the complexity of

the tasks, which has an interplay with the environment the experiment took place in. The chat based interface shows many different messages to many individual subjects. The amount of chat messages directly raises the possibility that one or more messages could be interpreted from one person to another in different ways. Once a message is interpreted by a worker in a certain way, it will affect the perception of any messages that follows. This creates an unique experience for every worker and this experience could lead to a noise factor which could bias the results and make it harder to see the effect of the actual tested conditions. Another threat is the order effect in the rating tool. It might be the case that people have rated 5 bad hypotheses. After that, they receive a mediocre hypothesis. This mediocre hypothesis might get a much higher rating, because the people are biased from what they have seen previously. People who rate the hypotheses mighe get tired or unmotivated after rating a certain amount of hypotheses. This could have an influence on the quality of the received ratings.

7.3 External Validity

The generalizability of these results is dependent on the field of application. While these results can be used to generate hypotheses of different quality in a chat based interface, with different levels of cognitive load for the workers, it might be applicable elsewhere too. There is a connection between visualisations versus textual information representation and the cognitive load visible from the results, that could possible be used in other fields of research as well. A factor that strengthen the external validity is the fact that people from all around the world could participate through MTurk in the experiment. What poses threats to this validity is the language barrier, because the experiments were only available in English. Furthermore the workers had to be people who own a computer or other device that is capable of completing task in the MTurk environment, therefore the device also must have a connection to the internet. Furthermore, the workers must have an MTurk account, and thus have some level of digital affinity. All these points account for the limiting factor of the generalizability of the presented results.

8

Future Work

The experiment and the analysis that was conducted in this project gives a basic platform to build upon and conduct further research. To strengthen the validity of the results presented in here and to give it statistical relevance, the same procedure as shown in this experiment can be done but on a larger scale. It would be also a good option to introduce more datasets, on which the different conversational styles and information elements will be tested. This will result in the possibility of a larger comparison across the conditions and may help to minimize the influence of the dataset content on the result.

A next step that could be done is to try and reduce the complexity, as it is described in the limitations. Here, a possibility would be to either try the experiment with less chat messages. The amount of chat messages raises the possibility that one or more messages could be interpreted from one person to another in different ways, and thus creating a possible external influencing factor through this. Another way to reduce complexity is to move away from the chat based approach and try another form of communication, while testing the different conversational styles and information elements.

Also an important step would be to statistically make sure that the complexity of each condition is the same across the whole experiment, rather than just relying on pilot studies and think-aloud sessions as in this project. This would require extensive testing of each feature in each condition, meaning each chat message, each message chain and the interaction between them, and each information element such as visualisation needs to be evaluated with a statistically relevant amount of participants and make sure that their level of complexity stays the same.

Furthermore the quality formula used in this paper could be improved. This, as the other mentioned points above, could be a project on its own: to find a formula that holds reasonable criterias to evaluate the quality of hypotheses on a general, interdisciplinary and data-independent level.

A true understanding of why textual information elements in a chat based environment leads to lower cognitive load could not be reached through this project and the analysis of the conducted experiment. A next project could dive deeper in this niche and explore the true connection between cognitive load and textual information elements.

9

Conclusions

The thesis focuses on the utility of conversational crowdsourcing in the complex task of generating data science hypotheses. The study looked at how conversational styles and informational elements affected crowd workers and the quality of their work on the hypothesis generation task that used a chat-based interface. In this task, the workers had to generate hypotheses about a specific dataset presented to them. The goal was to discover which conversational styles and informational elements would lead to a higher quality of generated hypotheses and a lower cognitive load for the workers. An experiment was conducted to reach this goal. In the experiment, two conversational styles and two combinations of information elements were evaluated across two datasets.

The conversational styles tested were "mixed conversational style," a style that uses human conversational elements such as emojis and a friendly and appreciative voice for talking with the worker. In this "mixed conversational style," the voice does not use human conversation elements when conveying crucial information about the dataset to the worker. In contrast to this style, there is the "machine-like conversational style," which does not rely on the previously mentioned human conversation elements.

The two sets of information elements used in this experiment are the following. The first is abbreviated to "text" or "textual information elements" for convenience. It uses text and tables to convey information about the dataset. The second is abbreviated to "viz" or "visual information elements." It contains the same information elements as "text" but is enriched by visualisations about the data.

In the results, only a small selection of comparisons across the eight different experimental conditions show statistical significance. Thus, with many of the observed differences, it is not possible to answer conclusively the hypotheses stated in the introduction of this thesis.

The results of the thesis may give the following indications about the effect of information elements: (1) Text-based information elements have a lower cognitive load on the worker. (2) Text-based information elements can lead to higher-quality hypotheses when combined with a machine-like conversation style. And (3) visual information elements can lead to higher-quality hypotheses, when combined with a mixed conversational style.

The indications from the results of this thesis for the conversational styles are as follows: (1) A mixed conversational style can lead to a lower cognitive load for the crowd worker. (2) A mixed conversational style can also lead to a higher quality score when

combined with visual information elements.(3) A machine-like conversation style can lead to higher-quality hypotheses when combined with text-based information elements.

It was also demonstrated in the results section that all of the aforementioned indicators are dependent on other conditions. That means that the effect of information elements might be dependent on a certain conversation style, and vice versa.

The statistical significance and interdependencies between the different condition combinations must be noted for all the conclusions drawn. The interdependencies mean that one experimental condition might hold such individuality that it becomes hard to compare individual conditions to each other. This is also mentioned in the limitations. There, it was suggested that as a first step, the experiment may be repeated with more unique datasets to minimise the effect of the datasets on each condition.

Bibliography

2014. Mental Health in Tech Survey Kaggle. *https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey* Retrieved 2022-12-01.

2018a. Amazon Mechanical Turk. *https://www.mturk.com/* Retrieved 2022-12-01.

2018b. Amazon Mechanical Turk. *https://www.mturk.com/privacy-notice* Retrieved 2022-12-01.

2018. Ghana Health Facilities Kaggle. *https://www.kaggle.com/datasets/citizen-ds-ghana/health-facilities-gh?select=health-facility-tiers.csv* Retrieved 2022-12-01.

2019. Google Play Store Apps Kaggle. *https://www.kaggle.com/datasets/lava18/google-play-store-apps* Retrieved 2022-12-01.

2020. TLX @ NASA Ames - Home. *https://humansystems.arc.nasa.gov/groups/TLX/* Retrieved 2022-12-01.

2020. Vega-Altair: Declarative Visualization in Python — Altair 4.2.0 documentation. *https://altair-viz.github.io/* Retrieved 2022-12-01.

2020. World Health Statistics 2020CompleteGeo-Analysis Kaggle. *https://www.kaggle.com/datasets/utkarshxy/who-worldhealth-statistics-2020-complete* Retrieved 2022-12-01.

2022. Cross-Origin Resource Sharing (CORS) - HTTP MDN. *https://developer.mozilla.org/en-US/docs/Web/HTTP/CORS* Retrieved 2022-12-01.

2022. Kaggle: Your Machine Learning and Data Science Community. *https://www.kaggle.com/* Retrieved 2022-12-01.

2022. Plotly Python Graphing Library. *https://plotly.com/python/* Retrieved 2022-12-01.

2022. QualificationRequirement - Amazon Mechanical Turk. *https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/ApiReference_QualificationRequirementDataStructureArticle.html#ApiReference_QualificationType-IDs* Retrieved 2022-12-01.

2022. /r/SampleSize: Where your opinions actually matter! *https://www.reddit.com/r/SampleSize/* Retrieved 2022-12-01.

Sungeun An, Robert Moore, Eric Young Liu, and Guang Jie Ren. 2021. Recipient Design for Conversational Agents: Tailoring Agent's Utterance to User's Knowledge. *ACM International Conference Proceeding Series* (7 2021). *https://doi.org/10.1145/3469595.3469625*

Sonia Baee, Mark Rucker, Anna Baglione, Mawulolo K Ameko, and Laura Barnes. 2020. A Framework for Addressing the Risks and Opportunities In AI-Supported Virtual Health Coaches. *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare* (2020). *https://doi.org/10.1145/3421937*

Gaoping Huang, Meng Han Wu, and Alexander J. Quinn. 2021. Task Design for Crowdsourcing Complex Cognitive Skills. *Conference on Human Factors in Computing Systems - Proceedings* (5 2021). *https://doi.org/10.1145/3411763.3443447*

Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. CrowdForge: Crowdsourcing complex work. *UIST'11 - Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (2011), 43–52. *https://doi.org/10.1145/2047196.2047202*

Sandeep Kaur Kuttal, Jarow Myers, Sam Gurka, David Magar, David Piorkowski, and Rachel Bellamy. 2020. Towards Designing Conversational Agents for Pair Programming: Accounting for Creativity Strategies and Conversational Styles. *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC* 2020-August (8 2020). *https://doi.org/10.1109/VL/HCC50065.2020.9127276*

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (3 1947), 50–60. *https://doi.org/10.1214/AOMS/1177730491*

Lars Müller, Thomas Wetzel, Hans Christoph Hobohm, and Thomas Schrader. 2012. Creativity support tools for data triggered hypothesis generation. *Proceedings - 2012 7th International Conference on Knowledge, Information and Creativity Support Systems, KICSS 2012* (2012), 24–27. *https://doi.org/10.1109/KICSS.2012.12*

Heather L. O'Brien and Elaine G. Toms. 2010. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology* 61, 1 (1 2010), 50–69. *https://doi.org/10.1002/ASI.21229*

Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It's How You Say It: Identifying Appropriate Register for Chatbot Language Design. *Proceedings of the 7th International Conference on Human-Agent Interaction* (2019). *https://doi.org/10.1145/3349537*

Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020a. Estimating Conversational Styles in Conversational Microtask Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (5 2020). *https://doi.org/10.1145/3392837*

Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020b. Improving Worker Engagement Through Conversational Microtask Crowdsourcing. *Conference on Human Factors in Computing Systems - Proceedings* (4 2020). *https://doi.org/10.1145/3313831.3376403/FORMAT/PDF*

Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020c. TickTalkTurk: Conversational Crowdsourcing Made Easy. *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing* (10 2020). *https://doi.org/10.1145/3406865*

Mary Ellen Quinn and Kenneth D. George. 1975. Teaching hypothesis formation. *Science Education* 59, 3 (7 1975), 289–296. *https://doi.org/10.1002/SCE.3730590303*

Jorge Ramírez, Burcu Sayin, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, and Gianluca Demartini. 2021. On the State of Reporting in Crowdsourcing Experiments and a Checklist to Aid Current Practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021). *https://doi.org/10.1145/3479531*

Daniela Retelny, Michael S. Bernstein, and Melissa A. Valentine. 2017. No workflow can ever be enough: How crowdsourcing workflows constrain complex work. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (11 2017). *https://doi.org/10.1145/3134724*

Petra Schneider, W. Patrick Walters, Alleyn T. Plowright, Norman Sieroka, Jennifer Listgarten, Robert A. Goodnow, Jasmin Fisher, Johanna M. Jansen, José S. Duca, Thomas S. Rush, Matthias Zentgraf, John Edward Hill, Elizabeth Krutoholow, Matthias Kohler, Jeff Blaney, Kimito Funatsu, Chris Luebkemann, and Gisbert Schneider. 2019. Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery 2019 19:5* 19, 5 (12 2019), 353–364. *https://doi.org/10.1038/s41573-019-0050-3*

S S Shapiro and Aotj M B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3-4 (12 1965), 591–611. *https://doi.org/10.1093/BIOMET/52.3-4.591*

Micol Spitale and Franca Garzotto. 2020. Towards Empathic Conversational Interaction. *Proceedings of the 2nd Conference on Conversational User Interfaces* (7 2020), 1393–1400. *https://doi.org/10.1145/3405755*

Antonela Miruna Stan. 2020. *Talking with chatbots : the influence of visual appearance and conversational style of text-based chatbots on UX and future interaction intention - University of Twente Student Theses.* Ph. D. Dissertation. *http://essay.utwente.nl/83151/*

Student. 1908. The Probable Error of a Mean. *Biometrika* 6, 1 (3 1908), 1. *https://doi.org/10.2307/2331554*

Menasha Thilakaratne, Katrina Falkner, and Thushari Atapattu. 2019. A Systematic Review on Literature-based Discovery. *ACM Computing Surveys (CSUR)* 52, 6 (12 2019). *https://doi.org/10.1145/3365756*

Ruilin Wang, Somayajulu Gowri Sripada, and Nigel A Beacham. 2021. Auto-generating Textual Data Stories Using Data Science Pipelines. *2021 4th International Conference on Algorithms, Computing and Artificial Intelligence* (12 2021), 1–8. *https://doi.org/10.1145/3508546.3508642*

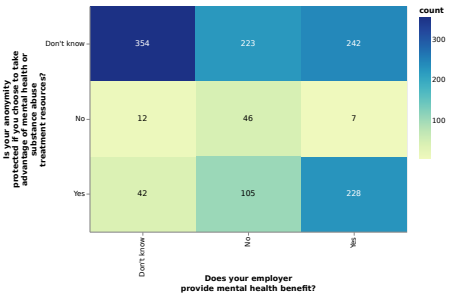
Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2012. Strategies for crowdsourcing social data analysis. *Conference on Human Factors in Computing Systems - Proceedings* (2012), 227–236. *https://doi.org/10.1145/2207676.2207709*

A

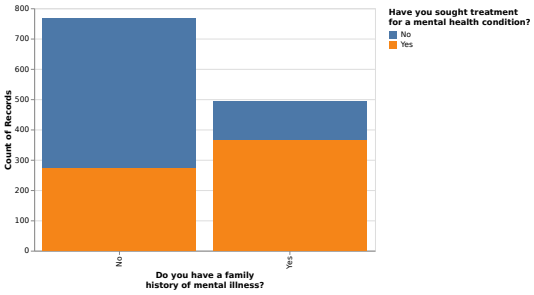
Experiment Resources

A.1 Data Visualisations

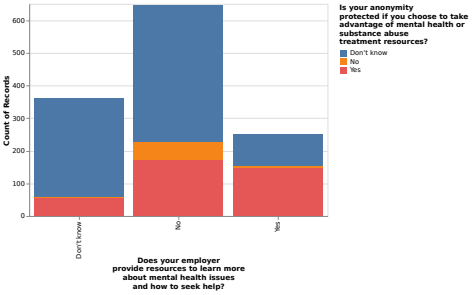
The visualisations in figures A.1 and A.2 were used in the experiment:



(a) Visualisation 1



(b) Visualisation 2

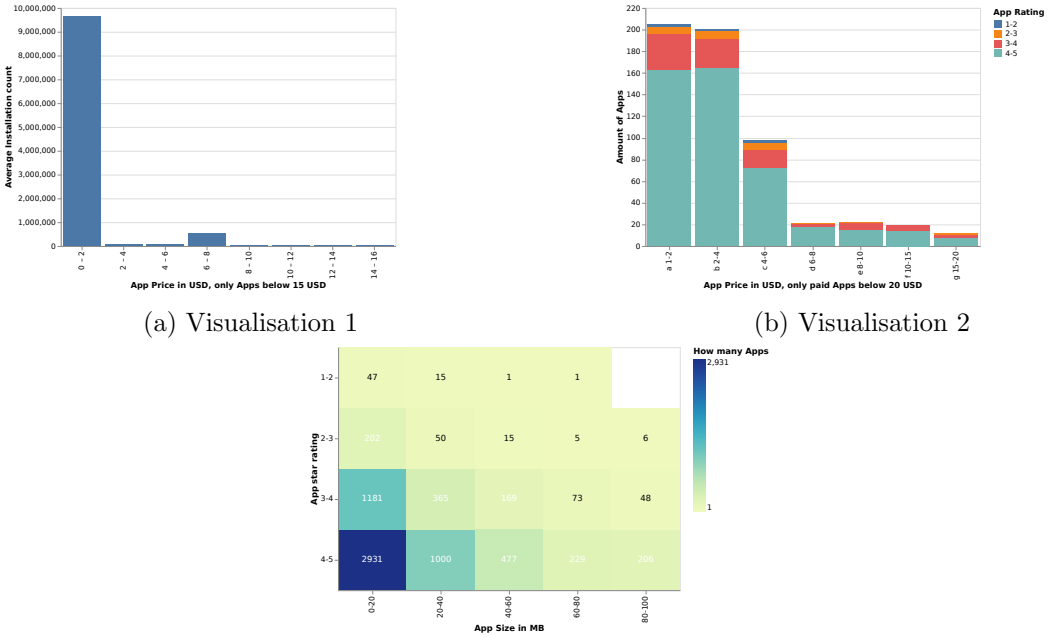


(c) Visualisation 3

Figure A.1: Dataset 1 Visualisations

A.2 Best and Worst Hypotheses

In table A.1 are the 10 best and the 10 worst hypotheses, together with their quality scale.



(a) Visualisation 1

(b) Visualisation 2

(c) Visualisation 3

Figure A.2: Dataset 2 Visualisations

A.3 Gitlab Project URL

The whole project is on the UZH IFI Gitlab server. The URL is the following:
https://gitlab.ifi.uzh.ch/ddis/Students/Theses/2022-Emanuel-Graf

Hypothesis	Mean Quality Score
Integrity and Discipline are to be act in line with our values ,even when it feels uncomfortable.	0.0
the lower the overall rating of the product.	0.16666666666666666
It both help to understand where you came from and where you wnat to go wit theraphy and beyond.	0.25
The higher the version number of the product	0.5833333333333334
Certainly the graphs and spreadsheets demonstrated deepen a research rich in data and new studies.	1.0833333333333333
A person might want to be anonymous and not speak with their coworkers	1.125
A person that knows about a wellness program and about how to seek help would also imply that you know what your care options are.	1.2916666666666667
Having a mental health condition leads to an increase in mental health consequences.	1.3333333333333333
the largest being between 1-2 USD.	1.3333333333333333
and the lowest download fee are apps between 14-16 USD.	1.5
...	...
less number of installation caused less number of ratings	4.458333333333333
the more installs an app has, the more reviews it will have	4.5
The number of installs of an app directly impacts the number of times it is rated	4.5
Review has a strong positive correlation with both Rating and Last Update.	4.5
more number of app installation increased number of reviews	4.541666666666667
The price of the app has a great impact on the installation count	4.583333333333333
Those with no history of mental illness in the family are less likely to seek treatment	4.583333333333333
Apps with a higher amount of installations are more likely to have more reviews	4.666666666666667
Cheaper apps are more likely to have a higher number of installations	4.708333333333333
Employees who work for an employer that provides mental health benefits are more likely to seek treatment.	4.75

Table A.1: The 10 worst and the 10 best hypotheses

B

Crowdsource Checklist

The checklist that was defined by Ramírez et al. (2021) was used for the experiment carried out in this thesis. The checklist is visible on the following three pages.

If the paper reports on multiple studies involving the crowd as subjects, fill out the checklist per experiment. Besides, if an experiment uses different (potentially interconnected) micro-tasks, then report the Task and Quality control sections for each task.

Item	Item No.	Recommendation	Section Title
Experimental design			
Input dataset	1	Describe how the input dataset for the experiment was obtained and if it is publicly available. Also, touch on its reputation and difficulty (if applicable)	"Dataset Selection"
Allocation to experimental conditions	2	State how the participants were assigned to the experimental conditions or treatments, and how this step was implemented in the crowdsourcing platform	Worker Assignment to Task
Experimental design to task mapping	3	Describe what research design was used in the experiment and how were the experimental conditions mapped to crowdsourcing tasks	"Experiment Combinations"
Execution of experimental conditions	4	Report how the crowdsourcing tasks, representing the experimental conditions, were executed (e.g., in parallel, sequentially, or mixed)	Execution of the experiments
Execution timeframe	5	State over what timeframe the experiment was executed	Execution
Pilots	6	Describe if pilot studies were performed before the main experiment	of the experiments
Returning workers	7	Report the strategies used to prevent returning workers, i.e., workers who finish the experiment and then reenter it later because the study was still running	Quality Control
Crowd			
Target population	8	Describe the criteria used to determine the workers who are allowed to participate (e.g., acceptance rate, tasks completed, demographics, working environment). And also include the strategy used to identify such workers.	Quality Control
Sampling mechanism	9	Report what strategies were used to recruit a diverse or representative set of workers from the target population	Quality Control
Task			
Task interface	10	Report and show the task interface as seen by workers	Task Interface
Task interface source	11	Provide a link to an online repository with the source code of the task interface (typically a combination of HTML, CSS, and JavaScript)	Gitlab Project URL
Instructions	12	Describe and show the instructions of the task as seen by workers	Task Interface
Reward strategy	13	State the mechanisms used to reward and motivate workers (e.g., payments)	Rewards
Time allotted	14	Report if a time constraint was defined for workers to complete the task (if so, describe also how much)	Rewards

If the paper reports on multiple studies involving the crowd as subjects, fill out the checklist per experiment. Besides, if an experiment uses different (potentially interconnected) micro-tasks, then report the Task and Quality control sections for each task.

Item	Item No.	Recommendation	Section Title
Quality control			
Rejection criteria	15	State the criteria used to accept or reject a contribution from a worker (e.g., workers can be allowed to submit the task and reject it afterward, submissions can be blocked based on prior rejections or on time spent on the task)	Quality Control
Number of votes per item	16	Describe, if applicable, how many workers solved the same item or data unit	Experiment Conditions
Aggregation method	17	Report, if applicable, how the contributions from workers were aggregated (e.g., majority voting)	Hypothesis Quality Measurement
Training	18	State if workers performed a training session or pre-task qualification test. If so, describe 1) the training, 2) the items used as the training set, and 3) if it was performed before or as part of the task	Quality Control
In-task checks	19	Report the mechanisms embedded in the task to guard the quality of the results. Also, state if and how workers were allowed to revise their answers. In case gold items or attention checks were used, describe how these items were selected, how frequently they appear, and the threshold used to filter out workers underperforming on these items.	Force workers to keep bigger picture
Post-task checks	20	Report the steps performed upon task completion to safeguard the quality of the results (e.g., post hoc analysis)	Quality Check
Dropouts prevention mechanisms	21	Indicate the strategies used to deal with worker dropouts (i.e., workers who leave the task unfinished)	Worker Assignment to Task
Outcome			
Number of participants	22	Indicate how many workers participated in the experiment (in total and per condition)	Results
Number of contributions	23	Report the number of contributions (e.g., votes) in total and per condition	
Excluded participants	24	Indicate the number of participants not considered for the data analysis, including the reason for exclusion.	
Discarded data	25	State the number of contributions excluded before the data analysis	
Dropout rate	26	Describe the dropout rate of the participants in the experimental conditions. If applicable, also show breakdowns per milestone of progress within the task (e.g., after 2, 3, and 5 questions).	
Participant Demographics	27	Report the demographics of the participants (e.g., age, country, language)	Quality Control

If the paper reports on multiple studies involving the crowd as subjects, fill out the checklist per experiment. Besides, if an experiment uses different (potentially interconnected) micro-tasks, then report the Task and Quality control sections for each task.

Item	Item No.	Recommendation	Section Title
Data processing	28	Report any data transformation, augmentation, and/or filtering step performed on the raw dataset obtained from the crowdsourcing platform.	Pipelines and data preprocessing
Output dataset	29	Provide a link to the dataset resulting from the experiment. Also, indicate if the dataset contains the aggregated or individual contributions from workers	Gitlab Project URL
Requester			
Platform(s) used	30	Indicate the crowdsourcing platform(s) selected for the experiment	Workers
Implemented features	31	Report any additional feature implemented to support the experiment, covering missing functionality from the selected platform(s)	Quality Control
Fair compensation	32	State whether workers were compensated fairly and according to legal minimum wage	Rewards
Requester-Worker interactions	33	Describe concrete requester-worker interactions taking place as part of the experiment	
Privacy & Data Treatment	34	Report any relevant privacy regulations and methods used to comply, especially if the output is put online (e.g., the data could be anonymized to meet privacy policies).	Introductional Popup
Informed consent	35	Indicate if an informed consent was used	
Participation awareness	36	State if workers were informed they took part in an experiment	
Ethical approvals	37	Report if the study received ethical approval from the corresponding institutional authority	Quality Control

List of Figures

List of Tables