Much Ado about Accessibility: Exploration of online information seeking tools' responses to Autistic users

Msc Thesis Computer Science & Engineering

Hrishita Chakrabarti

TU Delft

Much Ado about Accessibility

Exploration of online information seeking tools' responses to Autistic users

Thesis report

by

Hrishita Chakrabarti

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on 06 August 2024.

Thesis committee:

Chair:	Sole Pera
Supervisor:	Sole Pera
External examiner:	Jorge Martinez Castaneda
Place:	Faculty of Electrical Engineering, Mathematics, Computer Science, Delft
Project Duration:	20 November 2023 - 05 August 2024
Student number:	5759048

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Faculty of EEMCS · Delft University of Technology

Abstract

Autism Spectrum Disorder (ASD) is a neuro-developmental disorder reported to affect around 58 million people across the world. Due to their struggles with social communication in real life, they prefer to use online mediums to form social connections and are better able to maintain meaningful relationships through their communications over these online mediums. However, empirical studies have revealed that autistic users are not as efficient in using these online mediums when they have to search for information. While studies in the past have resulted in a growing list of guidelines to improve the web search experience of autistic users, whether existing online ISTs adhere to these guidelines has not been studied extensively. Thus while we know what the autistic users need to efficiently look for information on the web, we do not know if and how the ISTs are catering to their accessibility needs.

To advance knowledge in the accessibility of popular online ISTs when catering to the information needs of autistic users, we conducted an empirical exploration focusing on the accessibility of the textual information in the IST responses. We specifically examined the responses on three aspects of text accessibility for autistic users - (1) text structure, (2) text readability, and (3) text concreteness using "accessibility indicators" i.e. quantifiable features created by us based on widely accepted web content accessibility guidelines for autistic users. Due to the lack of standard query datasets for autistic users, we had to generate synthetic queries representative of autistic users' information needs to collect responses from four popular ISTs Google, Bing, Gemini, and ChatGPT. We examined the differences in the way ISTs respond to queries characteristic to autistic users in contrast to queries typically asked by the general public, as well as the differences in IST responses across different kinds of queries asked by autistic users. We also investigated if an autistic user were to reformulate their query and explicitly state that they are autistic in their query would nudge the IST to respond differently to their query.

Juxtaposing the IST responses generated for control and ASD group queries revealed that the ISTs not only responded differently to the queries asked by the two user groups, but the responses generated for the control group were surprisingly more accessible for autistic users than the IST responses generated for their queries. For queries typically asked by autistic users, none of the ISTs produced responses suitable for autistic users. However, each IST had its strengths in the context of the accessibility which can employed in the other ISTs to improve the general accessibility of the responses generated by all ISTs. We also found that our query and prompt reformulation strategies affected the accessibility of each IST differently. In general, the reformulations did nudge the ISTs to generate responses that are more suitable for autistic users. This presents an optimistic solution to the issue of ISTs not catering to the accessibility needs of autistic users when they search for information on the web. Our results highlight areas where popular ISTs can be improved to make their responses more accessible to autistic users. Our systematic empirical investigation pipeline can also be extended to investigate IST responses on other factors that influence the accessibility of web content for autistic users.

Acknowledgments

This thesis began with the motivation to improve the accessibility of digital sources for non-traditional users. However, my idealistic and vague motivations could not have resulted in this extensive yet rigorous empirical study without the consistent support and guidance of my supervisor Dr Sole Pera. What started as a project to help others ended up helping me discover things about myself that I was unaware of. There were more than a few moments where I questioned my abilities, where the light at the end of the tunnel seemed too far for me to reach. But your consistent motivation and strong belief in me kept me going. Thank you for making me realise my aptitude and love for research, and building me into the researcher I am today.

To Adithi, my childhood friend and chaos-enabler, thank you for listening to me rant about my research and helping me decide on the statistical significance tests despite me confusing you with all of my experiments. You were right, the huge number of results nearly made me lose my mind, but I persevered thanks to a group of four brilliant young women that I had the fortune of meeting here in TU Delft - *Dream Team*, your cut-throat rivalry during board games and endless support otherwise, played a key role in keeping my sanity intact during my master's program, and I cannot thank all of you enough.

To my flatmate, Parvathi, thank you for being my biggest cheerleader throughout this entire process. Fixing your LaTEX issues and hearing you describe my project to others gave me the confidence boost I needed to keep believing in my research.

Finally, I would have never had this incredible research experience if not for the love and support of my family. Thank you for letting me take the risk and thank you for believing in me when I didn't. I wouldn't be here without you.

Contents

List of Figures

List of Tables

1

Introduction

The Internet is one of the largest sources of information in the world today. With just a few words entered into their preferred search engine **(SE)**, users are exposed to an endless stream of information about their topic of interest, in any form of media they desire. The rise in the popularity of AI chatbots based on powerful Large Language Models **(LLMs)**, such as ChatGPT, Gemini, etc. has introduced a new way for users to search for information on the web [89, 36]. With these chatbots, users can learn about their topic of interest by engaging in a human-like conversation as the chatbot generates answers in real-time based on the user's prompt, and can even modify its responses as per the user's request [49]. With over 5.4 billion Internet users across the world [63], the World Wide Web has disseminated information at an unprecedented scale. Yet not every user has the same experience when looking for information on the web, especially users with disabilities which affect their web search experience. Therefore, the World Wide Web Consortium **(W3C)** maintains a comprehensive list of guidelines to improve the accessibility of web content and technologies for all users [85]. These guidelines cover the needs of a very large population which may lead to the needs of non-traditional user groups being overshadowed by the needs of the more traditional crowd. For our study, we focus on one such group of non-traditional users - people on the Autism spectrum.

Autism Spectrum Disorder **(ASD)** is a neuro-developmental disorder that affects the way a person interacts and communicates with the people and environment around them [16]. As the name suggests, the abilities of autistic people can vary immensely, i.e. their abilities lie on a *spectrum* with every individual's condition evolving with time [62]. While some people require considerable help to carry out their day-to-day activities, others can live with little to no support at all. Despite the varying abilities of autistic people[1], most of them face issues with attention, stimulus sensitivity, and language and visual comprehension [4, 28]. These issues make face-to-face social interactions overwhelming and anxiety-provoking for autistic people [7] due to which they tend to drift to online mediums to make social connections and talk about their interests [31]. Yet autistic users do not display a strong preference for using the Internet to search for information to bridge their knowledge gaps, for instance when they need to learn about a topic related to their work or school [81, 31], which is an odd contradiction.

Empirical studies investigating the lack of interest of autistic users in looking for information on the web have revealed that there exist several barriers that hinder autistic users' access to the content on the web [24, 53, 70]. The issues due to which autistic users turn to online platforms to form social connections and share their interests are the same reasons that prevent them from efficiently looking for information on the web. For instance, a common issue for autistic users is their struggle with adapting to unexpected changes [28] but with online Information Seeking Tools **(ISTs)**, such as the traditional SEs and the new LLM-based chatbots, always working on improving the way they present information to the users, autistic users can never predict how their result page might look like. The changes could be as small as introducing small icons next to the URL on the search engine's result page **(SERP)** as done by Google (Figure 1.1) to more prominent changes like the introduction of featured snippets in Google SERP which changed the entire format of the SERP (Figure 1.2).

[1] There is an ongoing debate on how Autism must be described, but in this manuscript we use "identity-first" language (e.g., autistic people or autistic users) instead of "person-first" language (e.g., individual with Autism or user with Autism) based on the preferences of autistic people reported in different studies [37, 11].

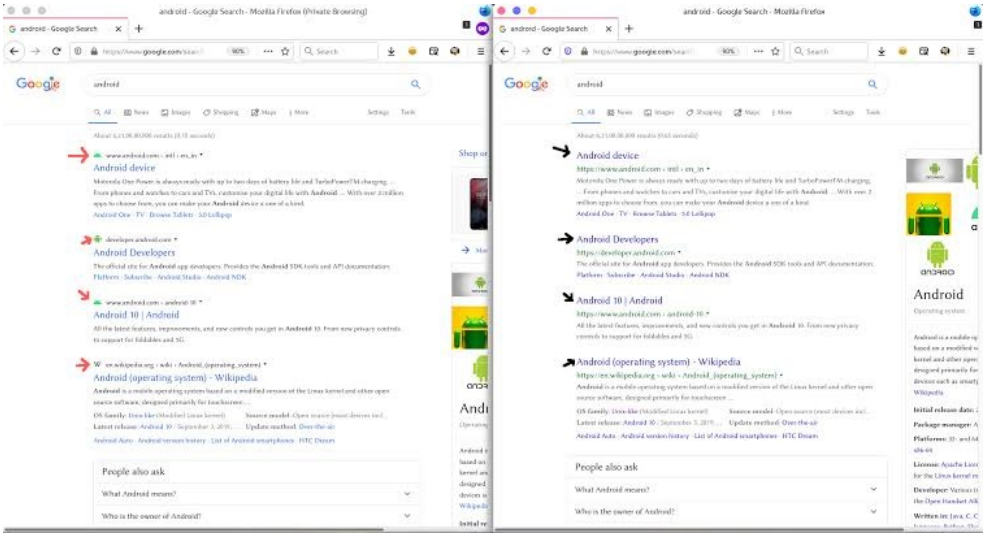
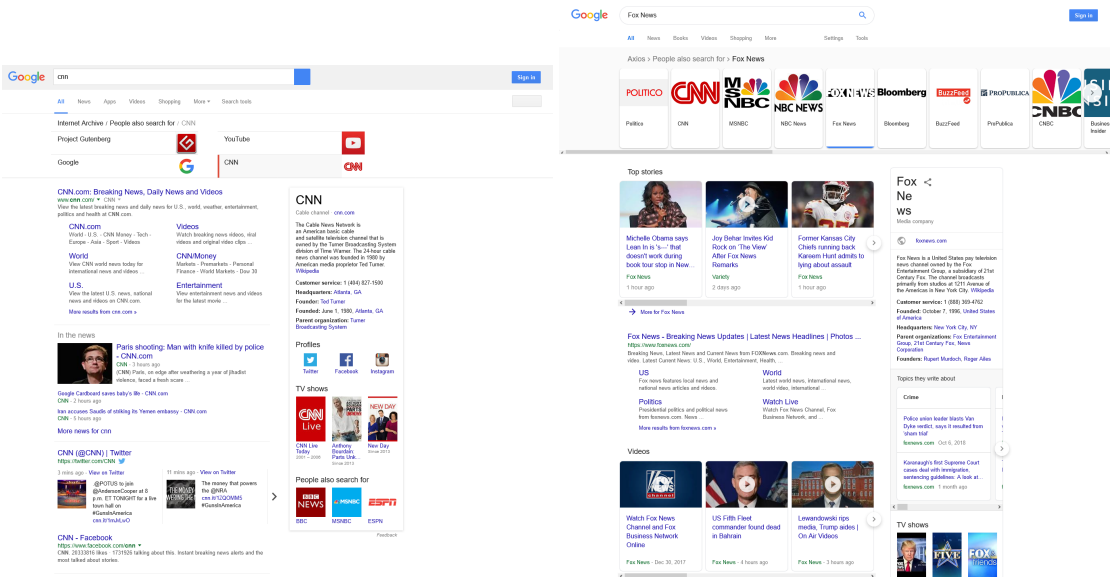


Figure 1.1: Introduction of favicons next to URL links on Google's SERP (left) and the previous design (right) [71].



(a) (b)

Figure 1.2: Google SERP in (a) 2015 and (b) 2018; taken from the dataset created by Oliveira et al [61]

Regardless of whether the change in the design of the IST's result page interface is minor or prominent, these unexpected changes can be anxiety-provoking for autistic users [2] who prefer predictable interactions and require considerable time to adapt to changes [62]. Furthermore, crowded or "busy" websites also pose a challenge for autistic users, as their attention issues and struggles with visual comprehension hinder them from assessing the relevance of an element on a web page [2, 24].

Besides the impact of visual factors of a website, even the phrasing of the textual matter can affect the search experience of an autistic user. For example, vague metaphors and "fancy" language which may make the information engaging for the average user, end up confusing an autistic user [2, 95]. Long passages of text common in the recently popular chatbot responses may lead to information or sensory overload which can incite challenging behaviours or meltdowns in them [2]. These are but a few factors that influence the web search experience of an autistic user, and while empirical studies have been useful

in expanding the knowledge on the limitations and preferences of autistic people when accessing the information on the web, merely increasing the awareness of their limitations is not sufficient.

Access to information is integral to a person's freedom of expression [83], and for a community that forms about 0.72% of the world's population [80], to be unable to access information on the Internet efficiently, one of the most popular sources of information, presents a problem that requires urgent attention. While assistive technological interventions for autistic people have made use of the existing guidelines to inform and evaluate their design choices, the existing interventions are yet to address the struggles of autistic users when looking for information on the web [44, 14, 21]. Thus, in this study, we focus on the existing online ISTs, specifically SEs and LLM-based chatbots, and investigate whether they cater to the information needs of autistic users. For this purpose, we designed an evaluation pipeline to answer the question: ***Are popular online Information Seeking Tools (ISTs) catering to the needs of autistic users in a way accessible to them?***

To guide our investigation, we pose two research questions:

RQ1.1: Is searching for information more accessible for an autistic user on a traditional search engine or through an LLM-based chatbot?

RQ1.2: Are the responses produced by popular ISTs more accessible when the query/prompt makes it explicit that the user is an autistic user?

To answer our research questions we conducted an in-depth empirical exploration of the accessibility of the responses provided by popular ISTs to autistic users. For better contextualization of our results, we compare the IST responses for autistic users' queries i.e. ASD group queries with the responses produced by the same ISTs for queries of the general population i.e. control group queries, to investigate whether the ISTs respond differently to control group and ASD group queries in the context of accessibility of the information in the IST responses.

We limit our scope to the textual information in the IST responses for our empirical exploration. Autistic people face difficulties processing huge blocks of text [32] and struggle with understanding complex sentences with niche jargon and abstract concepts [30, 59]. Thus, for textual content to be accessible to autistic users, it should be kept concise and easy to read i.e. the content should be broken down into small paragraphs or a list of bulleted points and the content itself should be written in simple and plain English as much as possible [95]. Therefore, we investigate the responses on three broad categories of web content accessibility - (1) structure of the textual content, (2) readability of the response, and (3) concreteness of the response.

Due to the lack of standard datasets of queries typically asked by autistic users, we had to generate synthetic queries/prompts that represent the information needs as well as the linguistic style of autistic users when looking for information on the web. The synthetic queries were generated using the data collected from Reddit posts under subreddits dedicated to autistic people by following the procedure used in similar studies conducted for individuals with mental health issues [56, 22]. To represent the information needs of the general public, we picked a random sample of queries from the Yahoo! Search Query Tiny sample dataset [91] with the same query length distribution as the ASD group queries.

To simulate a typical web search experience for an autistic user, we focused our investigation on the most popular ISTs, i.e. Google and Bing for traditional SEs, and ChatGPT and Gemini for LLM-based chatbots. We based our simulation on the Information Search Process **(ISP)** proposed by Kuhlthau [42] which describes a six-stage process when a user searches for information on the web. The six stages of the ISP are (1) Initiation (noticing a lack of knowledge), (2) Selection (identifying the general topic of inquiry), (3) Exploration (collecting information on the general topic), (4) Formulation (refining initial query to a specific topic), (5) Collection (collecting information on a specific topic), and (6) Presentation (present or use the collected information). For our investigation, we intended to evaluate the responses produced by the ISTs and so we assume that the autistic user is already aware of their information need and have a query that they believe is appropriate to answer their question. Therefore we skipped the Initiation and Selection stages and conducted our experiments on the IST functionalities an autistic user would use during the Exploration, Formulation, Collection, and Presentation stages.

To represent the responses offered by SEs at the different ISP stages, we followed the mapping between ISP stages and SE functionalities described by Milton and Pera [56]. We collected the web

snippets from the first 10 responses of the SERP for the Exploration and Formulation stages. For the Collection and Presentation stages, we collected text content from the top 10 Retrieved Resources **(RR)**. In the case of a chatbot, we collected the first response generated by the chatbot in response to the user's prompt. The collected responses were evaluated on how much their textual content adheres to the widely accepted web content accessibility guidelines for autistic people using quantifiable indicators we created based on the guidelines which we henceforth refer to as **Accessibility indicators**.

The results of our experiments reveal that ISTs respond differently to control and ASD group queries, with neither the SEs nor chatbots meeting the requirements of autistic users for the ASD group queries, especially the readability and concreteness of the textual information they provide. investigation into the effect of the query length did reveal variations in the responses across different query lengths although the variations were largely unique for search engines and LLM-based chatbots individually. Our results also revealed that the responses provided by the ISTs were even less suited to the needs of autistic users when the queries were directly related to Autism. The only exception to this trend was the structure of the responses provided by Google. Thus it is evident from our studies that not only are the popular SEs and LLM-based chatbots unable to cater to the needs of autistic users but the information related directly to Autism is even less accessible to autistic users.

In our experiments, we also observed that reformulating the query/prompt to explicitly state that the user is autistic led to the SEs producing responses that were much more readable and concrete. However, we must also take into account that Autism is a largely under-diagnosed disorder with even reports of WHO stating to be an estimated average at best with no information about its prevalence in low- and middle-income countries [62]. Due to this, many web users across the globe may not even be aware of their diagnosis [43, 58] which means they would never reformulate their queries or prompts to explicitly state that they are autistic. Hence ISTs' algorithms must be modified to detect implicit signals in the user's query and produce simpler and more accessible responses when required.

While the other domains work on improving awareness and diagnostic criteria and therapeutic interventions for Autism, information systems should strive to make the information accessible to all users regardless of their disabilities which they may or may not be aware of. Our study sheds light on the different ways popular online ISTs respond to queries typically asked by a traditional and an autistic user. The results of our experiments also reveal the strengths and limitations of four popular ISTs when responding to queries posed by autistic users in the context of accessibility. Based on our results, we suggest how popular ISTs could improve their respective algorithms to ensure that the information they provide is accessible to autistic users. Our accessibility evaluation pipeline can also be extended to other web content accessibility guidelines and can be moulded to fit other information search behaviours on the web.

In the rest of the manuscript, we first discuss the background and related literature to offer context to our study and inform our experiment design (Chapter 2). In Chapter 3 we elaborate on the evaluation pipeline we designed for our exploration detailing our approach to synthesising queries and prompts characteristic to autistic users, the quantifiable indicators we created to investigate the accessibility of the responses collected from the ISTs, and the experiments designed to answer our research questions. Since we collect data from public posts shared by autistic people on Reddit, we explain the data collection ethics followed to maintain the privacy of autistic users in Chapter 4. We then report the results of our experiments in Chapter 5 and discuss the implications of the results in Chapter 6. We conclude our work in Chapter 7 and mention the limitations of our work which could serve as inspiration for future work.

2

Background and Related Work

Empirical research has revealed the needs and preferences of autistic users when searching for information on the web. However, whether the ISTs used by autistic users are sensitive to their needs remains understudied which is the focus of our exploration. In this chapter, we discuss related work and literature to offer context and inform our experimental designs for our empirical investigation into the accessibility of responses provided by popular ISTs to autistic users.

2.1. Autism Spectrum Disorder (ASD)

Autism Spectrum Disorder (ASD) is a neurological developmental disability which causes people to communicate and behave in ways that are different from most people. People with ASD can begin showing symptoms as early as the age of 2 years, but the symptoms and effects of the disorder vary from person to person and evolve with age [28, 62]. The abilities of people with Autism are described to lie on a spectrum and categorised on three levels wherein people with level 1 Autism require low-level support mostly with social communication while people with level 3 Autism require substantial support to carry out even day-to-day activities [4].

Despite the heterogeneity in their abilities, there are some common challenges that autistic people face such as issues related to attention, language and visual comprehension, and stimulus sensitivity [4]. Autistic people also prone to restricted and repetitive actions [28] and tend to focus on individual details which prevent them from perceiving the bigger picture [32, 64]. Due to these challenges and preferences, autistic people often find face-to-face social interactions overwhelming and anxiety-provoking which hinders them from forming and maintaining meaningful social relationships [7] and are also reported to have a lower quality of life than non-autistic people [48, 52].

Although the scientific knowledge about Autism has advanced substantially over the past few years [29, 51], ASD remains a largely under-diagnosed disorder [60, 62], especially amongst women [15, 23], with many individuals receiving a proper diagnosis much later in their adolescence or even adulthood [58, 43]. The late diagnosis hinders autistic people from receiving the appropriate attention they require to reach their full potential which in turn has been reported to have a long-term negative impact on their employment rates, social relationships, and mental and physical health [35]. Studies have revealed that autistic people show a natural affinity to technology to develop their abilities and aid them in their daily life [47]. Specifically, autistic people have shown a keen interest in using online communication mediums to form new social connections and learn about their interest as they lack the complexities of face-to-face social communication [10]. Researchers have made use of the behavioural data collected from autistic users' presence on online platforms and investigated the use of machine learning (ML) and artificial intelligence (AI) algorithms to facilitate early diagnosis of Autism [92, 45, 86]. However, the impact of these technological interventions remains limited [78, 9].

2.2. ASD and the Online World

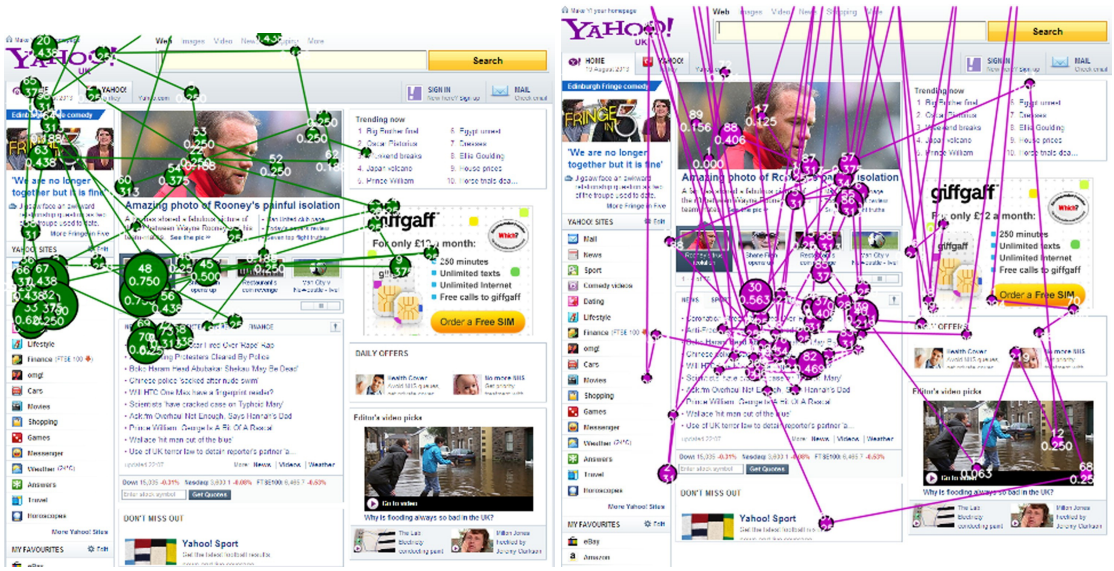
Due to their struggles with social interactions, autistic people have a very restricted friend circle due to which they are reported to show higher levels of loneliness compared to the general public [77, 50]. Social media has proven to be a great tool for them to abate this problem with studies revealing a strong preference amongst autistic people to maintain their social contacts and even create new ones [31, 87].

Online communications provide autistic people with a sense of control over the conversation that autistic people found to be lacking in a face-to-face conversation [5]. Communication via online mediums also allows autistic people to take a break from a conversation if they get too overwhelmed and thereby allow them to control their availability for a conversation [5, 41].

Advantages like better control over conversation, lack of nonverbal feedback, and inherent anonymity of online communications have enabled a strong presence of autistic people on social media platforms. Researchers have in turn made use of their posts and their interactions on platforms like Reddit or Twitter to detect linguistic and behavioural signals characteristic to autistic people which can be used in detection models to facilitate early diagnosis of Autism [39, 72]. For our investigation into the accessibility of IST responses to autistic users, we make use of the posts shared in subreddits dedicated to autistic users on Reddit to generate synthetic queries that would represent the information needs of autistic users using the framework followed in similar studies conducted for users with mental health issues [22, 56].

2.3. Information seeking with ASD

While autistic users' engagement with social media platforms has been useful in discovering characteristic behaviours of autistic users [54, 84], Gillespie-Lynch et al. conducted an online survey on the preferences of autistic users when using computer-mediated communications and they reported a huge disparity in the enthusiasm amongst autistic users when accessing the web to engage with information about their interests and look for information related to their work [31]. Controlled eye-tracking experiments have revealed that autistic adults are also not efficient when looking for information on web sources. The autistic users were reported to scan more items on the web page provided to them to answer a question, most of which were irrelevant to the question. They were also reported to return to items more often than the control group [24, 26]. Figures 2.1b and 2.1a highlight the stark difference between the search behaviour of autistic users and non-autistic users with the autistic user's eyes scanning the entire web page while the non-autistic user, having quickly recognised the relevant elements, had a much more focused scan path.



(a) Scan path of non-autistic user's eye movements **(b)** Scan path of autistic user's eye movements

Figure 2.1: Scan paths of the eye movements of a non-autistic user (left) and an autistic user when looking for information on Yahoo! homepage for the study conducted by Eraslan et al. [24]

Some researchers attribute the inefficiency of autistic users when looking for information on the web to the difference in the attention patterns between autistic users and non-autistic users as well as the information-processing issues common in autistic people [4, 88]. Furthermore, interviews with autistic adults revealed that most autistic adults reported an aversion to text-heavy elements and a preference for image-based elements[95]. Although the image-based elements were only effective when they were

directly related to the textual content, as the presence of abstract images and icons had a negative effect on the autistic users' efficiency [95, 70] when searching for information on a web page.

While these experiments improved the understanding of the needs of autistic adults when accessing web resources, the users were provided with ready-made queries based on the resources pre-selected by the researchers to assess the effect of different aspects of web design on the accessibility of information on the web to autistic users. This kind of controlled experiment process therefore may not truly reflect the information needs of the autistic adults. Research so far has focused on how the autistic users react to the web resources presented to them, however, in this study we switch the target of assessment and inspect whether the system can efficiently respond to the information needs of an autistic user.

2.4. Information Search Process (ISP)

When looking for information on the web through online ISTs, the user is exposed to various functionalities of the IST each of which could be mapped to different stages of their ISP as done by Milton and Pera [56]. To ensure users are able to find a satisfying answer to their questions, ISTs must respond to their information needs in a way that is accessible to them. However, with over 5.44 billion internet users worldwide (as of April 2024) [63], serving the needs of such a large population requires some generalisation. However, generalisation often leads to the overshadowing of minority groups i.e. groups of non-traditional users whose needs may not be largely different from the more traditional users. In our work, we focus specifically on one such group of non-traditional users, i.e. autistic users, and due to the lack of a model specific to the search patterns of autistic users when searching for information on the web we base our work on the model proposed by Kuhlthau [42].

The Information Search Process (ISP) as defined by Kuhlthau [42], is a six-stage process which begins with the user realising that they have an information need i.e. the *Initiation* stage when the user becomes aware that they lack knowledge or understanding of a particular topic. Upon realisation of their need, a user moves to the next stage, the *Selection* stage wherein the user attempts to ascertain the general topic their information need is about. During this stage due to the user's lack of knowledge, they may not be able to express their information need properly to the information system which hinders the communication between the user and the system. This is where the search engines' query suggestions come in handy.

Once the user gains some understanding of the general topic, they move on to the next stage - the *Exploration* stage, the user begins to look for information about the general topic they ascertained in the previous stage to gain further understanding about their information need. Through their exploration, the user can refine their information need and thus enter the *Formulation* stage where they define what their exact information need is which manifests as the user's search query when interacting with a digital information system. During these two stages, a traditional SE aids the user by presenting a ranked list of resources on its results page (SERP) which its algorithm computes to be relevant to the user's information needs [56].

Finally, the user is in the *Collection* stage and must select the retrieved resources (RR) they wish to look into and collect the information they need to answer their query. Once satisfied with the information they have acquired, the user moves to their final stage *Presentation* where they present the information they had gone through.

For our empirical study, we aimed at simulating the ISP for autistic users when using popular types of ISTs, specifically SEs and LLM-based chatbots. The purpose of this study is to inspect the accessibility of IST responses when an autistic user uses it to answer a question, hence we assume that the autistic user is aware of their information need (Initiation stage) already formulated an initial query (Formulation stage) which we represent by the synthetic queries we generate using the framework followed by [22] and [56]. We thus conduct our investigation by focusing on the four latter stages of the ISP i.e. Exploration, Formulation, Collection, and Presentation stages.

3

Experimental Setup

To investigate the accessibility of responses produced by popular online ISTs, in this work, we simulate the last 4 stages of the ISP for autistic users, namely the Exploration, Formulation, Collection, and Presentation stages. We investigate the accessibility of the responses generated for each of these stages by four ISTs in particular - Google, Bing, Gemini, and ChatGPT. In this chapter, we elaborate on the sequence of steps undertaken to conduct our investigation. All code files can be found in the dedicated GitLab repository

3.1. Generation of synthetic queries and prompts

To investigate the accessibility of IST responses to autistic users, we first need queries typically asked by autistic users to collect responses from the ISTs, but due to the lack of publicly available databases containing query logs of autistic users, we had to opt for an alternative method and generate synthetic queries. To represent the information need of autistic users appropriately we needed to capture their topics of interest and their linguistic patterns. People on the spectrum often struggle with social interactions and thereby tend to prefer online platforms to talk about their interests [31, 33] preferring written forms of communication over face-to-face communication [34, 5]. So we turn to a popular social media website - Reddit. As per their reports, [68], Reddit observed 82.7 million daily users in the first quarter of 2024. Users (called Redditors) interact with each other through communities called “Subreddits” wherein the Redditors share information and opinions on topics related to the subreddit via posts and comments creating a “thread” under each post. Each subreddit is moderated by a voluntary group of redditors who ensure that all users follow the community’s rules and that the threads under the subreddit are relevant to it. The moderators are also aided by the active feedback of all the redditors who interact with the posts and decide how relevant the post is via “upvotes” and “downvotes”. Thus each subreddit can be considered an online community controlled fully by the people interacting with the subreddit.

We found several subreddits dedicated to the interests and daily struggles of autistic people. We collected the names of some of the most popular subreddits related to Autism, making sure that they were related to autistic people and their topics of interest *directly* and were not intended for family, friends, or caretakers of autistic people. We compiled a list of 16 subreddits (listed in Appendix A) and made use of PRAW [1]- the official Python wrapper for the official Reddit API to crawl the top 1000 posts from each subreddit. Posts that had no text content in both the title and body were removed. For the remaining posts, the text content of the post’s title and body were concatenated to create a text sample for each post. Thus we obtained around 11,100 text samples used as the raw data to generate our synthetic queries.

Finding keyphrases. All text samples were cleaned to remove the name of subreddits, URLs, escape sequences, dates and time notations, and then parsed using `KeyBERT`[2] to extract key phrases. `KeyBERT` is a keyword extraction technique that makes use of BERT embeddings to create keywords and key phrases that best represent the text sample. We made use of the default transformer used by `KeyBERT` i.e Hugging Face’s sentence transformer model `all-MiniLM-L12-v2` but for the length of the key phrases we made use of `KeyphraseVectorizers`[3] which enabled `KeyBERT` to extract grammatically accurate key phrases.

[1] https://github.com/praw-dev/praw

[2] https://github.com/MaartenGr/keyBERT

[3] https://github.com/TimSchopf/KeyphraseVectorizers

We maintained a count of how many posts the extracted key phrases occurred in. Phrases which were the singular or plural forms of each other were merged and the sum of all the similar key phrases represented the final frequency of each key phrase. Any unigrams that were stopwords listed under NLTK or Spacy were also removed. This resulted in approximately 28,800 key phrases.

Selecting queries. To further narrow down our list, we arranged all the key phrases in descending order of their frequency within their n-gram category. We looked at the frequency distribution of queries grouped by their n-gram length [4]. We estimated the upper and lower frequency thresholds for each n-gram category (Table 3.1) following Zipf's law which determines the qualification of keywords from a dataset based on their frequency[65]. However certain dedicated key phrases were retained due to their high relevance to the user group. For key phrases longer than 4-gram, all of them occurred only in one post but for the sake of variety in the n-gram length we retained queries which contained dedicated words listed in Table 3.2. Through this process, we obtained around 7000 key phrases.

n-gram	minimum number of posts	maximum number of posts
1	1	400
2	1	100
3	1	10
4	1	4

Table 3.1: Frequency thresholds for queries of different n-gram lengths to select the most representative subset of popular n-gram queries.

Dedicated word(s)	ngram
autism	1
diagnosis	1
work	1
sensory issue	2
special interest	2
autistic person	2
mental health issue	3
higher support needs	3
autism spectrum disorder	3
low support need	3
world autism awareness day	4

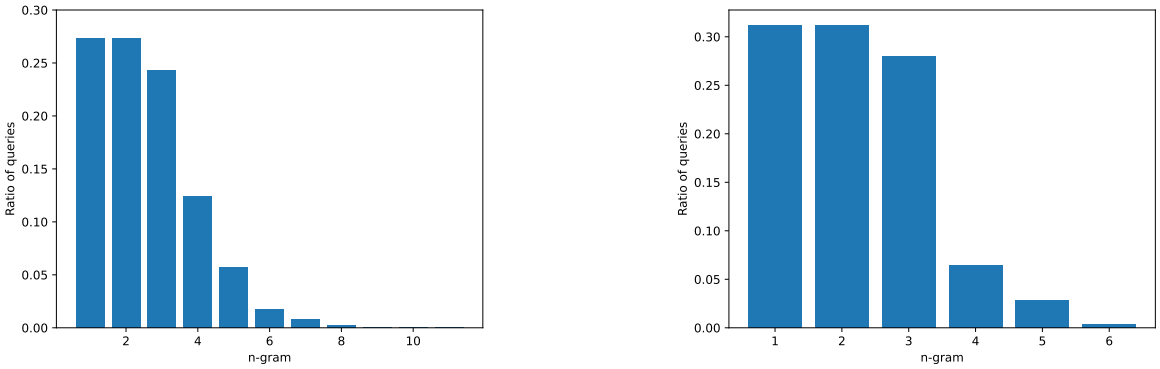
Table 3.2: List of dedicated words/phrases added to the final list of queries/prompts despite high frequency due to their relevance to autistic users.

For our study, we shortlisted a total of 250 queries to conduct our investigation. The final list of queries/prompts was categorised on two properties: n-gram length and whether the query/prompt was domain-specific.

Distribution of queries/prompts by n-gram length. For n-gram length distribution, since we do not have any publicly available information on the typical length of queries framed by autistic people, we resorted to following the distribution of n-grams in Yahoo! Search Query Tiny sample dataset [91]. Based on the distribution of the Yahoo! dataset (Figure 3.1a), we picked the top (frequency-wise) key phrases from each n-gram category to make a total of 250 queries for the SEs. It is to be noted that for the sake of consistency, we used the same list of key phrases as prompts when generating responses from the LLM-based chatbots. The distribution of queries/prompts by their n-gram length is illustrated in Figure

[4] The frequency distribution of key phrases is visualised in Appendix B

3.1b.



(a) Yahoo! Search Query Tiny sample dataset **(b)** Final list of queries/prompts for our study

Figure 3.1: Distribution of queries in (a) Yahoo! Search Query Tiny sample dataset and (b) Final list of queries/prompts for our study by n-gram.

Distribution of queries/prompts by domain specificity. We categorised a query/prompt to be domain specific if it contained any of the dedicated terms from Table 3.2 except for the keyword “work” as although relevant to autistic users, work is not specific to autism. For a more exhaustive list of autism-related terms, we scraped the glossary terms from ReframingAutism.org), a charity run by autistic people in Australia aimed to improve the understanding of Autism through education, resources, and research. The distribution of queries/prompts by their n-gram length is illustrated in Figure 3.2.

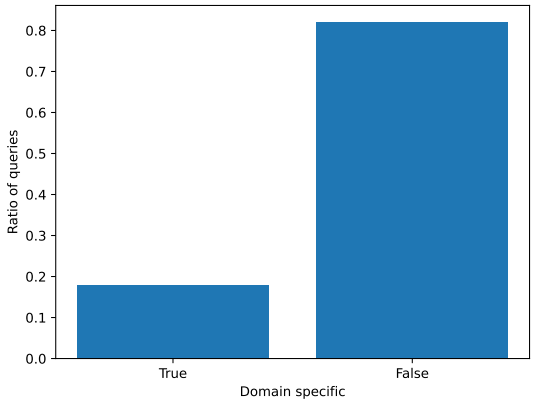


Figure 3.2: Distribution of final list of queries/prompts by domain specificity.

Control group queries. To investigate if ISTs respond differently for queries typically asked by the general public and for queries characteristic to autistic users, we needed a collection of queries to represent the information needs of our control group i.e. the general public. Thus we collected 250 queries from the Yahoo! Search Query Tiny sample dataset [91] through random sampling and kept the n-gram distribution of our selected queries the same as the distribution for the ASD group queries. By comparing the IST responses for control and ASD group queries, we can make inferences on whether the ISTs respond differently to queries made by the general public and autistic users, which would provide a better context for the implications of the results of our investigation into the accessibility of IST responses for autistic users.

Reformulating ASD group queries/prompts. Query reformulation by users during their web search tasks has been studied extensively [18, 6] and the results of the studies are being used to improve the query suggestions offered by SEs during a user's search [73, 90]. Similarly, prompt engineering to improve LLM responses is also an emerging area of research [75, 55]. Hence, in our study, we also investigated whether explicitly stating that the user is autistic by *reformulating the ASD group queries/prompts* affects the accessibility of responses produced by the SEs and Chatbots for autistic users.

To investigate the effect of the query/prompt reformulation on the responses produced by the SEs and Chatbots respectively, we reformulated the query for SEs by taking inspiration from the query categorisation technique followed by Yechiam et al. [96] for their study and appended every original query with the phrase `"I'm autistic"`. By this method of query reformulation, we explicitly stated that the user is autistic without adding any other distracting signals as "I'm" would be considered as stopwords by the SE's ranking algorithm when processing the reformed query to find and rank relevant documents. In the case of LLM-based chatbots, due to their conversational nature, we reformed the prompt to be conversational in nature as well i.e the reformed prompt would be of the form - `"I am autistic please explain "` + `prompt` + `"to me"`. Table 3.3 contains a few sample queries and prompts and their reformulated counterparts.

Query		**Prompt**	
Original	Reformulated	Original	Reformulated
autism diagnosis	autism diagnosis I'm autistic	autism diagnosis	I am autistic please explain autism diagnosis to me
common autistic trait	common autistic trait I'm autistic	common autistic trait	I am autistic please explain common autistic trait to me

Table 3.3: Examples of original and reformulated queries and prompts used to collect responses from SEs and Chatbots respectively

3.2. Collection of responses from popular online ISTs

Using the control group and ASD group queries and prompts, we collect responses from two types of popular online ISTs - SEs and LLM-based chatbots to compare and contrast the accessibility of the information provided by them to autistic users:

1. **SEs:**
   (a) Google: We made use of the Google Custom Search JSON API[5] to collect the top 10 results for each query.
   (b) Bing: We made use of the Bing Web Search API[6] to collect the top 10 results for each query.
2. **Chatbots**:
   (a) Gemini: We made use of the Gemini Pro model from Google's Vertex AI API[7] to collect responses from Gemini for every prompt.
   (b) ChatGPT 3.5: We made use of the GPT 3.5 Turbo model from OpenAI's GPT API[8] to collect the responses for ChatGPT 3.5 for every prompt.

For the SEs, we collected the first 10 responses for each query as that is the default number of responses presented on the first page of a SERP and most users never go past the first result page when searching for information [76]. We stored the title, snippet, and URL of each response, as well as the date the response was generated. For the Chatbots, we collected the responses generated by them for the prompt. In this case, prompts were the same as the counterpart queries used for the SEs. We made sure that each API call was independent of the other API calls to ensure that the responses acquired for each query were not affected by other queries.

[5]https://developers.google.com/custom-search/v1/overview

[6]https://learn.microsoft.com/en-us/bing/search-apis/bing-web-search/overview

[7]https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/gemini

[8]https://platform.openai.com/docs/models/gpt-3-5-turbo

For this experiment, we limited the search results to only the English language due to the researchers' fluency with the language thereby ensuring proper analysis of the textual content of the collected responses. Furthermore, we only looked at the textual content of the acquired responses. The textual content was cleaned to remove any non-ASCII characters and extra white spaces to enable proper analysis based on the markers described in the following subsection.

When searching for information on an SE, the user is exposed to information in two different ways - first the list of ranked resources on the SERP, and second the retrieved resource (RR) itself. We collected responses from both these functionalities and categorised them as two separate **response group types** for our comparisons, namely `SERP` and `RR`. In the case of an LLM-based chatbot, the user chats with the chatbot and all the information is provided to the user on a single interface. The responses generated by the chatbots are thus categorised under a single response group type which we refer to as `Chatbot`. Hence we have three response group types:

1. `SERP`: When a user enters a query on a SE, they are presented with a ranked list of documents which form the SERP. On the SERP each document is displayed as the title followed by a snippet of the document's content. Thus for our investigation of the accessibility of SERP responses, we represent each document as a concatenation of its title and snippet.
2. `RR`: When a user clicks on the link of one of the documents on the SERP, they can access the information presented on the Retrieved (web) Resource (RR). To investigate the accessibility of each retrieved resource, we scrape its HTML content using the Python package `BeautifulSoup`.
3. `Chatbot`: For the LLM-based chatbots, the user receives a generated text response for their prompt. Thus for each Chatbot, we collect the first response generated by the Chatbot for a given prompt.

Under each response group type, we compare the responses from two ISTs each of which forms a **response group** under the given response group type:

1. `SERP`:
   (a) `Google SERP`: The text content from the first 10 web snippets taken from Google SERP for a query.
   (b) `Bing SERP`: The text content from the first 10 web snippets taken from Bing SERP for a query.
2. `RR`:
   (a) `Google RR`: The text content of the top 10 Google results for a query.
   (b) `Bing RR`: The text content of the top 10 Bing results for a query.
3. `Chatbot`:
   (a) `Gemini`: The first response generated by Gemini for a user prompt.
   (b) `GPT 3.5`: The first response generated by ChatGPT 3.5 for a user prompt.

3.3. Investigation of collected responses for accessibility

To investigate the accessibility of the responses produced by the ISTs, we look into the web content accessibility guidelines for autistic users and translate them into quantifiable indicators, referred to as **accessibility indicators**, to use for our empirical analysis. We looked at the guidelines concerning the readability of the textual content on the web as we limited the scope of our investigation to only the textual content of IST responses. We focused on guidelines that have been empirically proven through studies involving diagnosed autistic users to ensure that the guidelines represent the needs of a significant portion of autistic users.

To create our accessibility indicators, we took inspiration from the following guidelines:

- "Be succinct, avoid writing long paragraphs and use markups that facilitate the reading flow such as lists and heading title" [12]
- "Use texts written in plain English. A general rule of thumb is that the text should have a score higher than 65 according to the Flesch-Reading Ease formula" [95]
- "Use a simple visual and textual language, avoid jargons, spelling errors, metaphors, abbreviations and acronyms, using terms, expressions, names and symbols familiar to users' context" [12]

Based on these guidelines, we created three categories of accessibility indicators - (1) Text structure, (2) Text readability, and (3) Text concreteness.

Investigating the structure of textual content. Texts presented in short paragraphs or divided into lists are more suitable for autistic users over long continuous bodies of texts [12, 95]. We investigate how the textual content is structured on the web page. Since the guideline [12] suggests textual content to be succinct, brief sentences are preferred over long-winded sentences. Thus, we counted the number of sentences and computed the average length of a sentence in terms of the average number of words per sentence.

For better readability, guidelines compiled by Britto and Pizzolato [12] suggested dividing textual content into paragraphs led by relevant headings and presenting information as bullet points whenever possible. Thus we also computed the ratio of sentences which are list items or headings to understand how the text has been divided. Since paragraphs are usually the longest component of a text, we also looked at the average length of a paragraph to investigate if each paragraph is readable for an autistic user individually.

The calculation of the metrics is explained below:

1. Number of sentences: Total number of sentences in the text. It is to be noted that for this metric individual list items and headings also count as individual sentences.
2. Average length of sentences: Average number of words per sentence
3. Ratio of list items: Number of sentences which are an item in a list
4. Ratio of headings: Number of sentences which are headings or sub-headings
5. Ratio of paragraphs: Number of sentences which are part of a paragraph but are not a list item or a heading. It is to be noted that in the case of in-paragraph headings, the entity is also counted as a paragraph.
6. Average length of paragraphs: Average number of words per paragraph. Note that in the case of in-paragraph headings, we remove the heading before counting the number of words in the paragraph.

Investigating the readability of textual content. Yaneva [95] suggests text written as simple as possible and has even suggested that the text should have a score of 65 or higher in the Flesch-Reading ease formula [40] which can be described as text that can be easily understood by 13-15-year-old students [27]. However, the Flesch Reading ease score formula, like several other readability score formulas was developed based on research conducted on children and no further research was done to ensure whether the reading ease was reflected in adult readers as well [69]. However Flesch Reading ease score continues to be a standard measure of readability and since prior research has already provided us with a threshold we use it to compare and contrast the collected responses.

We also calculate the Coleman-Liau readability index [20] of the textual content as an additional readability measure due to its similarity to the Automated Readability Index which was found to be suitable for estimating the accessibility of a text document for autistic users [93] and due to the suitability of Coleman-Liau readability index for digital texts [3]. A Flesch Reading Ease score of 65 represents a test easily understood by 13-15-year-old students which corresponds to 8th - 9th grade in the US school system. In the Coleman-Liau readability index, the index value for 11-14-year-old students lies between 5-8, and the index value for 14-17-year-old students lies between 8-11, thus to represent the readability of students aged 13-15 years old we take the highest index value for group 11-14-year-old students and lowest index value for group 14-17-year-old students i.e ideally the Coleman-Liau index value for the text should be 8 or lower. We make use of Python's Textstat library to compute the readability scores[9].

Thus the two readability scores we used as accessibility metrics are:

1. Flesch Reading ease score (minimum threshold: 65)
2. Coleman Liau readability index (maximum threshold: 8)

Investigating the concreteness of textual content. One common issue faced by autistic users when reading textual content is understanding figurative language and abstract words [95, 93]. Therefore to quantify the presence of abstract phrases in the collected responses, we make use of the concreteness

[9] https://github.com/textstat/textstat

ratings estimated by Brysbaert et al. [13]. The dataset created as a result of their research is a list of 40 thousand generally known English lemmas with a mean concreteness rating for each lemma estimated based on the ratings collected from over 4000 participants. The concreteness ratings range from 1: maximally abstract to 5: maximally concrete. We use these concreteness ratings to compute the average concreteness of a response group sample. If a word from the sample did not exist in the dataset then we ignored it when computing the average concreteness.

Since the dataset cannot be exhaustive for the words used by an English-speaking user, we expected there to be a portion of words that do not have a concreteness rating assigned to them. Therefore the average concreteness computed would not be an accurate measure of concreteness of the text. Furthermore, average values are also heavily affected by extreme values, thus for a better understanding of the concreteness of the text, we also estimated the ratio of concrete and abstract words.

During their study, to ascertain the concreteness of a word, Brysbaert et al. [13] presented the participants with a 5-point rating scale with ratings between 1-3 categorised as abstract and ratings 4-5 categorised as concrete. Thus for our experiments, we considered a word to be concrete only if its concreteness rating was above 4. Otherwise, the word was categorised to be abstract.

The calculation of the metrics is described below:

1. Average concreteness of text: Average concreteness score of all the rated words in the text
2. Ratio of concrete words in the text: Number of words rated concrete (concreteness rating $>= 4$) out of all the words in the text
3. Ratio of abstract words in the text: Number of words rated abstract (concreteness rating < 4) out of all the words in the text

3.4. Quantitative analysis of accessibility indicators

For our investigation, we aggregated the multiple responses per query for the SEs and took the average value for each accessibility indicator to represent the accessibility of an SE's response for a given query. In the case of chatbots, no such aggregation was necessary as we collected only the first response per prompt. If a query/prompt resulted in no responses from an SE or Chatbot, the query/prompt was given a score of 0 for all the accessibility indicators to facilitate significance testing during comparisons.

Our quantitative analysis was broadly divided into two main analyses to answer each of our research questions introduced in Chapter 1.

RQ 1.1: Is searching for information more accessible for an autistic user on a traditional SE or through an LLM-based chatbot? To investigate the accessibility of the responses produced by SEs and LLM-based chatbots, we first compared the IST responses for control and ASD group queries on their scores for the accessibility indicators described in Section 3.3. The comparisons are conducted on the response group type level i.e. between the `SERP`, `RR`, and `Chatbot`. We perform a pairwise comparison for each response group type to investigate whether the ISTs respond differently to control and ASD group queries.

We further investigated the IST responses for ASD group queries by comparing the accessibility indicator scores according to response group types and response groups described in Section 3.2 in the following ways:

1. `SERP`, `RR`, and `Chatbot`
2. `Google SERP` and `Bing SERP`
3. `Google RR` and `Bing RR`
4. `Gemini` and `GPT 3.5`

We also looked at the potential effect of the nature of the ASD group query, given a response group, on the accessibility indicators' score distributions. We specifically categorised our queries in the following two ways:

1. n-gram length
2. domain specificity: domain-specific vs. general queries

RQ1.2: Are the responses produced by popular ISTs more accessible when the query makes it explicit that the user is an autistic user? To investigate the effect of reformulating the ASD group queries and prompts on the IST responses (reformulation method described in Section 3.1), we first measured the change in the responses produced by each IST.

To measure the differences between the responses produced by the SEs for the original ASD group queries and the reformulated ASD group queries, we compute the Rank Biased Overlap (RBO) [74] between the ranked list of URLs produced by a SE for a given query. RBO is a similarity measure that computes the similarity between two ranked lists of entities and returns a numeric value between 0 and 1. A value of 0 would mean that the two lists are completely different while a value of 1 would mean that the two lists are identical. For implementation purposes, we made use of the python package by the same name - rbo[10].

We took into account the possibility of the SE ranking different landing pages of the same web resource differently for the original and reformed query, thus for the computation of RBO, we extracted the homepage URL from each web resource and thus compared the list of homepages to compute the RBO between the resources retrieved by a SE for a given query and its reformed alternative.

For Chatbot-generated responses, we compared the similarity of the responses with the similarity measure Recall-Oriented Understudy for Gisting Evaluation i.e. ROUGE [46]. ROUGE is typically used to compare an automatically generated summary of a document to a gold-standard summary (usually human-made), and it works by computing the recall and precision of the generated summary compared to the gold-standard summary. In our work, to compute the effect the query reformulation has on the response, we take the response generated for the original prompt as the "gold standard".

We computed the ROUGE-L score due to its reported effectiveness on single document summarization tasks [46] and since the metric takes the longest common subsequence between the two texts, we did not need to decide the n-gram length and consequently ended up influencing the scores. We evaluated ROUGE-L scores on both the sentence level and summary level using the Python package torchmetrics[11]. We then compared the ROUGE-L score distributions between `Gemini` and `GPT 3.5` on both sentence and summary levels.

We compared the change in the accessibility indicator scores when the query/prompt was reformulated as described in Section 3.1 between the following pairs:

1. `Google SERP` and `Bing SERP`
2. `Google RR` and `Bing RR`
3. `Gemini` and `GPT 3.5`

Testing for statistical significance: All analyses were tested for significance using the Kruskal-Wallis H test ($p < 0.05$) for comparison between > 2 groups and using the two-sided Mann-Whitney U (with $p < 0.05$) test for pairwise comparisons. The significance tests were implemented using Python's scipy package[12].

[10] https://github.com/changyaochen/rbo

[11] https://github.com/Lightning-AI/torchmetrics

[12] `https://github.com/scipy/scipy`

4

Ethical Considerations

The project is intended for a targeted group of web users i.e. autistic users. To represent their information needs, we had to acquire information available publicly on the Web, and thus we designed our project keeping in mind the ethical concerns that may arise during and after the course of this project. In this chapter, we elaborate on the data management practices and ethics practised during the project.

4.1. Data Management

We received a sample of the queries from Uitdenbogerd et al. [82] through personal communication with the researchers. These queries were phrased by the participants in their study who were diagnosed with Autism, and we obtained them to gain insight into the linguistic style of query formulation of autistic users. To generate synthetic queries for autistic users we scraped data from public Reddit posts using PRAW - the Python wrapper for the official Reddit API. The public posts were scraped for their post title and textual content in the post's body. It is to be noted that none of the user's details including their username and user id were scraped for this work. The posts were scraped from 16 public subreddits dedicated to Autism and autistic users directly. The names of the subreddits can be found in Appendix A.

The synthetic queries generated using the Reddit data were then used to retrieve responses from SEs (specifically Google and Bing) and LLM-based chatbots (specifically Gemini and ChatGPT 3.5). The responses from SEs and Chatbots were also collected using the official APIs namely, Google's Custom Search JSON API[1], Bing's Web Search API[2], Google Gemini's Model API in Vertex AI[3], and OpenAI's GPT 3.5-turbo[4]. The process of response collection can be found in Section 3.2.

All the data collected was stored in a private repository on the GitLab server maintained by the Faculty of Electrical Engineering, Mathematics and Computer Science at TU delft. Following the Reddit Data API guidelines [67], the data scraped from public Reddit posts were discarded after the completion of the project. We also did not have permission to share the sample queries obtained from Uitdenbogerd et al. [82] hence the sample queries were also removed from the repository after the completion of the project.

4.2. Ethics

The project did not involve any direct human participation. The only data subjects (indirectly) involved in the project were Reddit users aged 13+ who had made public posts under the subreddits listed in Appendix A. It is not feasible to contact all the Reddit users whose public posts will be collected to ask for their consent. However, the data collection was conducted for academic research and we are relying on public interest as our legal ground for any potential personal data processing.

Furthermore, we did not collect any personal information of the users including their username and user ID, and the posts were processed only to extract common phrases from all the collected posts. Hence we only used the anonymised aggregated data from the collected Reddit posts for the project.

[1] https://developers.google.com/custom-search/v1/overview

[2] https://learn.microsoft.com/en-us/bing/search-apis/bing-web-search/overview

[3] https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/gemini

[4] https://platform.openai.com/docs/models/gpt-3-5-turbo

5

Results

In this chapter, we report the results of our experiments described in Section 3.4. We compared and contrasted the scores of different response groups (described in Section 3.2) across various accessibility indicators described in Section 3.3. The comparisons were made over three broad categories - (1) the structure, (2) the readability, and (3) the concreteness of the textual content of the collected responses. All analyses were also tested for statistical significance using the Mann-Whitney U test for pairwise comparisons and the Kruskal-Wallis H test for multiple groups with $p < 0.05$. Note that all results were significant unless stated otherwise.

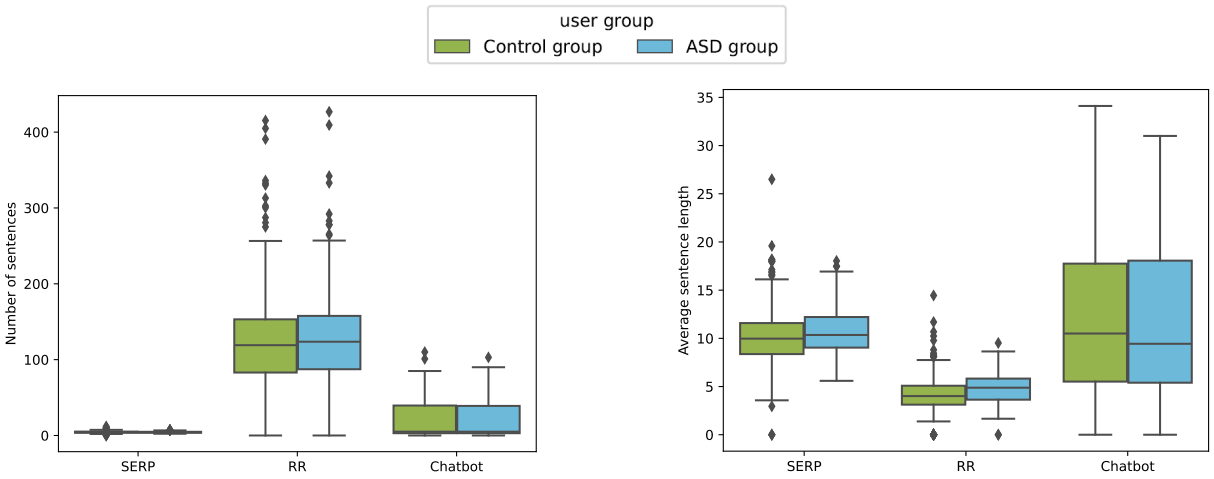
5.1. Comparing the accessibility of IST responses

We compared and contrasted the accessibility of SEs with that of LLM-based chatbots. We also delve into the potential effect of the query length and domain specificity of queries on the accessibility of the responses produced by the ISTs.

5.1.1. Comparing IST responses for Control and ASD group queries

Comparing the IST responses for ASD group queries and control group queries on the response group type level (`SERP`, `RR`, and `Chatbot`) resulted in many significant differences in the accessibility indicator scores.

Text Structure. The number of sentences was similar for control and ASD group queries across all response group types and the distributions were also not significantly different. However, the sentences were significantly shorter for SE responses for control group queries (9.97 for `SERP` and 4.01 for `RR` responses) than for ASD group queries (median value of 10.35 for `SERP` and 4.88 for `RR` responses). The `Chabot` responses were longer for Control group queries (median value of 10.5 for control and 9.44 for ASD group queries) but the difference was not statistically significant.

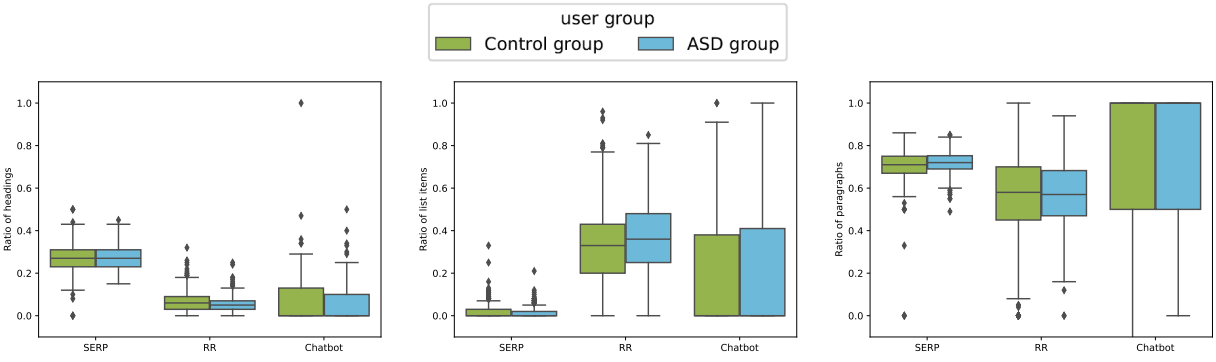


(a) Number of sentences (One outlier removed from `Chabot` for the control group queries for visualisation purposes) (Distributions were not significantly different for any response group type).

(b) Average sentence length (Distributions were not significantly different for `Chatbot`).

Figure 5.1: Analysis of text structure (sentence-level) based on `SERP`, `RR`, and `Chabot` for Control and ASD group queries.

The ratio of headings in the `SERP`, `RR`, and `Chabot` were similar for the control (median value of 0.27 for `SERP`, 0.06 for `RR`, and 0 for `Chatbot`) and ASD group (median value of 0.27 for `SERP`, 0.05 for `RR`, and 0 for `Chatbot`) queries, but the control group queries resulted in a (statistically) significantly lower ratio of list items in `RR` responses (median value of 0.33 for control group and 0.36 for ASD group queries). The ratio of paragraphs however was not significantly different for control and ASD group queries except for `SERP` responses wherein the control group queries had a significantly lower ratio of paragraphs (median value of 0.71) than ASD group queries (median value of 0.72).



(a) Ratio of headings (Distributions were significantly different only for `Chabot`, $p<0.05$)

(b) Ratio of list items

(c) Ratio of paragraphs (Distributions were significantly different only for `SERP` responses, $p<0.05$).

Figure 5.2: Analysis of text structure (full body) based on `SERP`, `RR`, and `Chabot` for Control and ASD group queries.

The paragraphs of `SERP` and `RR` responses were significantly shorter for control group queries. The median value of average paragraph length for the control group was 17.75 while for the ASD group, it was 19.25. In the case of `RR` responses, control group queries had a median value of 19.49 and 22.56 for ASD group queries. The paragraphs of `Chabot` were also longer for control group queries (median value of 27) than for ASD group queries (median value of 25) but the difference was not statistically significant.

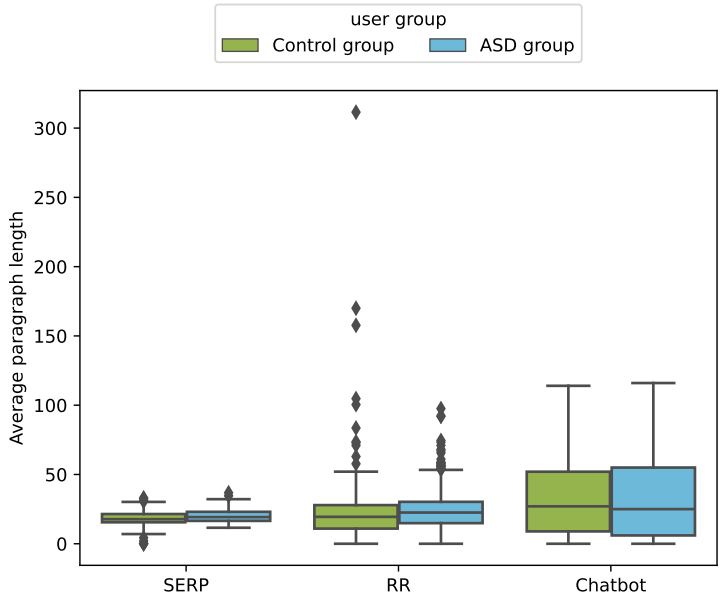
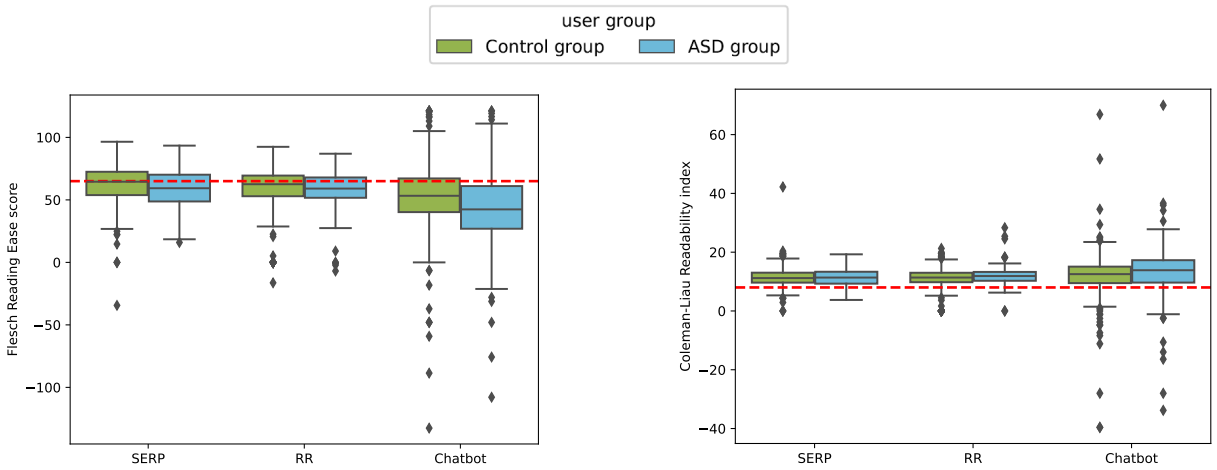


Figure 5.3: Average paragraph length of `SERP`, `RR`, and `Chabot` for Control and ASD group queries (Two outliers from `Google RR` responses for the control group queries and one outlier from `Bing RR` responses for the ASD group queries have been removed for visualisation purposes) (Distributions for `Chabot` not significantly different).

Text Readability. The Flesch reading ease score was significantly higher for control group queries (median value of 64.55 for `SERP`, 62.52 for `RR`, and 52.29 for `Chatbot`) than for ASD group queries (median value of 59.32 for `SERP`, 59.06 for `RR` responses, and 42.41 for `Chatbot`) with `SERP` and `RR` responses for control group queries nearly surpassing the minimum threshold of the indicator for the text to be readable by autistic users.

The Coleman-Liau readability index scores did not have a huge margin of difference as the difference in the Flesch reading ease scores for the control and ASD group queries. The control group queries led to responses with a lower index value (median value of 11.19 for `SERP`, 11.38 for `RR`, and 12.53 for `Chatbot`) than ASD group queries (median value of 11.36 for `SERP`, 11.88 for `RR`, and 13.86 for `Chatbot`) although the difference in `SERP` responses was not statistically significant.



(a) Flesch Reading Ease score

(b) Coleman-Liau Readability Index (Distributions for `SERP` responses not significantly different).

Figure 5.4: Analysis of text readability based on `SERP`, `RR`, `Chabot` for Control and ASD group queries. (One outlier has been removed from `Google RR` responses for Control group queries from both readability scores for visualisation purposes).

Text Concreteness. All responses for control group queries were significantly more concrete than responses for ASD group queries with the median value for average concreteness for control group queries being 2.5 for both `SERP` and `RR` responses and 2.49 for `Chabot` responses. In the case of ASD group queries, the median values for average concreteness were 2.37 for `SERP`, 2.42 for `RR`, and 2.36 for `Chatbot`.

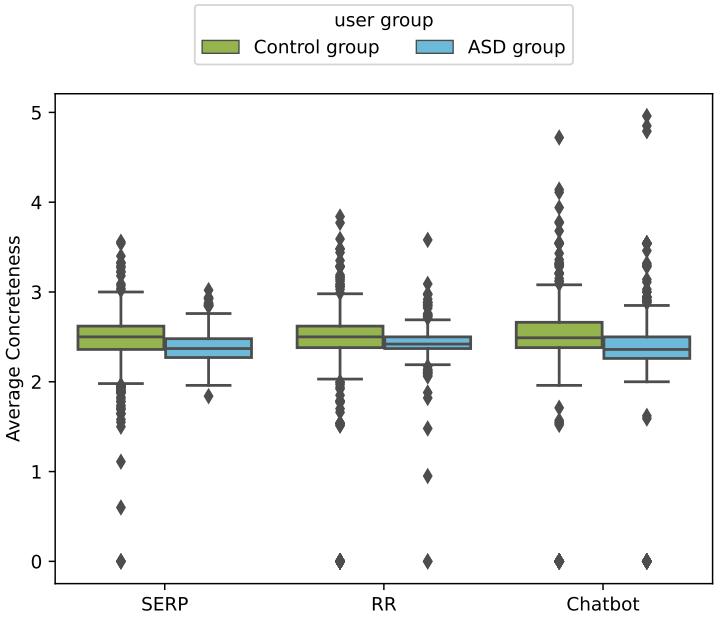
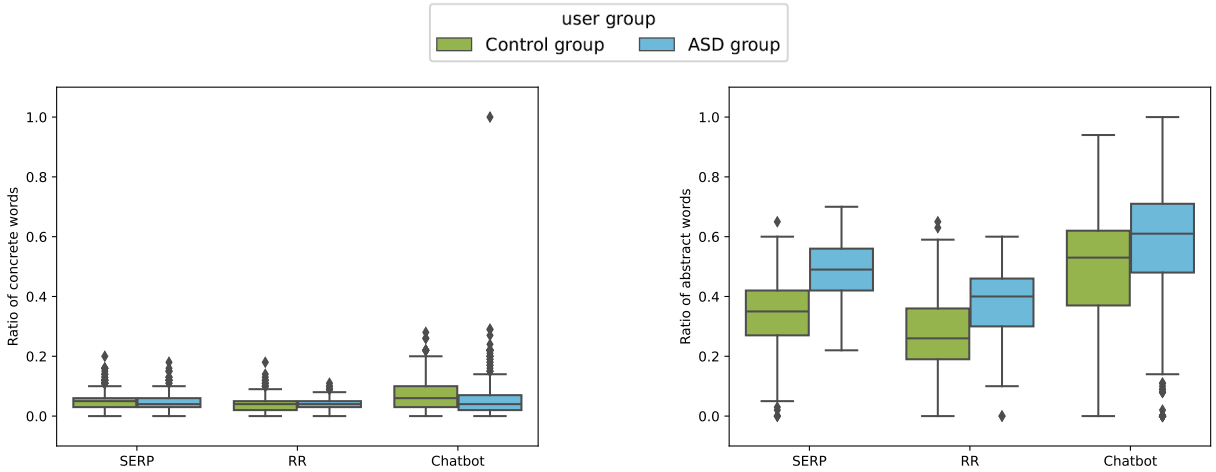


Figure 5.5: Average concreteness of `SERP`, `RR`, and `Chabot` for Control and ASD group queries.

Furthermore, the ratio of concrete words was significantly higher for `SERP` and `Chabot` for the control group (median value of 0.05 for `SERP` and 0.06 for `Chatbot`) queries compared to ASD group queries (median value of 0.04 for both `SERP` and `Chatbot`). Although the median value of the ratio of concrete words in `RR` responses was the same for control and ASD group queries.

The ratio of abstract words was significantly lower for control group queries with a median value of 0.35

for SERP, 0.26 for RR, and 0.53 for Chabot respectively, compared to ASD group queries with a median value of 0.49 for SERP, 0.4 for RR, and 0.61 for Chatbot.



(a) Ratio of concrete words

(b) Ratio of abstract words

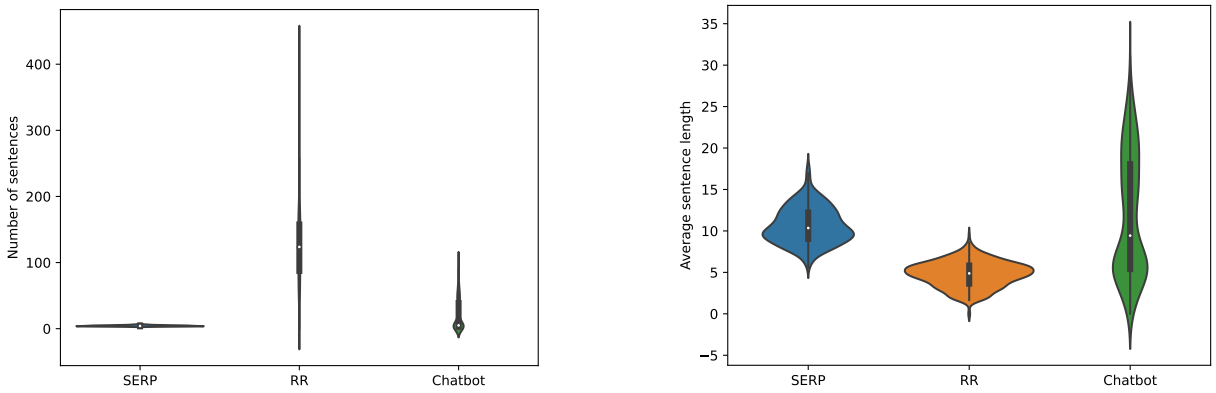
Figure 5.6: Analysis of text concreteness based on SERP, RR, and Chabot for Control and ASD group queries.

5.1.2. Comparing SERP, RR, and Chatbot responses for ASD group queries

To investigate which type of online IST catered to the needs of autistic users better, we first compared the textual content of the responses collected from the three response group types (as described in Section 3.2), namely the SERP, RR, and Chatbot responses. The results of our comparisons have been divided into three parts to report the differences in the (1) structure, (2) readability, and (3) concreteness of the responses provided by the three response group types. Note that all distributions were found to be significantly different.

Text Structure. For a response to be considered accessible to autistic users, the textual content must be kept brief and structured into lists and short paragraphs preceded by relevant headings.

We observed that RR and Chabot had significantly longer sentences (median: 10.35 and 9.44 respectively) on average than RR (median: 4.88). Chabot also had a large variance in its distribution as seen in Figure 5.7b. However in Figure 5.7a we see that the number of sentences in SERP and Chabot is much lower (median: 4.13 and 5 respectively) in comparison to the number of sentences in RR (median: 123.64). Although SERP and Chabot had slightly longer sentences than RR responses, RR responses had substantially more sentences than SERP and Chabot, thus we conclude that RR responses were the most content-heavy among the three response group types.

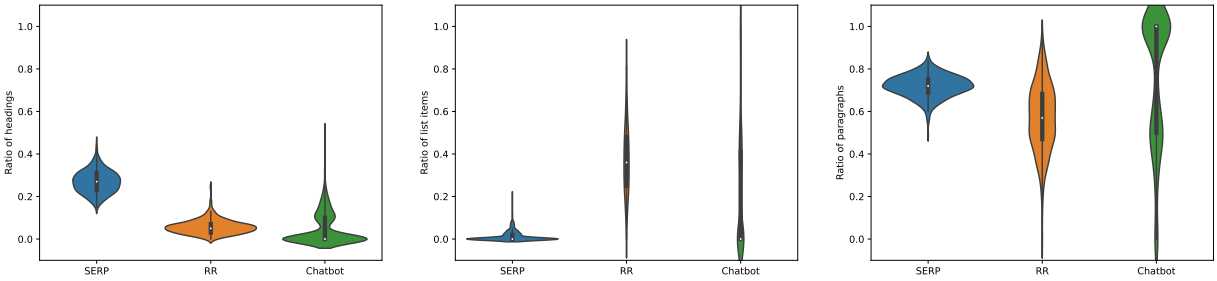


(a) Number of sentences

(b) Average sentence length

Figure 5.7: Analysis of text structure (sentence-level) based on responses collected from `SERP`, `RR`, and `Chatbot`.

From Figures 5.8a-5.8c we see that sentences in responses from `SERP` are mostly divided as headings and paragraphs (median: 0.27 and 0.72 respectively). The sentences in `RR` on the contrary are structured more often as list items and paragraphs (median: 0.36 and 0.57 respectively) although we observed a large variance in the ratio of list items for `RR` responses as seen in Figure 5.8b. In the case of chatbots, the sentences appear to be predominantly divided into paragraphs (median: 1.00) with very few responses having sentences structured as headings and list items (median: 0 for both). In Figure 5.9, we see that the average paragraph length of `Chabot` is mostly lower than that of `SERP` and `RR`, although we also observe a large variance in the distribution for average paragraph length of `Chatbot`.



(a) Ratio of headings

(b) Ratio of list items

(c) Ratio of paragraphs

Figure 5.8: Analysis of text structure (full body) based on responses collected from `SERP`, `RR`, and `Chatbot`.

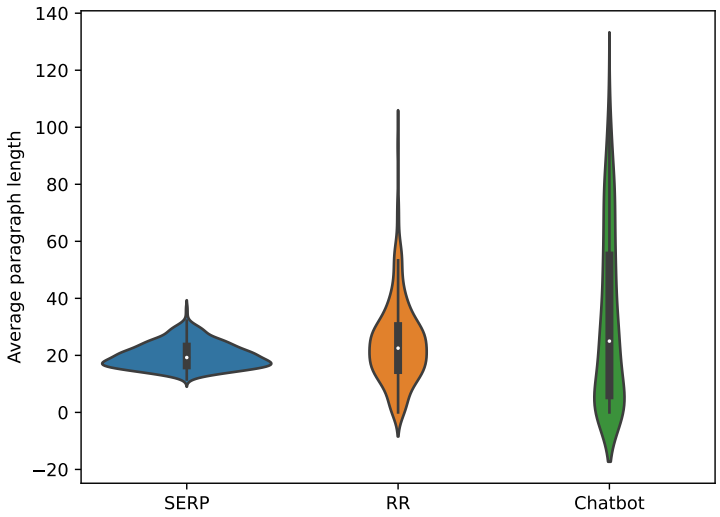
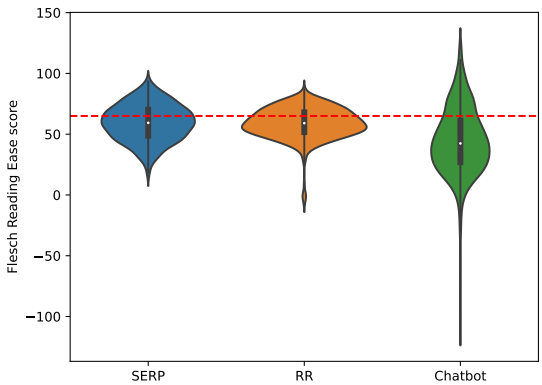


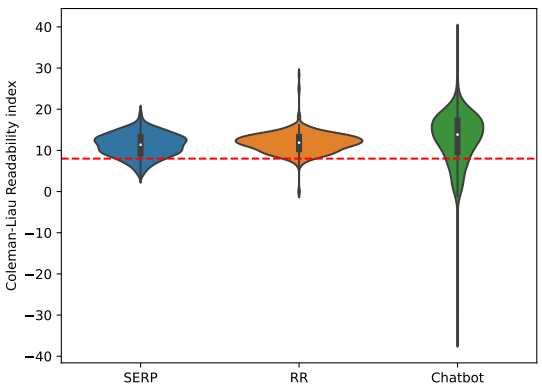
Figure 5.9: Average paragraph length of SERP, RR, and Chabot (An outlier has been removed from RR for visualisation purposes).

Text Readability. As discussed in Section 3.3, for a text to be readable to autistic users the Flesch reading ease score should be greater than or equal to 65 [95] i.e. the text should be readable by 8^{th} and 9^{th} graders [27]. However from Figure 5.10a we can see that most responses collected from all three response group types scored much lower than the ideal value (as indicated by the red dashed line). The Chatbot responses specifically had a much lower Flesch reading ease score with a median value of 42.41 which would be considered a text suitable only for a college student or higher. SERP and RR with median values at 59.32 and 59.06 respectively were comparatively easier to read although still not accessible to autistic users.

For a text to be readable by 8^{th} and 9^{th} graders, which is the ideal readability level for autistic users [95] the Coleman-Liau readability index of the text should be 8 or below [20]. From Figure 5.10b we can see that nearly all responses collected from all three response group types scored much higher than 8 (indicated by the red dashed line in the violin plot). The responses of SERP and RR had a median value of 11.36 and 11.88 respectively while the Chatbot responses had a median value of 13.86 which would mean the text is suitable for 11^{th} grade and higher students. Thus according to their Coleman-Liau readability index values, nearly all the responses collected from all three response group types are inaccessible to autistic users.



(a) Flesch Reading Ease score



(b) Coleman-Liau Readability Index (An outlier has been removed from Chabot for visualisation purposes)

Figure 5.10: Analysis of text readability based on SERP, RR and Chabot responses.

Text concreteness. For a text to be readable by autistic users, abstract words should be kept at a minimum and the text should be as concrete as possible. The average concreteness of responses collected from all three response group types displayed a similar distribution as observed in Figure 5.11 with `SERP`, `RR`, and `Chabot` having median values at 2.37, 2.42, and 2.36 respectively. This indicates that the textual content of `RR` is the most concrete on average but only by a very small margin compared to the other two response group types.

Even the ratio of concrete words is similar for all three response group types as observed in Figure 5.12a with all their median values at 0.04. However from Figure 5.12b we can see that the `Chatbot` responses had a much larger ratio of abstract words (median: 0.61) compared to `SERP` and `RR` (median: 0.49 and 0.4 respectively).

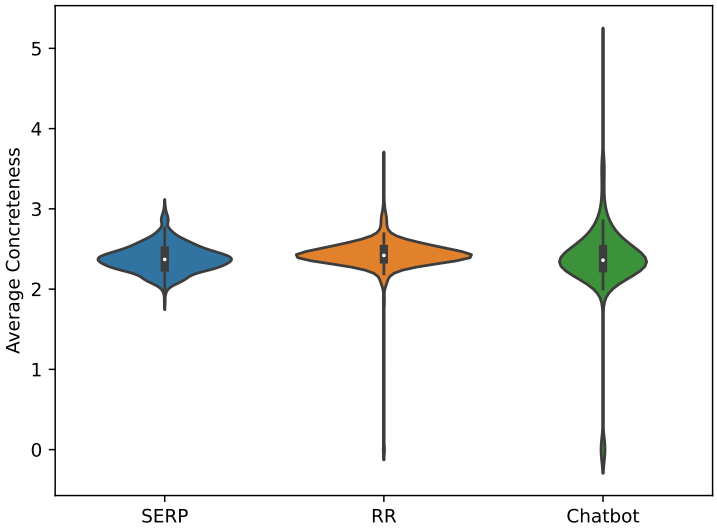
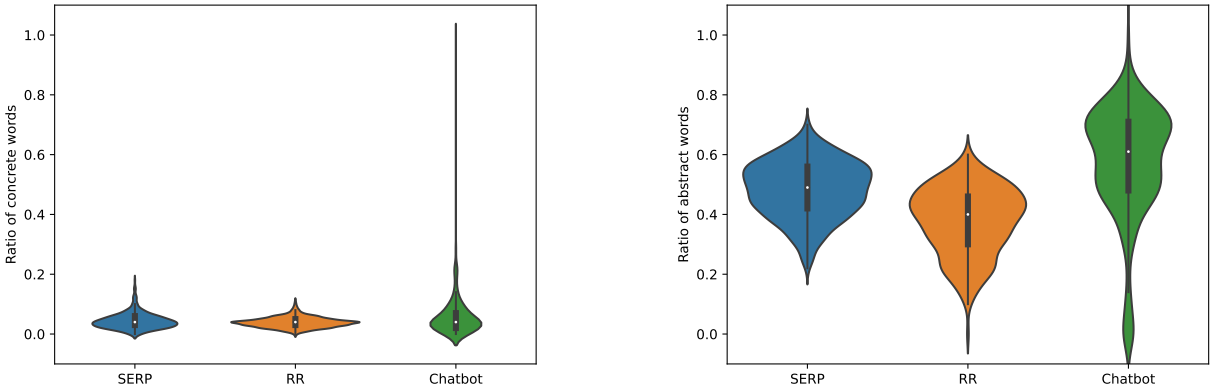


Figure 5.11: Average concreteness of `SERP`, `RR`, and `Chatbot`.



(a) Ratio of concrete words

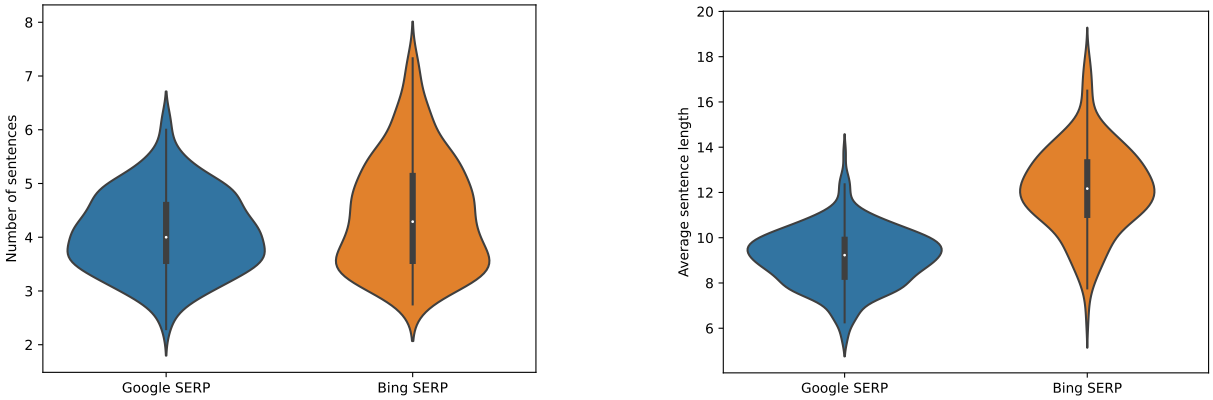
(b) Ratio of abstract words

Figure 5.12: Analysis of text concreteness based on `SERP`, `RR` and `Chabot` responses.

5.1.3. Comparing `Google SERP` and `Bing SERP` response for ASD group queries responses

To investigate the variance in the scores of the accessibility indicators for `SERP` responses, we dig deeper and compare the accessibility indicator scores for `Google SERP` and `Bing SERP` responses. The results are once again divided into three parts that compare the responses over the three accessibility categories described previously.

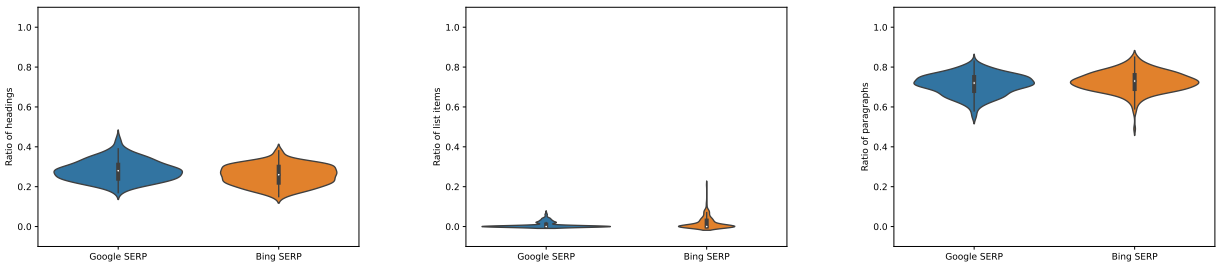
Text structure. `Google SERP` and `Bing SERP` had significantly different distributions for the number of sentences (as seen in Figure 5.13a), but their median values differed only by a small margin with `Google SERP` responses having lesser number of sentences (4 for `Google SERP` and 4.29 for `Bing SERP`). The sentences were also significantly shorter on average for `Google SERP` (median: 9.23) compared to `Bing SERP` (median: 12.17) as seen in Figure 5.13b.



(a) Number of Sentences **(b)** Average sentence length

Figure 5.13: Analysis of text structure (sentence-level) based on `Google SERP` and `Bing SERP` responses.

From Figures 5.14a-5.14c we can see that the distribution of sentences amongst headings, list items and paragraphs is similar for `Google SERP` and `Bing SERP` although the distributions tested to be significantly different from each other for all three ratios. On the other hand, `Google SERP` had significantly shorter paragraphs (median: 16.73) compared to `Bing SERP` (median: 23.06) as seen in Figure 5.15. Thus overall the textual content is similarly structured for both `Google SERP` and `Bing SERP` however `Google SERP` had significantly shorter sentences and paragraphs on average compared to `Bing SERP`.



(a) Ratio of headings **(b)** Ratio of list items **(c)** Ratio of paragraphs

Figure 5.14: Analysis of text structure (full body) based on `Google SERP` and `Bing SERP` responses.

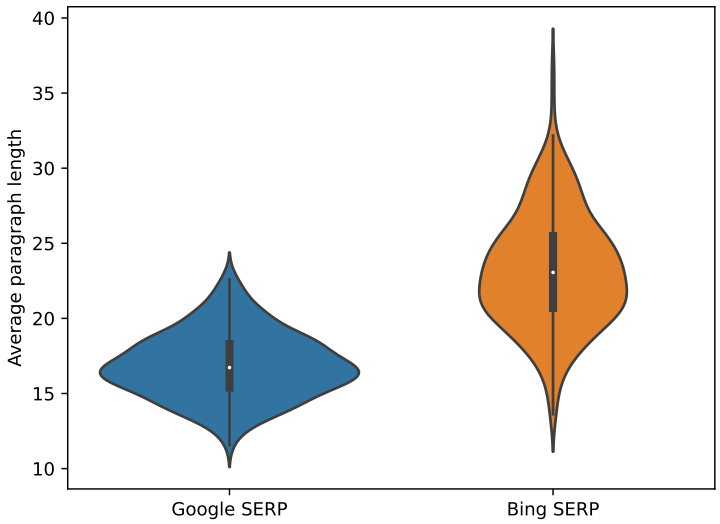
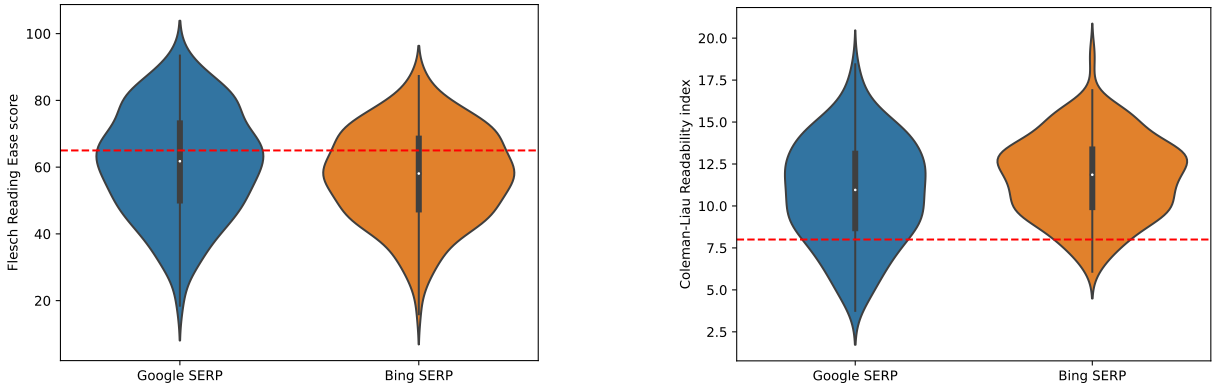


Figure 5.15: Average paragraph length of responses collected from `Google SERP` and `Bing SERP`.

Text readability. From Figure 5.16a we can see that most of `Google SERP` and `Bing SERP` responses scored much below the ideal value of 65 for the Flesch reading ease score. However, `Google SERP` had a significantly higher median at 61.76 compared to `Bing SERP` whose median score was at 58.08. `Google SERP` also scored slightly better on the Coleman-Liau readability index with a median value of 10.96 compared to `Bing SERP` with a median value of 11.86. However for the text to be accessible to autistic users the index value should be 8 or lower and thus both the SEs `SERP` failed to achieve ideal scores for both readability scores.



(a) Flesch Reading Ease score

(b) Coleman-Liau Readability Index

Figure 5.16: Analysis of text readability based on `Google SERP` and `Bing SERP` responses.

Text concreteness. `Google SERP` and `Bing SERP` had similar distributions for average concreteness (Figure 5.17) and their median values were also nearly the same (2.4 and 2.34 respectively). The distributions were also found to be not significantly different from each other. `Google SERP` and `Bing SERP` responses had similar ratios of concrete words as seen in Figure 5.18a (both medians at 0.04) but `Bing SERP` had a significantly higher ratio of abstract words with median score of 0.55, compared to `Google SERP` whose median score was 0.44. (Figure 5.18b).

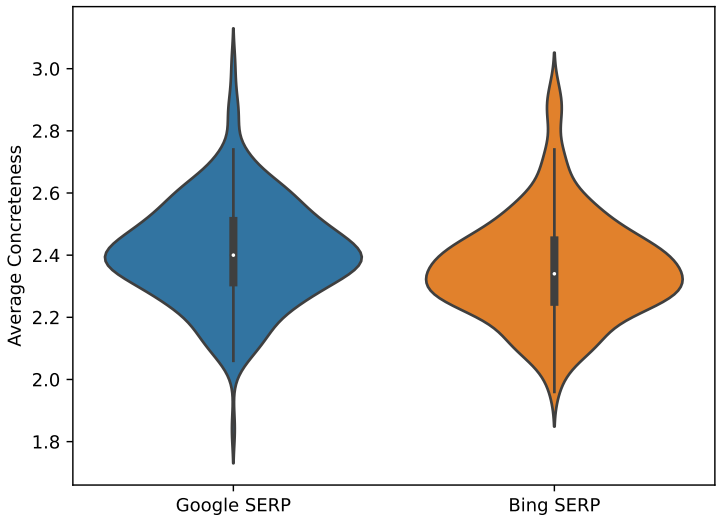
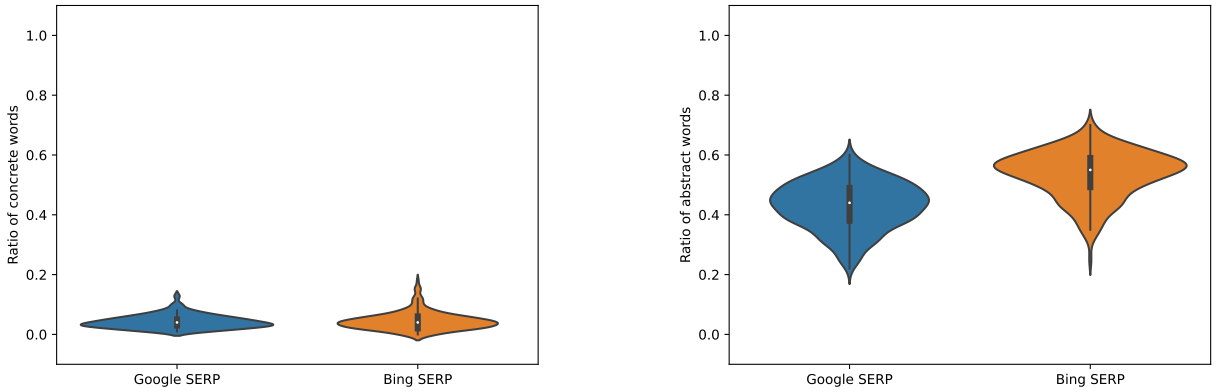


Figure 5.17: Average concreteness of `Google SERP` and `Bing SERP` responses (Distributions were not significantly different).



(a) Ratio of concrete words

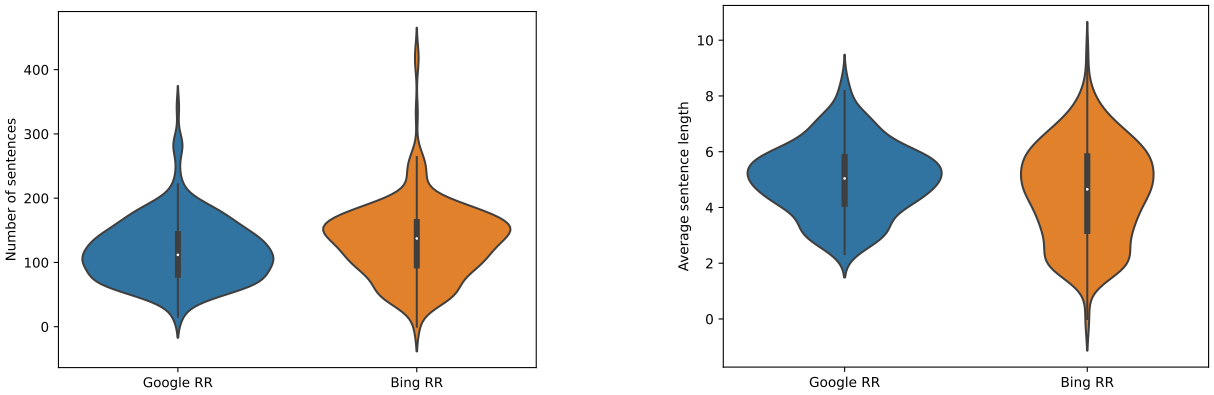
(b) Ratio of abstract words

Figure 5.18: Analysis of text concreteness based on `Google SERP` and `Bing SERP` responses.

5.1.4. Comparing `Google RR` and `Bing RR` responses for ASD group queries

Given the significant differences in accessibility indicator scores for `Google SERP` and `Bing SERP` responses, we also investigated the difference in the indicator scores for `Google RR` and `Bing RR` responses. We report the results of our comparison under the same three categories of accessibility as described previously.

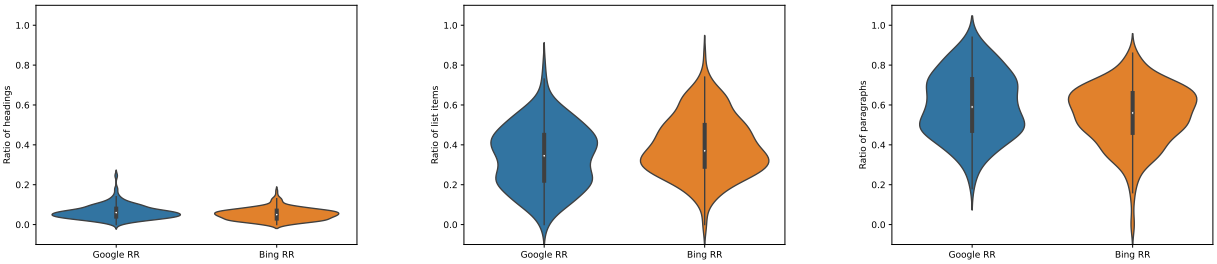
Text structure. `Google RR` had a significantly lesser number of sentences (median: 111.8) compared to `Bing RR` (median: 137.29) (Figure 5.19a) but the sentences were significantly longer on average for `Google RR` responses (median: 5.04) compared to the sentences in `Bing RR` responses (median: 4.65). The distribution of average sentence length was also spread over a wider range for `Bing RR` (with the minimum reaching 0 due to our code possibly failing to scrape the textual content from certain web resources) than for `Google RR` as seen in Figure 5.19b.



(a) Number of Sentences **(b)** Average sentence length

Figure 5.19: Analysis of text structure (sentence-level) based on `Google RR` and `Bing RR` responses.

The distributions of the ratios of headings, list items, and paragraphs were found to be significantly different for `Google RR` and `Bing RR` responses. The median value for the ratio of headings was 0.06 for `Google RR` and 0.05 for `Bing RR` responses respectively. For the ratio of list items, a lot of responses from `Bing RR` centred around a lower ratio of list items compared to `Google RR` as observed in Figure 5.20b. Their median values however were found to have a very small margin of difference (0.34 for `Google RR` and 0.37 for `Bing RR` responses). From Figure 5.20c we can see that `Google RR` had a lot of responses centered around a lower ratio of paragraphs compared to `Bing RR` although the median value for `Google RR` was significantly higher than `Bing RR` (0.59 and 0.56 respectively).



(a) Ratio of headings **(b)** Ratio of list items **(c)** Ratio of paragraphs

Figure 5.20: Analysis of text structure (full body) based on `Google RR` and `Bing RR` responses.

The distribution of average paragraph length for `Google RR` and `Bing RR` responses were also found to be significantly different from each other. The average paragraph length for most responses for `Google RR` is concentrated around a lower value compared to `Bing RR`. However, `Google RR` has a significantly higher median at 22.78 compared to `Bing RR` with a median value of 22.44 although the difference is only by a small margin.

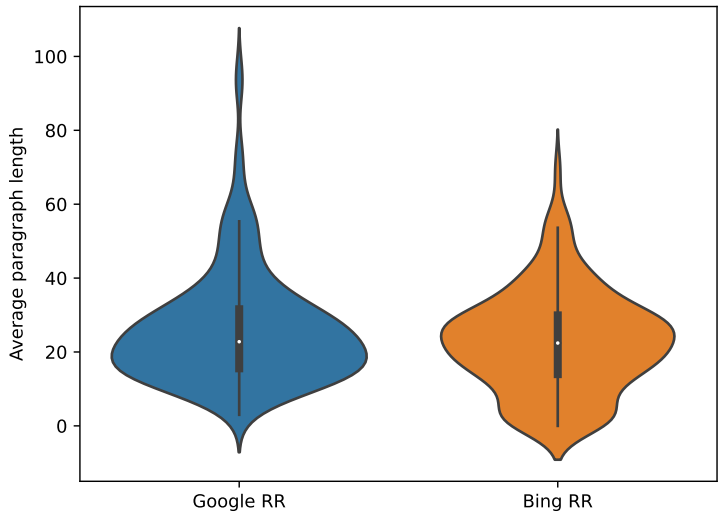
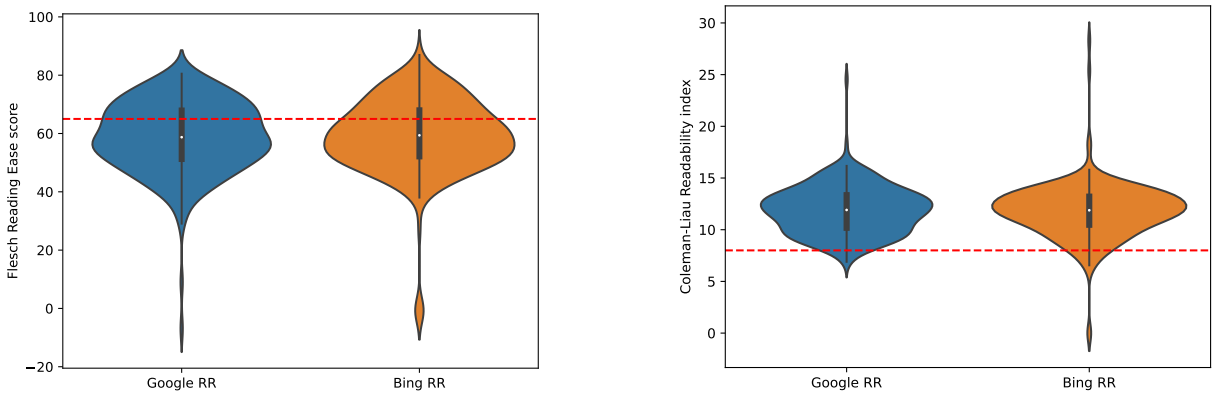


Figure 5.21: Average paragraph length of `Google RR` and `Bing RR` responses (An outlier has been removed from `Bing RR` for visualisation purposes) (Distributions were not significantly different).

Text readability. Both `Google RR` and `Bing RR` had similar distributions for Flesch reading ease score and the distributions were also not significantly different from each other. Most of the responses for both `Google RR` and `Bing RR` scored below 65 as seen in Figure 5.22a with the median value at 58.73 for `Google RR` and 59.38 for `Bing RR`. Even for the Coleman-Liau readability index most responses collected from `Google RR` and `Bing RR` scored higher than the ideal value of 8 as seen in Figure 5.22b with `Google RR`'s median value at 11.91 and `Bing RR`'s median value at 11.88.



(a) Flesch Reading Ease score

(b) Coleman-Liau Readability Index

Figure 5.22: Analysis of text readability based on `Google RR` and `Bing RR` responses (Distributions were not significantly different for `Google RR` and `Bing RR` responses for either text readability indicators).

Text concreteness. `Google RR` and `Bing RR` had similar distributions for all three markers for text concreteness i.e. for average concreteness, the ratio of concrete words and the ratio of abstract words (Figures 5.23-5.24a) with each pair of distribution not significantly different from each other. Even their median values were found to be similar for all three metrics - 2.42 for average concreteness, 0.04 for ratio of concrete words, and 0.4 and 0.38 for ratio of abstract words for `Google RR` and `Bing RR` respectively.

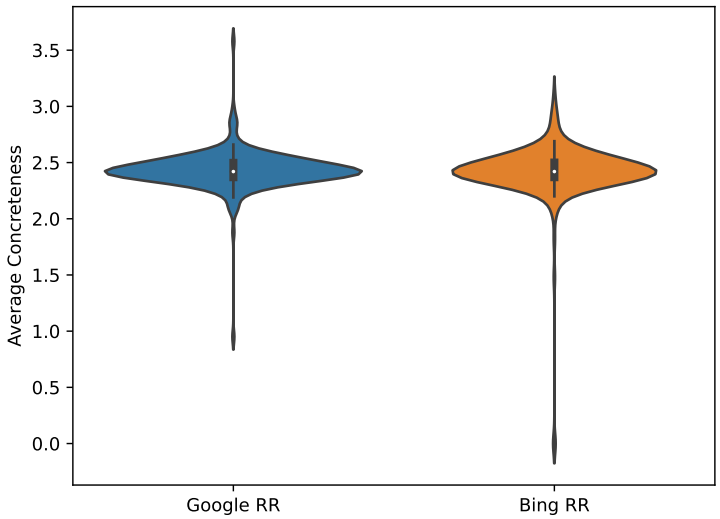
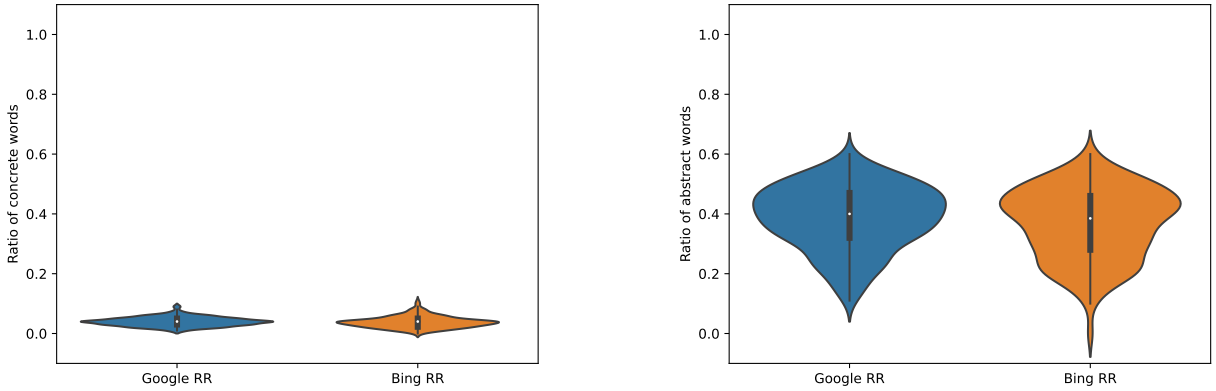


Figure 5.23: Average concreteness of `Google RR` and `Bing RR`.



(a) Ratio of concrete words

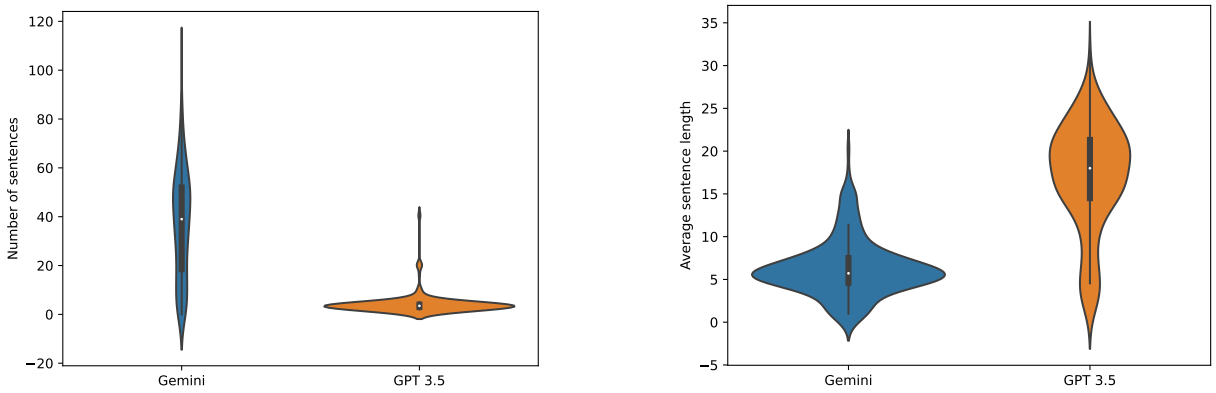
(b) Ratio of abstract words

Figure 5.24: Analysis of text concreteness based on `Google RR` and `Bing RR` responses (Distributions were not significantly different for `Google RR` and `Bing RR` responses for any of the text concreteness indicators).

5.1.5. Comparing `Gemini` and `GPT 3.5` responses for ASD group queries

The large variances in the distribution of several accessibility indicators for `Chabot` responses in the general comparisons (as reported in Section 5.1.2) necessitated further investigation to find potential reasons for the variance. Thus, we compare the responses generated by `Gemini` and `GPT 3.5` to investigate any significant differences in the accessibility indicators.

Text structure. `Gemini` responses had significantly more sentences (median: 39) compared to `GPT 3.5` (median: 3.5) and `Gemini` responses also had a much larger variance in the number of sentences compared to `GPT 3.5` responses as seen in Figure 5.25a. However, the sentences in `Gemini` responses were significantly shorter on average (median: 5.72) than `GPT 3.5` (median: 18) and from Figure 5.25b we can see that `GPT 3.5` had more variance in its responses for average sentence length compared to `Gemini` response.

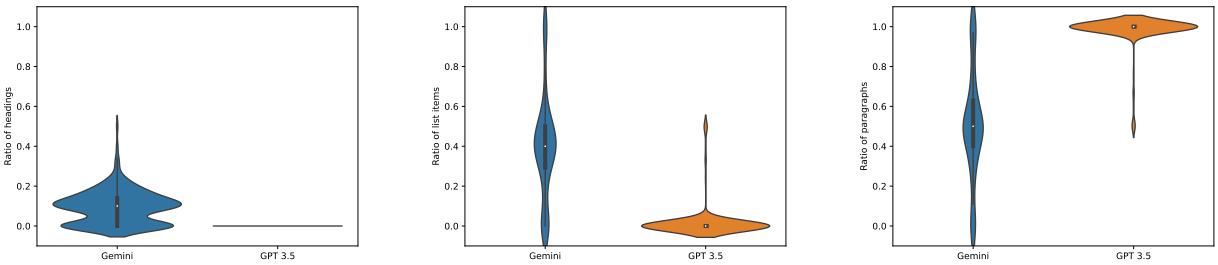


(a) Number of sentences

(b) Average sentence length

Figure 5.25: Analysis of text structure (sentence-level) based on `Gemini` and `GPT 3.5` responses.

The distributions for all three ratios under text structure - the ratio of headings, the ratio of list items, and the ratio of paragraphs, were significantly different for `Gemini` and `GPT 3.5` responses. `Gemini` responses had more diversity in their structure with a median value of 0.1 for the ratio of headings, 0.4 for the ratio of list items and 0.5 for the ratio of paragraphs. On the other hand, `GPT 3.5` responses were predominantly structured as paragraphs with a median value of 0 for both ratios of headings and of list items and a median value of 1 for the ratio of paragraphs. However, the distributions for the ratio of list items and the ratio of paragraphs were observed to have a high variance for `Gemini` responses as can be seen in Figures 5.26b and 5.26c respectively while all 3 ratios for `GPT 3.5` responses had lesser variance compared to `Gemini` responses as seen in Figures 5.26a - 5.26c.



(a) Ratio of headings

(b) Ratio of list items

(c) Ratio of paragraphs

Figure 5.26: Analysis of text structure (full body) based on `Gemini` and `GPT 3.5` respones.

`Gemini` responses also had significantly shorter paragraphs (median: 9) on average compared to `GPT 3.5` (median: 55) with `GPT 3.5` responses showing a much larger variance than `Gemini` responses.

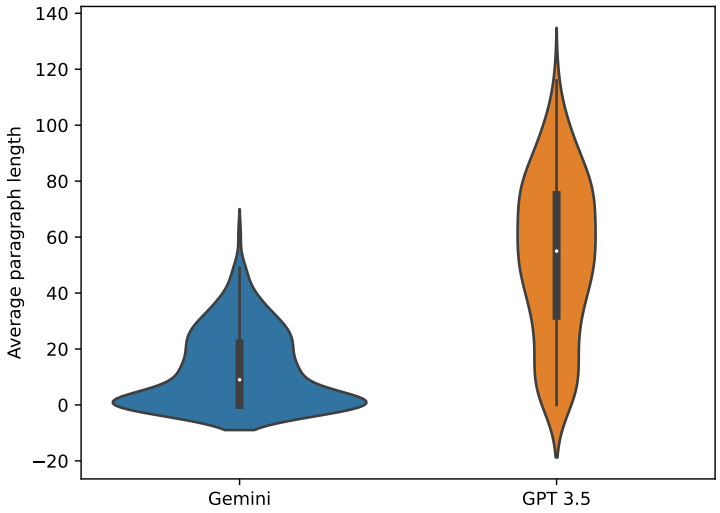
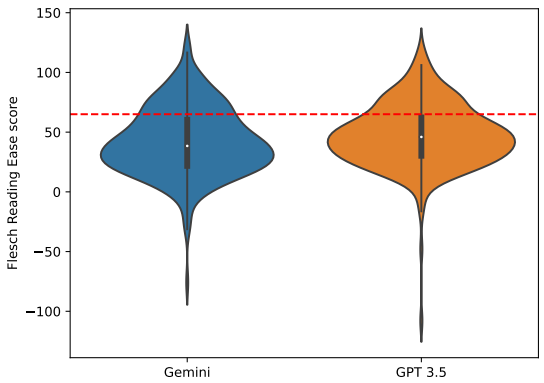


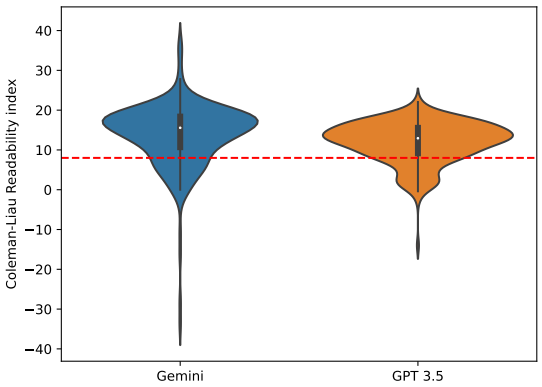
Figure 5.27: Average paragraph length of `Gemini` and `GPT 3.5` responses.

Text readability. Both `Gemini` and `GPT 3.5` responses scored much more poorly on the readability scores compared to the responses collected from `Google SERP` and `Bing SERP` and `Google RR` and `Bing RR`.

`Gemini` responses had a significantly lower median value for the Flesch reading ease score (38.45) than `GPT 3.5` responses (45.98). But both of the response groups scored much below the ideal score of 65 which means that most of the responses generated by both the chatbots are readable only by college students. For the Coleman-Liau readability index, the distributions were significantly different with the `GPT 3.5` responses being relatively closer to the ideal value of 8 (median value of 12.97) while the `Gemini` responses have a median score of 15.57. The Coleman-Liau readability index values also indicate that most of the responses generated by both chatbots are readable by only college students. Thus neither of `Gemini` responses `GPT 3.5` responses were accessible to autistic users.



(a) Flesch Reading Ease score



(b) Coleman-Liau Readability Index (An outlier has been removed from `GPT 3.5` for visualisation purposes)

Figure 5.28: Analysis of text readability based on `Gemini` and `GPT 3.5` respones.

Text concreteness. `Gemini` and `GPT 3.5` responses had similar average concreteness with their median concreteness rating at 2.37 and 2.34 respectively. The distributions also had similar variance as seen from Figure 5.29 and the distributions were also not significantly different. The distribution of the ratio of concrete words and the ratio of abstract words were however significantly different for `Gemini` and `GPT 3.5` responses. The distribution for the ratio of concrete words had a lesser variance for both `Gemini` and `GPT`

3.5 responses with median values 0.04 for Gemini and 0.05 for GPT 3.5 (Figure 5.30a). It is to be noted that most of Gemini responses had a noticeably lower ratio of abstract words (median: 0.5) compared to responses generated by GPT 3.5 (median: 0.7) as seen in Figure 5.30b.

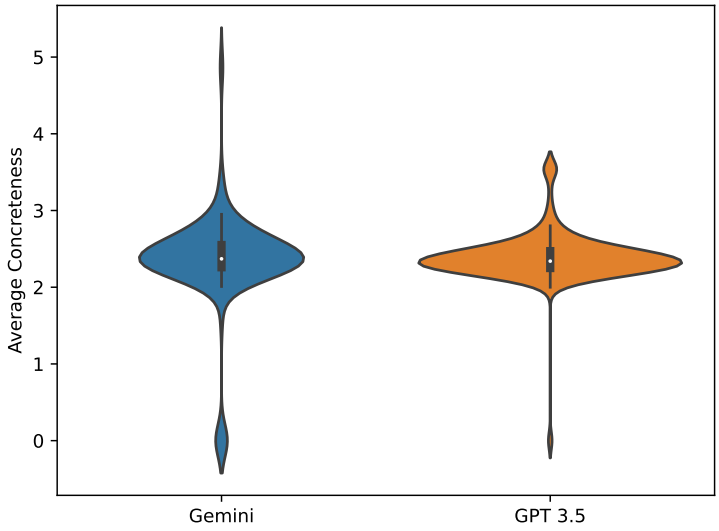
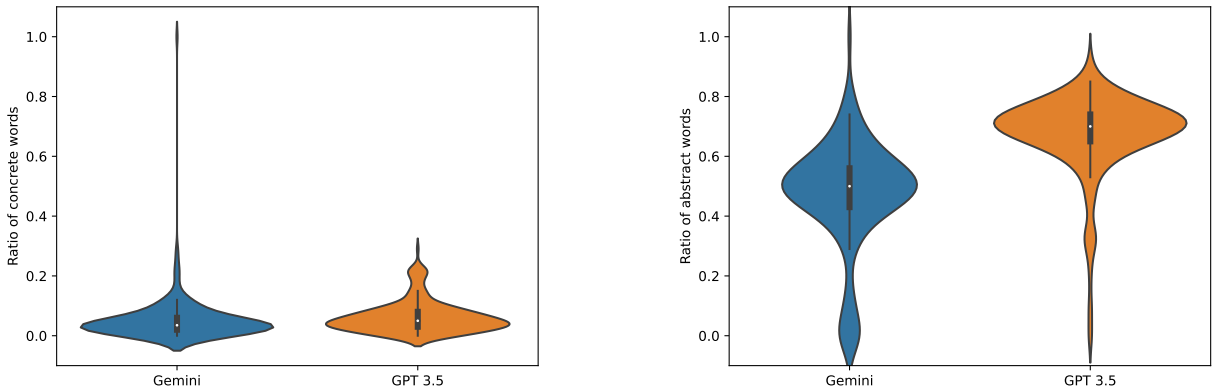


Figure 5.29: Average concreteness of Gemini and GPT 3.5 responses (Distributions were not significantly different).



(a) Ratio of concrete words

(b) Ratio of abstract words

Figure 5.30: Analysis of text concreteness based on Gemini and GPT 3.5 respones.

5.1.6. Investigating the effect of n-gram length of query/prompt on the accessibility of response groups for ASD group queries

While we did note significant differences in the SERP,RR, and Chatbot responses individually which could explain the variance in the accessibility indicators observed in general comparisons (reported in Section 5.1.2), it is also worth investigating whether the variance could also be due to the nature of the query. For our first investigation, we inspect the effect of the length of the query or prompt on the (1) structure, (2) readability, and (3) concreteness of all response groups described in Section 3.2.

Text structure. Google SERP, Bing SERP, and GPT 3.5 had consistent distributions for the number of sentences in their responses for varying query/prompt lengths. In contrast, Google RR, Bing RR and Gemini responses showed noticeable inconsistency across their distributions as seen in Figure 5.31. However, the distributions were not significantly different for Google RR and Gemini responses across

different query/prompt lengths. There also appears to be a general trend of the number of sentences being more for longer queries/prompts although for `Bing RR` the number of sentences drops dramatically for queries longer than 4-grams.

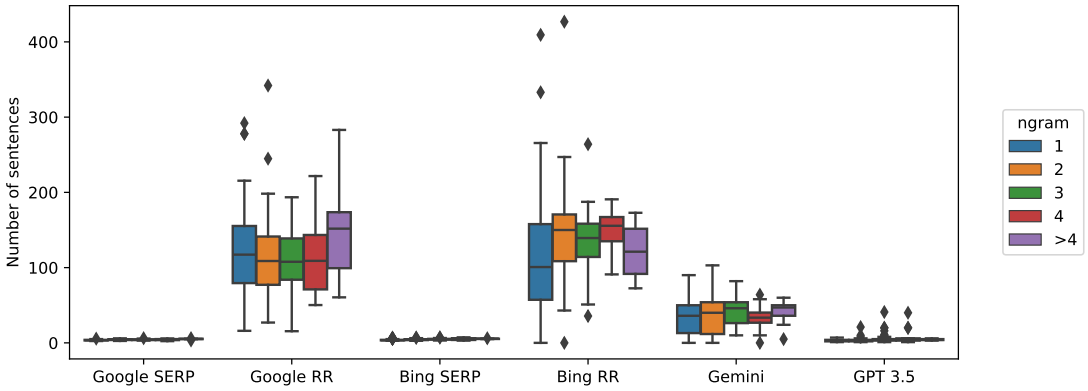


Figure 5.31: Number of sentences in each response group across varying query n-gram length (Differences across varied query/prompt lengths were not significant for `Google RR` and `Gemini`).

`GPT 3.5` responses had the largest variances for average sentence length across all the prompt length categories compared to the other response groups as seen in Figure 5.32. Unigram queries and prompts also resulted in shorter sentences on average across all response groups except for `Google SERP` and `Bing SERP` where the unigram queries resulted in longer sentences than other n-gram length queries. However, the distributions were not significantly different for `Bing SERP`.

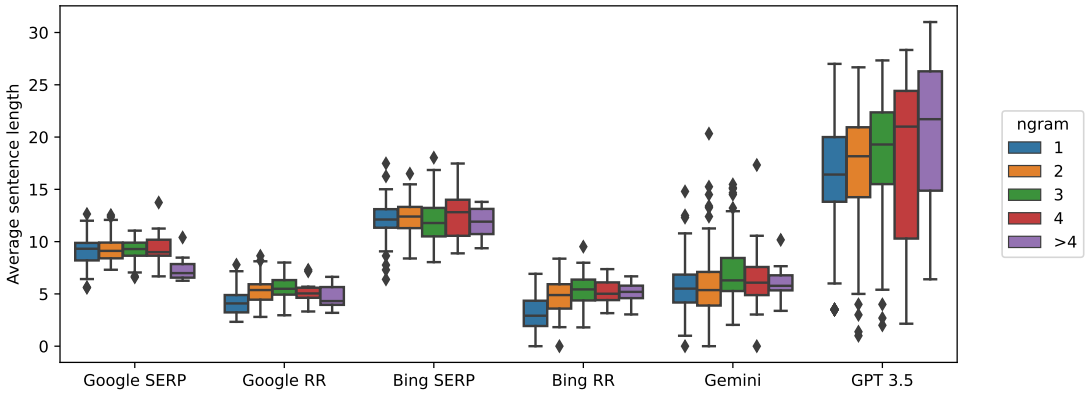


Figure 5.32: Average sentence length in each response group across varying query/prompt n-gram length (Differences across varied query/prompt lengths were not significant for `Bing SERP`).

The ratio of headings generally decreased with increasing n-gram length of queries or prompts as seen in Figure 5.33. `Google SERP` and `Bing SERP` displayed a higher ratio of headings across all n-gram lengths compared to the other response groups. The ratio of headings in `Gemini` responses was more varied than the responses from other response groups.

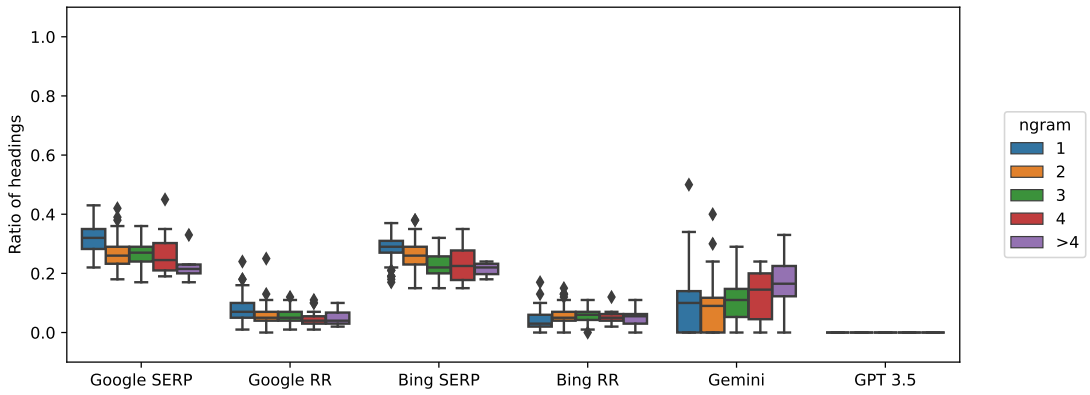


Figure 5.33: Ratio of headings in each response group across varying query/prompt n-gram length.

The distributions for the ratio of list items observed larger variances across different n-gram lengths for `Google RR`, `Bing RR` and `Gemini` responses with unigram prompts in `Gemini` responses having the largest variance (Figure 5.34). Despite the large variance, different n-gram prompts resulted in the same median value of the ratio of list items, and the distributions were also not significantly different for `Gemini`. Similarly, the distributions for `Google SERP` and `Google RR` respectively.

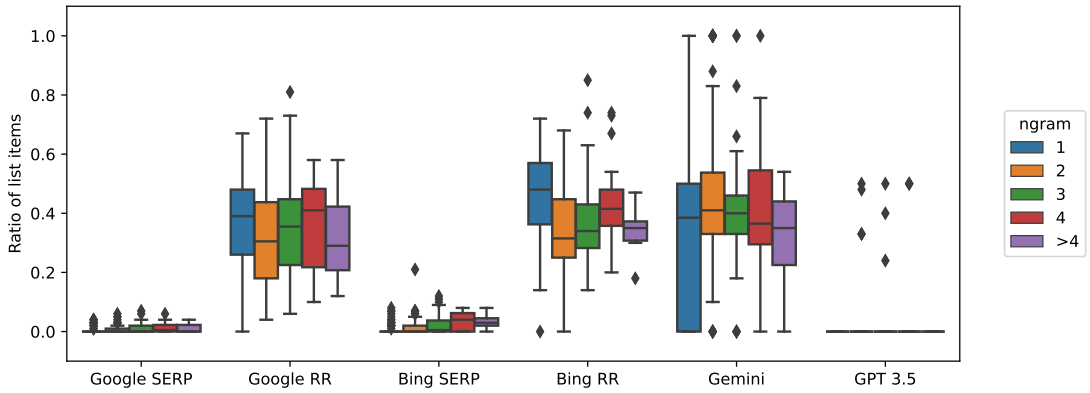


Figure 5.34: Ratio of list items in each response group across varying query/prompt n-gram length (Differences across varied query/prompt lengths were not significant for `Google SERP`, `Google RR` and `Gemini`).

For the ratio of paragraphs, all response groups had noticeable variance in their distributions across different n-gram lengths except for `GPT 3.5` responses which had consistent values centred at 1 across all n-gram length prompts. `Google RR`, `Bing RR`, and `Gemini` responses once again had the largest variances across their respective distributions compared to the other response groups with unigram prompts in `Gemini` responses displaying the largest variance as seen in Figure 5.35, although the distributions for `Gemini` responses were tested to be not significantly different from each other. The distributions for `Bing SERP` were concentrated at the highest ratio of paragraphs across all n-gram length queries.

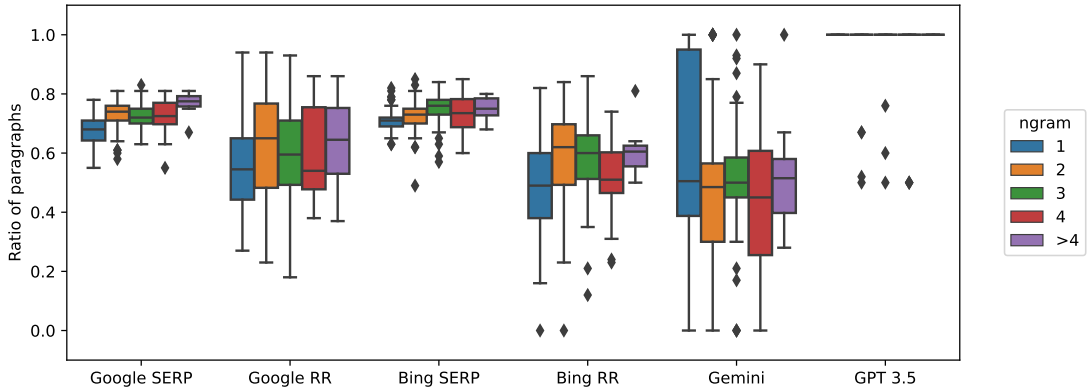


Figure 5.35: Ratio of paragraphs in each response group across varying query/prompt n-gram length (Differences across varied query/prompt lengths were not significant for `Gemini`).

The average paragraph length was the highest for `GPT 3.5` responses across all n-gram lengths and the distributions also had the largest variance as seen in Figure 5.36. `Google SERP`, `Google RR`, `Bing SERP`, and `Bing RR` all had shorter paragraphs on average for unigram queries than queries of longer lengths. In the case of `Chatbot`, both `Gemini` and `GPT 3.5` responses had the shortest paragraphs on average for bigram queries. However, the distributions for `Gemini` responses were not significantly different from each other.

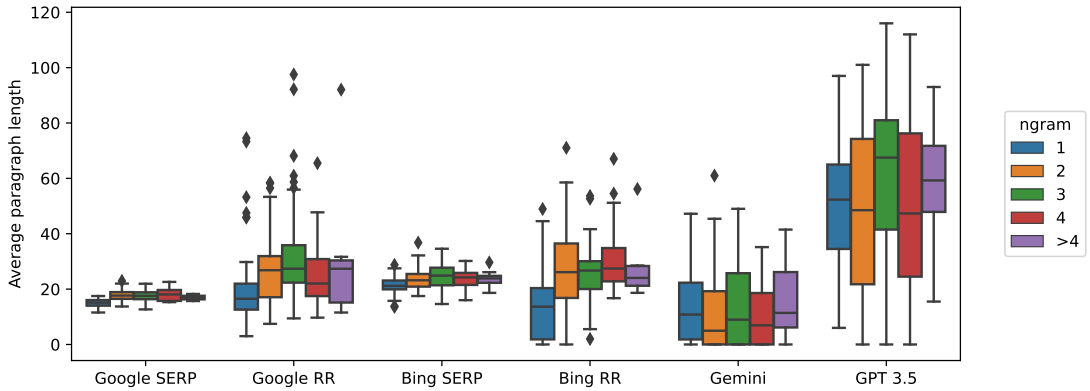


Figure 5.36: Average paragraph length in each response group across varying query/prompt n-gram length (An outlier has been removed from `Bing RR` for visualisation purposes) (Differences across varied query/prompt lengths were not significant for `Gemini`).

Text readability. As discussed previously, for a text to be readable by an autistic user, the text should have a Flesch reading ease score of at least 65 [95] and a Coleman-Liau readability index of at most 8.

Only the responses collected from `Google SERP` for bigrams and 4-gram queries had many responses that crossed the minimum threshold of 65 for the Flesch reading ease score as seen in Figure 5.37. `Gemini` and `GPT 3.5` scored much lower on the indicator than the other response groups. Furthermore, `Gemini` and `GPT 3.5` responses for unigram prompts observed the largest variance compared to all other distributions including those of other response groups. Their responses for prompts of length $\geq$ 3-gram scored much lower in Flesch reading ease score compared to all other distributions with the responses for prompts longer than 4-gram scoring the lowest. The distributions for `Google SERP`, `Google RR`, `Bing SERP`, and `Bing RR` are comparatively more consistent across varying query lengths. However, the distributions were found to be significantly different from each other for `Google SERP` and `GPT 3.5` responses only.

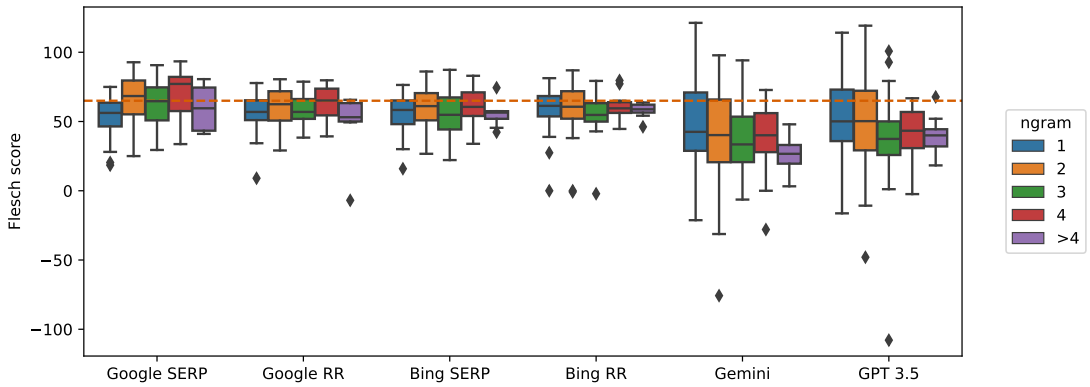


Figure 5.37: Flesch Reading Ease score in each response group across varying query/prompt n-gram length (Differences across varied query/prompt lengths were significant only for `Google SERP` and `GPT 3.5`, $p < 0.05$).

According to the observed distributions for the Coleman-Liau readability index, most of the responses collected from all the response groups are not readable by autistic users as all responses scored much higher than the maximum threshold of 8 as seen in Figure 5.38. `Gemini` and `GPT 3.5` scored much higher than the other response groups.

The scores for `Google SERP`, `Google RR`, `Bing SERP`, and `Bing RR` were mostly consistent across varying query lengths and the distributions were found to be not significantly different from each other for `Google RR` and `Bing SERP` responses respectively. On the other hand, `GPT 3.5` and `Gemini` showed substantial variance in their distributions, especially for responses generated for unigram and bigram queries. Furthermore, similar to the trend observed in the Flesch reading ease score, `Gemini` and `GPT 3.5` scored much higher in the Coleman-Liau readability index compared to all other distributions for prompts of length $\geq$ 3-gram, with responses generated for prompts longer than 4-gram scoring the highest.

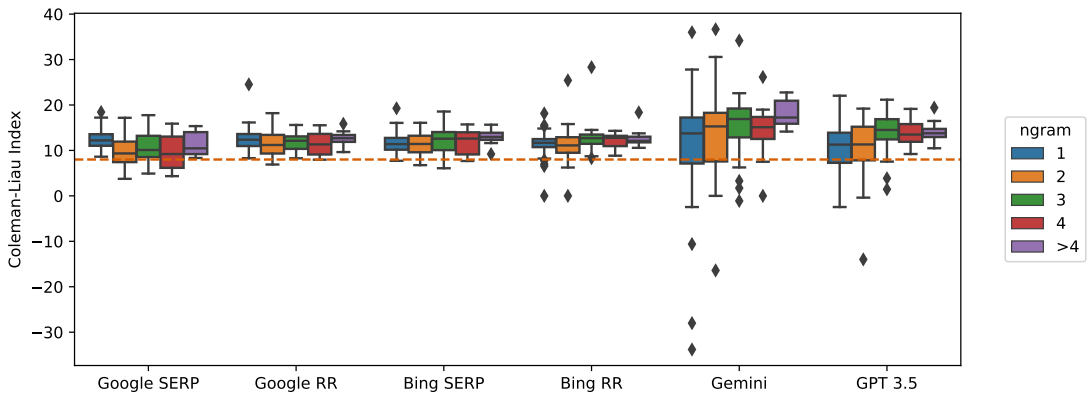


Figure 5.38: Coleman-Liau Readability Index in each response group across varying query/prompt n-gram length (An outlier has been removed from `GPT 3.5` for visualisation purposes) (Differences across varied query/prompt lengths were not significant for `Google RR` and `Bing SERP`).

Text concreteness. The average concreteness was mostly consistent irrespective of the query/prompt length for all response groups as seen in Figure 5.39 and only the distributions of `Google RR` were tested to be significantly different from each other. Similarly, the ratio of concrete words was also mostly consistent irrespective of different query/prompt lengths with the distributions being significantly different for `Google SERP`, `Google RR`, and `GPT 3.5` responses.

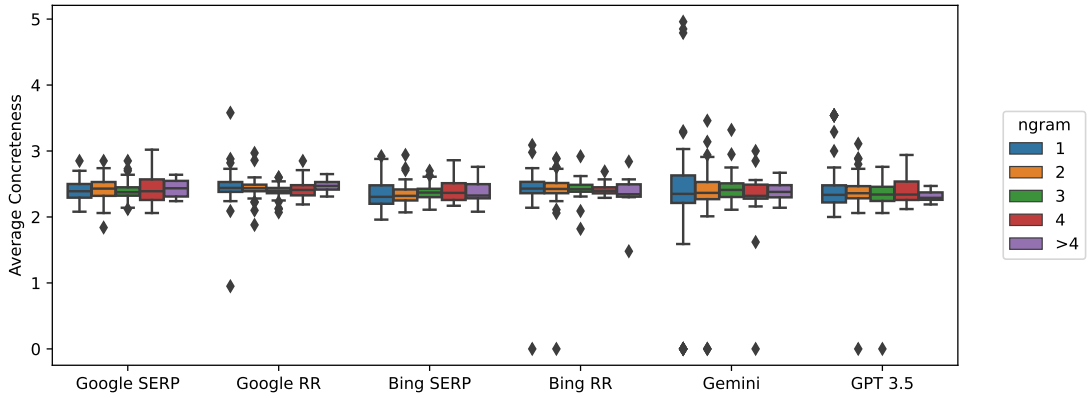


Figure 5.39: Average concreteness in each response group across varying query/prompt n-gram length (Differences across varied query/prompt lengths were significant only for `Google RR`, $p < 0.05$).

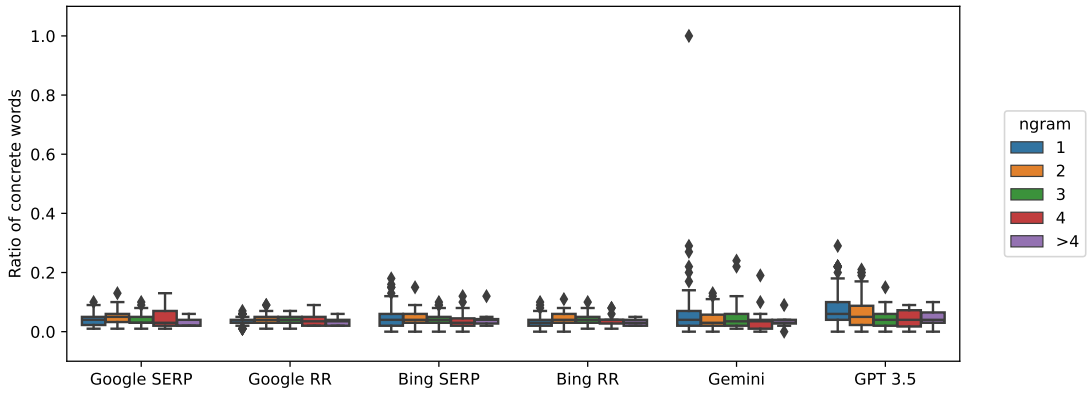


Figure 5.40: Ratio of concrete words in each response group across varying query/prompt n-gram length (Differences across varied query/prompt lengths were not significant for `Bing SERP` and `Gemini`).

All response groups had a noticeable variance in the ratio of abstract words across different n-gram lengths. The distributions were however not significantly different from each other for `Gemini` and `GPT 3.5` responses. `GPT 3.5` had the highest ratio of abstract words compared to all other distributions. For `Google SERP`, `Bing SERP`, `Google RR`, and `Bing RR`, the ratio of abstract words was the highest for trigram queries.

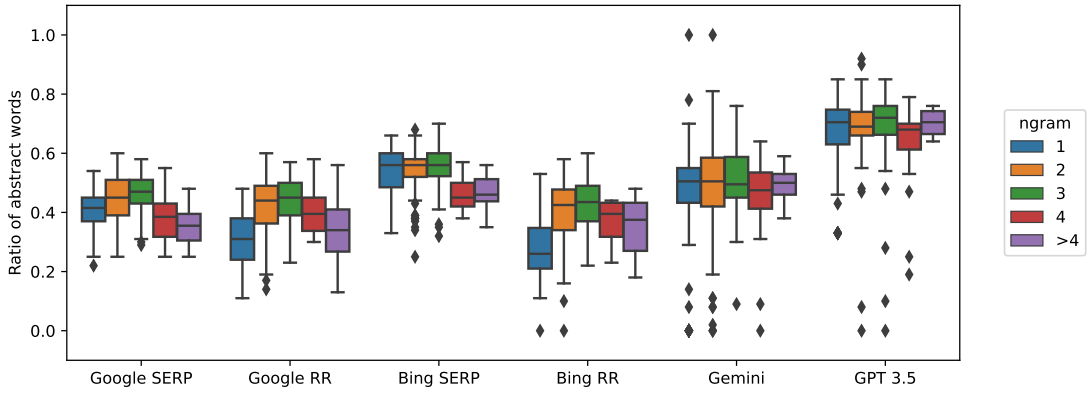


Figure 5.41: Ratio of abstract words in each response group across varying query/prompt n-gram length (Differences across varied query/prompt lengths were not significant for `Gemini` and `GPT 3.5`).

5.1.7. Investigating the effect of domain-specificity of queries on the accessibility of response groups for ASD group queries

Other than the n-gram length of the query/prompt we also categorised our queries as domain-specific queries/prompts (i.e. queries directly related to autism) and general queries/prompts. In this section, we report the effect of domain-specificity on the responses collected from all response groups (as described in Section 3.2).

Text structure. We observed a lesser number of sentences in the responses of `Google SERP` and `Gemini` for domain-specific queries/prompts, while all other response groups had more sentences in responses for domain-specific queries/prompts. However, the distributions were significantly different only for `GPT 3.5`. We also observed that the distributions had a lesser variance for responses for domain-specific queries/prompts in `Google RR`, `Bing RR`, and `Gemini` as seen in Figure 5.42.

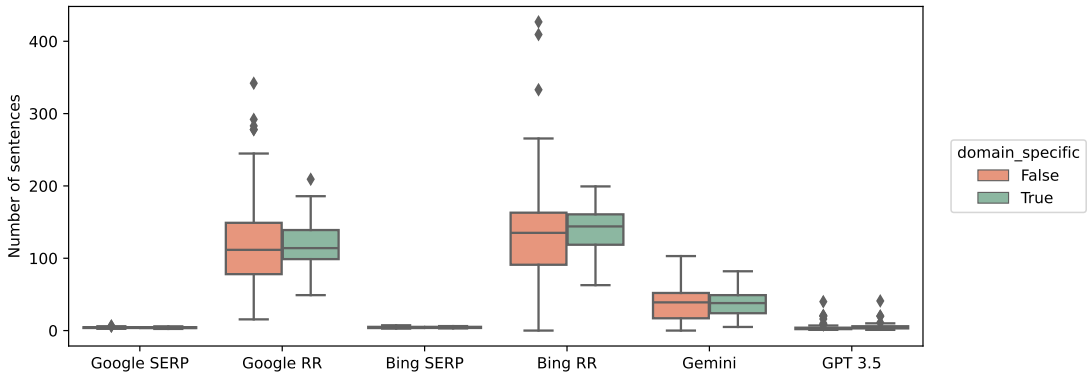


Figure 5.42: Number of sentences in each response group for domain-specific vs. general queries/prompts (Differences between domain-specific vs general queries were significant only when responses were generated by `GPT 3.5`, $p < 0.05$).

The sentences were shorter on average in responses for domain-specific queries/prompts only in `GPT 3.5` although the distributions were significantly different only for `Bing SERP`, `Bing RR`, and `Gemini` where the average sentence length was higher in responses for domain-specific queries/prompts than in responses for general queries/prompts.

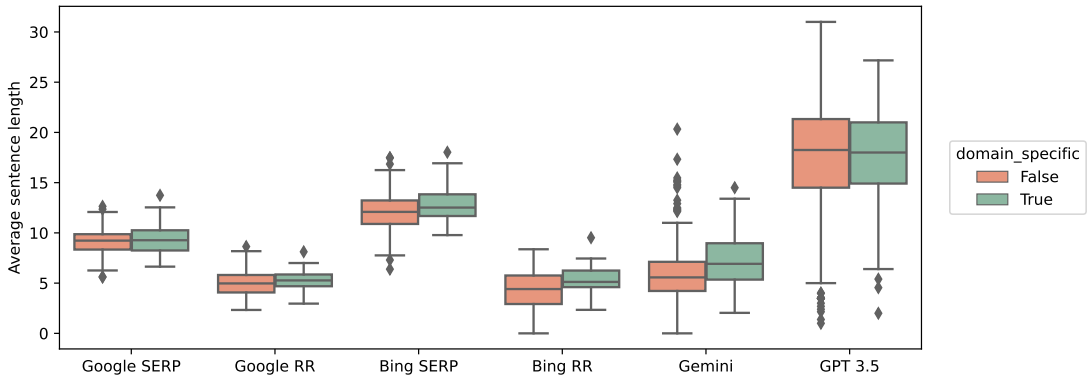


Figure 5.43: Average length in each response group for domain-specific vs. general queries/prompts (Differences between domain-specific vs general queries were not significant when responses were generated by `Google SERP`, `Google RR`, and `GPT 3.5`).

The ratio of headings was higher in responses for domain-specific queries/prompts for `Google SERP`, `Bing RR`, and `Gemini`. However, the distributions were only significantly different for `Bing RR`. In `Google RR` and `GPT 3.5` responses, the ratio of headings in the responses for domain-specific queries/prompts

was the same as the ratio in responses for general queries/prompts, although in `GPT 3.5` the ratio was 0 for nearly all responses.

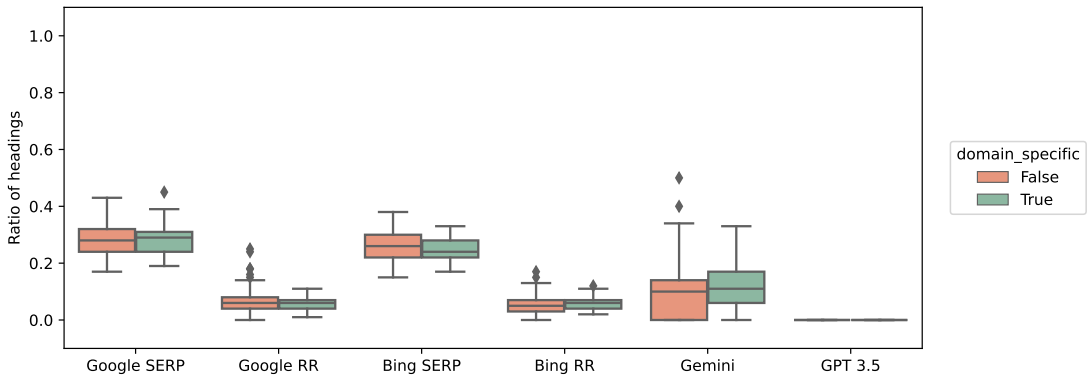


Figure 5.44: Ratio of headings in each response group for domain-specific vs. general queries/prompts (Differences between domain-specific vs general queries were significant only when responses were generated by `Bing RR`, $p < 0.05$).

The ratio of list items was significantly higher in responses for domain-specific queries/prompts for `Google RR` and `Bing RR`. In the case of the other response groups, the distribution of the ratio of list items was not significantly different for responses for domain-specific and for responses for general queries/prompts.

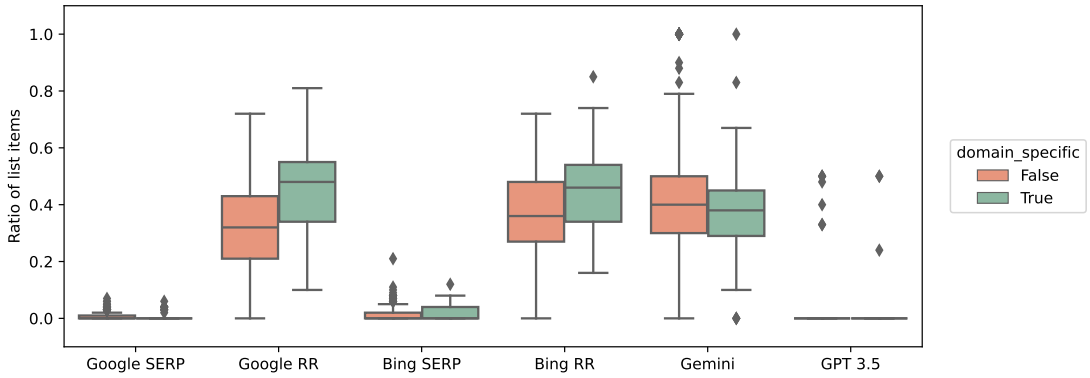


Figure 5.45: Ratio of list items in each response group for domain-specific vs. general queries/prompts (Differences between domain-specific vs general queries were significant only when responses were generated by `Google RR` and `Bing RR`, $p < 0.05$).

The ratio of paragraphs was lower in the responses for domain-specific queries/prompts in `Google SERP`, `Google RR` and `Bing RR` but the distributions were significantly different only for `Google RR` and `Bing RR`. Furthermore, the paragraphs were significantly shorter on average in responses for domain-specific queries/prompts only for `GPT 3.5`. The distributions were also significantly different for `Bing SERP` and `Bing RR` responses but the paragraphs were longer on average in responses for domain-specific queries/prompts than in responses for general queries/prompts in these response groups.

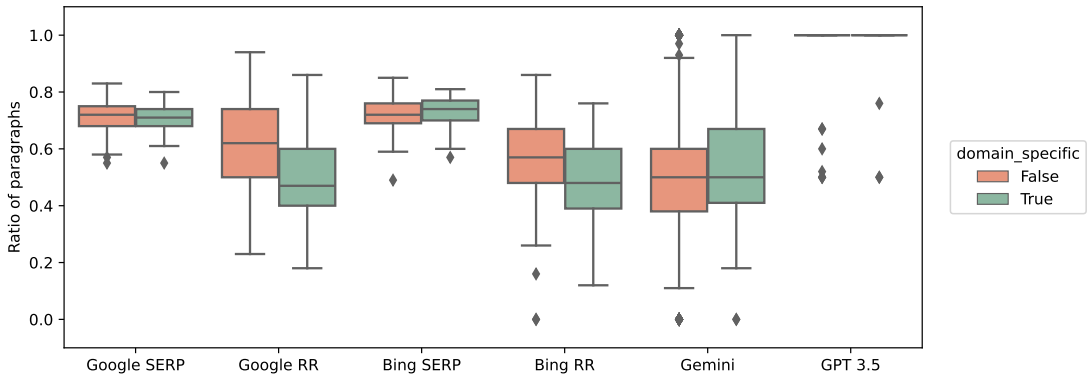


Figure 5.46: Ratio of paragraphs in each response group for domain-specific vs. general queries/prompts (Differences between domain-specific vs general queries were significant only when responses were generated by `Google RR` and `Bing RR`, $p < 0.05$).

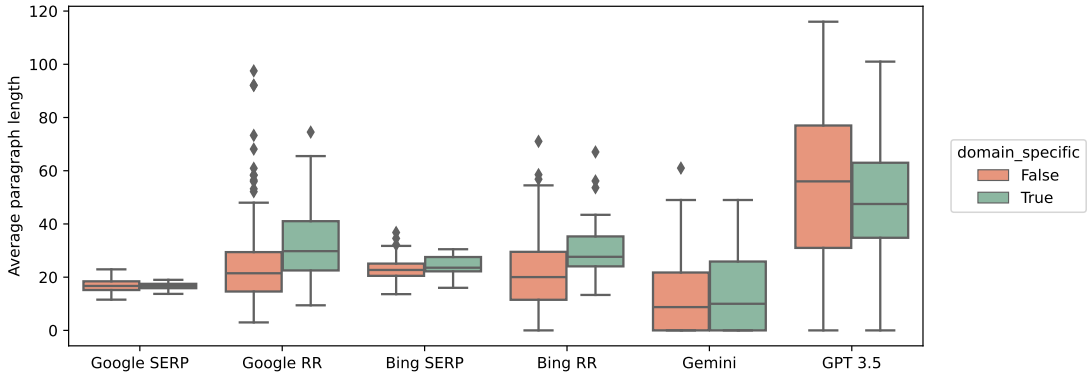


Figure 5.47: Average paragraph length in each response group for domain-specific vs. general queries/prompts (An outlier has been removed from `Bing RR` for visualisation purposes) (Differences between domain-specific vs general queries were not significant when responses were generated by `Google SERP`, `Gemini`, and `GPT 3.5`).

Text readability. The Flesch reading ease score was lower for responses for domain-specific queries/prompts for all response groups with the distributions being significantly different for all response groups except for `Google SERP`. The distributions had a lesser variance in responses for domain-specific queries/prompts for all response groups as observed in Figure 5.48. We also observed in Figure 5.48 that only `Google SERP` had many responses above the minimum threshold of 65 but the responses which met the criterion were all for general queries/prompts and none for domain-specific queries/prompts, hence only a few `Google SERP` for general queries could be considered to be readable by autistic users.

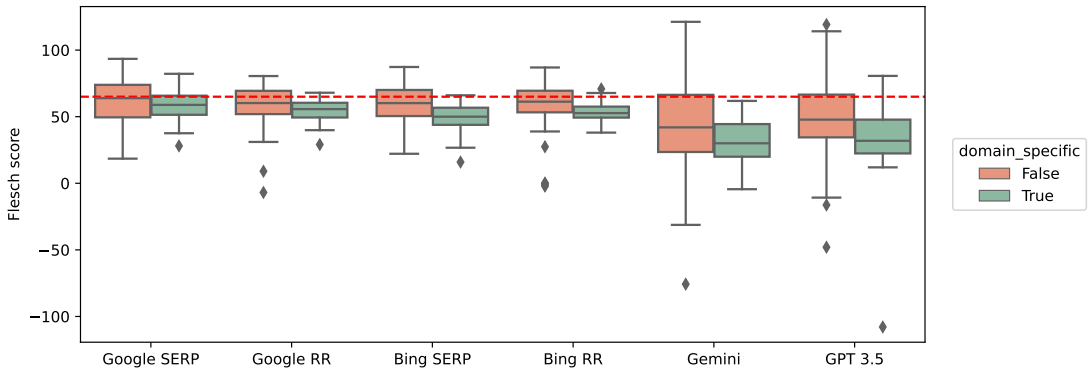


Figure 5.48: Flesch Reading Ease score in each response group for domain-specific vs. general queries/prompts (Differences between domain-specific vs general queries were not significant when responses were generated by `Google SERP`).

The Coleman-Liau readability index was significantly higher for responses to domain-specific queries/prompts for all response groups. The distributions also had a lesser variance in responses for domain-specific queries/prompts for all response groups as seen in Figure 5.49. Most of the responses also scored much higher than the maximum threshold of 8 for both general and domain-specific queries/prompts for all response groups as seen in Figure 5.49 hence regardless of the domain-specificity, responses from all response groups were not suitable for autistic users as per their Coleman-Liau readability index.

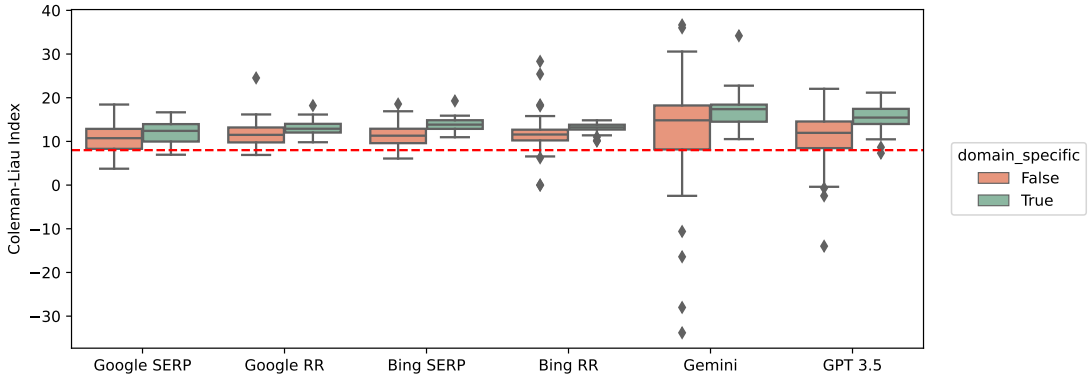


Figure 5.49: Coleman-Liau Readability Index in each response group for domain-specific vs. general queries/prompts (An outlier has been removed from `GPT 3.5` for visualisation purposes).

Text concreteness. The average concreteness was lower in responses for domain-specific queries/prompts across all response groups with the distributions significantly different for all response groups except for `Google SERP` and `Bing SERP`. The distributions also had a lesser variance in responses for domain-specific queries/prompts for all response groups as seen in Figure 5.50.

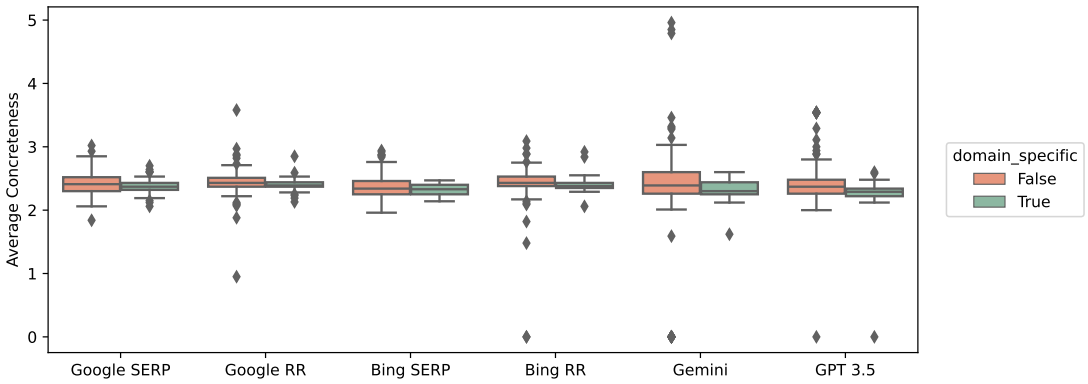


Figure 5.50: Average concreteness in each response group for domain-specific vs. general queries/prompts (Differences between domain-specific vs general queries were not significant when responses were generated by `Google SERP` and `Bing SERP`).

The ratio of concrete words was lower in responses for domain-specific queries/prompts for all response groups except for `Bing SERP` where the median ratio was the same for responses for general and domain-specific queries/prompts. The distributions were significantly different only for `Google SERP`, `Google RR`, and `GPT 3.5` responses. The distributions however had lesser variance in responses for domain-specific queries/prompts for all response groups as seen in Figure 5.51.

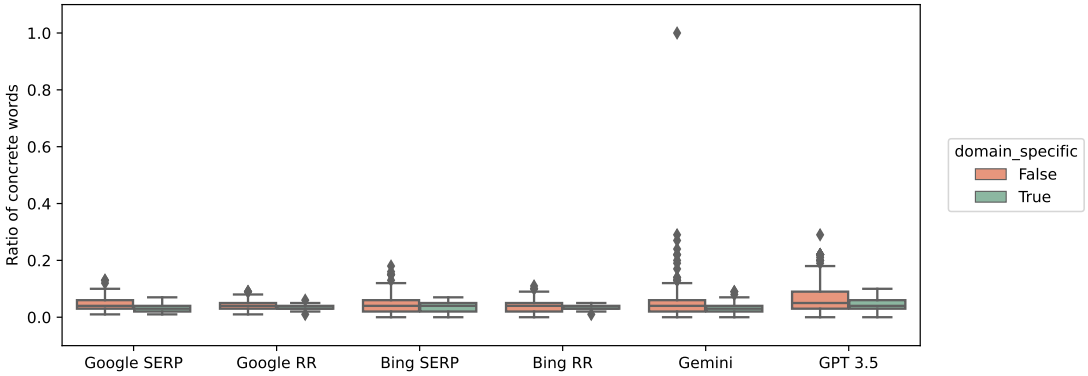


Figure 5.51: Ratio of concrete words in each response group for domain-specific vs. general queries/prompts (Differences between domain-specific vs general queries were not significant when responses were generated by `Bing SERP`, `Bing RR`, and `Gemini`).

The ratio of abstract words was higher in responses for domain-specific queries/prompts for all response groups except in the case `Bing SERP` where the median value was the same for responses domain-specific and general queries/prompts and in the case of `GPT 3.5` where the responses for domain-specific queries/prompts had a lower ratio of abstract words. The distributions however were not significantly different for any response groups but the variance was lesser in responses for domain-specific queries/prompts for all response groups except for `Google SERP` as seen in Figure 5.52.

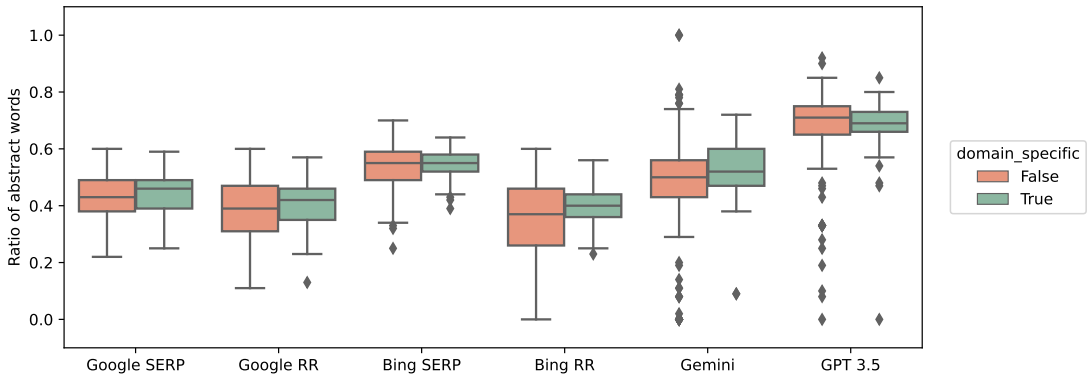


Figure 5.52: Ratio of abstract words in each response group for domain-specific vs. general queries/prompts (Differences between domain-specific vs general queries were not significant when responses were generated by all response groups).

5.2. Investigating effect of reformulation of ASD group queries and prompts on the accessibility of IST responses

As discussed previously in Section 3.1, query reformulation and prompt engineering have been studied extensively to improve the responses produced by SEs and LLM-based chatbots respectively [18, 75, 55, 90]. Hence we also investigated whether reformulating the ASD group queries and prompts to explicitly state that the user is autistic had any impact on the way the ISTs respond to the queries. In this section, we report the results of the pairwise comparisons of the change in the accessibility indicator scores of response groups under each response group type (as described in Section 3.2) when the query/prompt is reformulated. The comparisons are conducted on the same three categories of accessibility as in Section 5.1 i.e. the (1) structure, (2) readability, and (3) concreteness of the textual content of the responses collected for a given response groups.

5.2.1. Similarity measure between responses collected for original and reformed query/prompt

We first investigated whether the reformed query/prompt led to the IST responding differently than in the case of the original ASD group query/prompt. To measure the change in the responses provided by Google and Bing, we computed the RBO values of the list of the web URLs of the first 10 responses provided by the two SEs for the original and reformed query respectively.

Rank-biased overlap (RBO) is a similarity measure with a value ranging from 0 to 1. Responses collected from Google scored a slightly higher median value of 0.03 while the median value for the responses collected from Bing was 0.0. While the distribution of RBO values was not significantly different for Google and Bing, the low RBO value indicates that reformulating the query led to very different responses in both Google and Bing. It is to be noted that despite the lower median value, the RBO scores of the responses collected from Bing had a larger variance than the responses collected from Google.

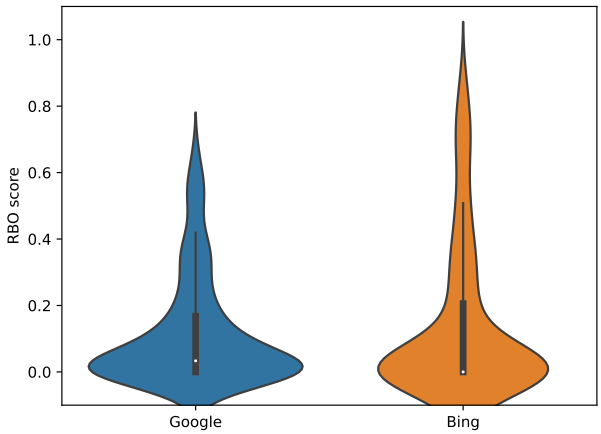
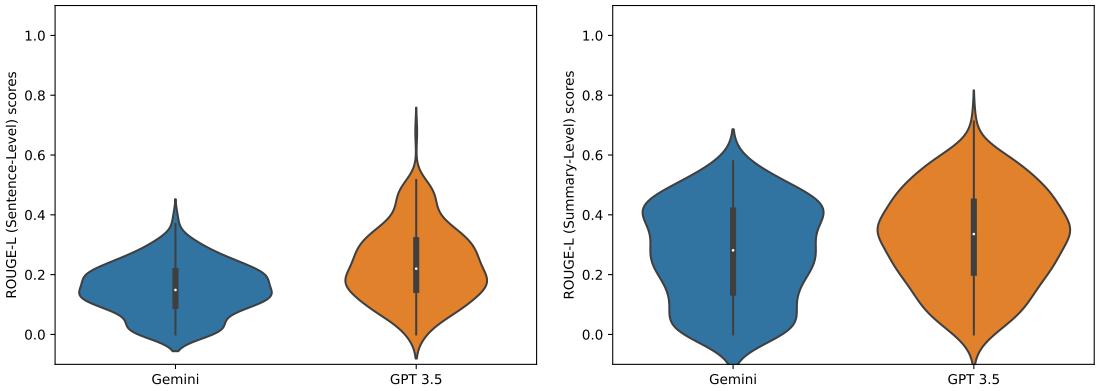


Figure 5.53: RBO scores for Google and Bing (Distributions were not significantly different).

In the case of Gemini and GPT 3.5, since we do not have a list of responses per prompt but instead a block of text, we measured the change in the generated responses using the ROUGE-L metric on both the sentence level and summary level of the responses.

Similar to RBO, ROUGE-L is also a similarity measure with a value ranging from 0 to 1. The similarity between the chatbot-generated responses for the original and reformulated prompts was found to be low on both the sentence and summary levels, with `Gemini` scoring a median value of 0.15 on the sentence level and 0.28 on the summary level; and `GPT 3.5` scoring a median value of 0.22 and 0.34 on the sentence and summary levels respectively. On both the sentence and summary levels, responses generated by Gemini and GPT 3.5 had significantly different ROUGE-L score distributions.



(a) ROUGE-L (sentence-level) scores for `Gemini` and `GPT 3.5`.

(b) ROUGE-L (summary-level) scores for `Gemini` and `GPT 3.5`.

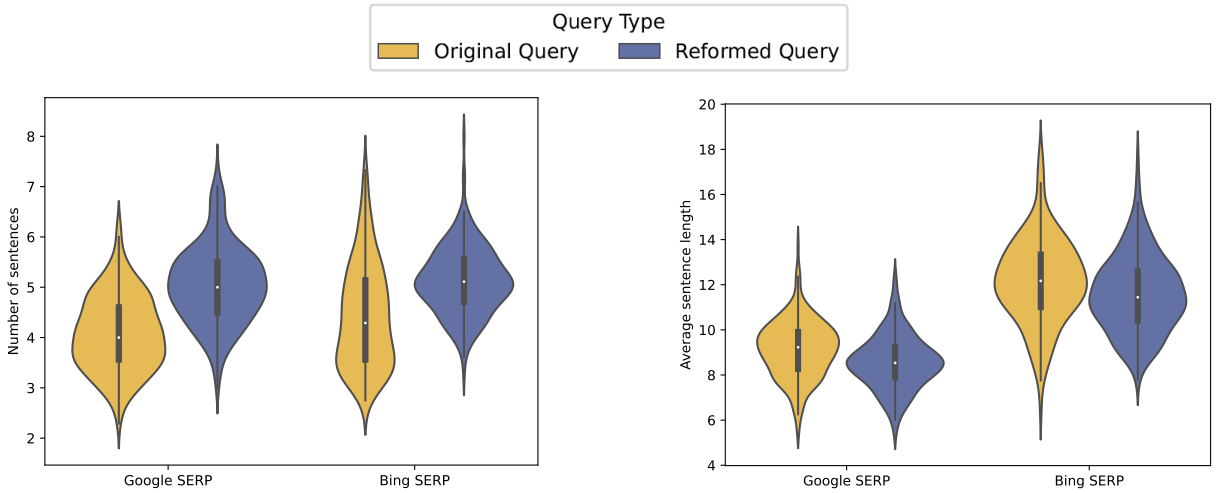
Figure 5.54: ROUGE-L scores for `Gemini` and `GPT 3.5` responses

5.2.2. Effect of query reformulation on `SERP`

With the RBO scores indicating that both Google and Bing produced very different responses for the reformulated queries, we investigated how the accessibility of `Google SERP` and `Bing SERP` responses were affected by the query reformulation.

Text structure. Both `Google SERP` and `Bing SERP` had a significant increase in the number of sentences when the queries were reformulated to explicitly indicate that the user is autistic. For `Google SERP` the median increased by 1 while for `Bing SERP` the median increased by only 0.82. From Figure 5.55a we can see that the distribution of the number of sentences for both `Google SERP` and `Bing SERP` had lesser variance after the query was reformulated. Despite the increase in the number of sentences, the average

length of the sentences dropped significantly for both `Google SERP` (median change: -0.7) and `Bing SERP` (median change: -0.72) as seen in Figure 5.55b.

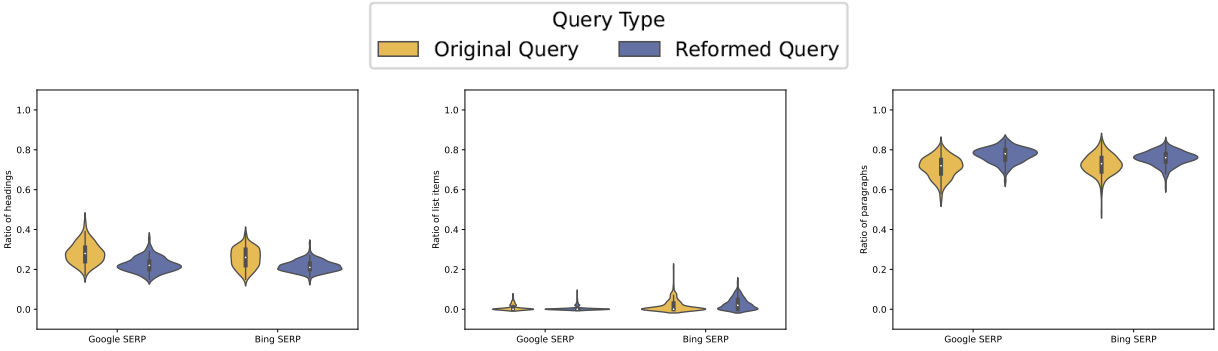


(a) Number of sentences

(b) Average sentence length

Figure 5.55: Analysis of text structure (sentence-level) based on responses collected from `Google SERP` and `Bing SERP` for original query and reformulated query.

The ratio of headings decreased significantly for both `Google SERP` (median change: -0.06) and `Bing SERP` (median change: -0.05) after the query was reformulated but the ratio of list items significantly increased for `Bing SERP` by 0.02 while remaining unchanged for `Google SERP`. On the other hand, the ratio of paragraphs increased significantly for both `Google SERP` (median change: 0.06) and `Bing SERP` (median change: 0.03) after the query was reformulated. From Figures 5.56a and 5.56c we also observed that the distributions for all three ratios also had a lesser variance for both `Google SERP` and `Bing SERP` after the query was reformulated.



(a) Ratio of headings

(b) Ratio of list items

(c) Ratio of paragraphs

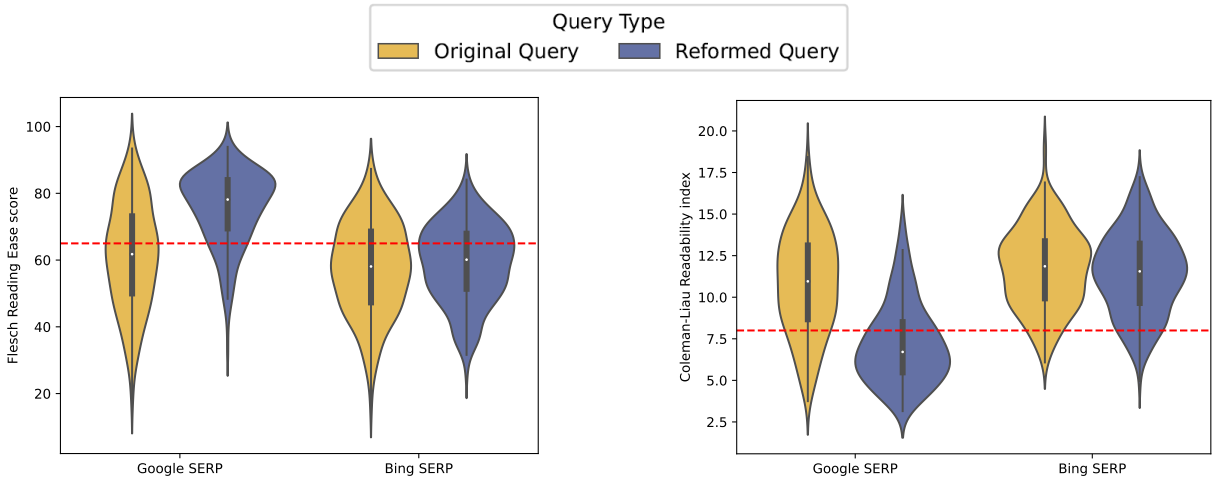
Figure 5.56: Analysis of text structure (full body) based on responses collected from `Google SERP` and `Bing SERP` for original query and reformulated query.

The average length of paragraphs increased significantly for `Google SERP` (median change: 3.07) and for `Bing SERP` (median change: 1.84) after the query was reformulated. Although the increase was by a much larger margin for `Google SERP` than for `Bing SERP` responses. From Figure 5.57 we can also see that the distribution had a lesser variance for `Google SERP` after the query was reformulated.

Figure 5.57: Average paragraph length in `Google SERP` and `Bing SERP` for original query and reformulated query.

Text readability. `Google SERP` had a huge (and significant) increase in the Flesch reading ease scores (median change: 16.37) after the query was reformulated with most of the responses scoring higher than the minimum threshold of 65 as seen from Figure 5.58a thus making them readable for autistic users. `Bing SERP` also observed an increase (2.04) although the increase was found to be not significant and most of the responses remained below the minimum threshold of 65 as seen in Figure 5.58a.

Similarly, we observed a significant and substantial drop in the Coleman-Liau readability index for `Google SERP` responses (median change: -4.24) after the query was reformulated with a lot of the responses scoring below the maximum threshold of 8 as seen in Figure 5.58b which indicated that a lot of `Google SERP` responses became readable for autistic users after the query was reformulated. `Bing SERP` also observed a decrease (median change: -0.3) but the decrease once again was tested to be not significant and most of the responses scored an index value much higher than the maximum threshold of 8 as seen in Figure 5.58b thus `Bing SERP` responses remained unreadable by autistic users even after the query was reformulated.



(a) Flesch reading ease score

(b) Coleman-Liau Readability index

Figure 5.58: Analysis of text readability based on responses collected from `Google SERP` and `Bing SERP` for original query and reformulated query (Distributions for both text readability indicators were not significantly different for `Bing SERP`).

Text concreteness. Query reformulation resulted in a significant increase in average concreteness for both `Google SERP` (median change: 0.14) and `Bing SERP` (median change: 0.08). The distribution of average concreteness of `Bing SERP` responses also had lesser variance after query reformulation as seen in Figure 5.59.

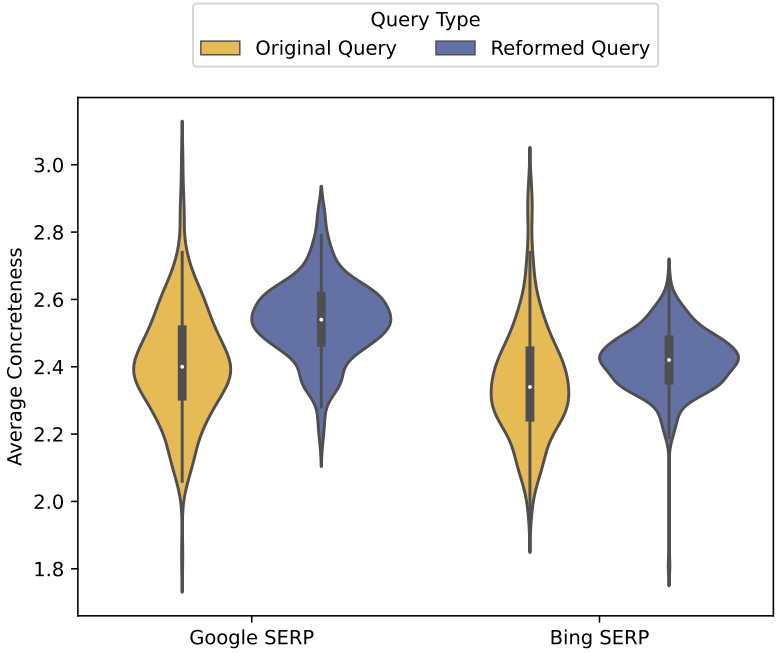


Figure 5.59: Average concreteness of `Google SERP` and `Bing SERP` for original query and reformulated query.

The ratio of concrete words remained unchanged for both `Google SERP` and `Bing SERP`, although the distributions were found to be significantly different for `Bing SERP`. The distribution of the ratio value also had a lesser variance for `Bing SERP` after query reformulation as can be seen from Figure 5.60a. For `Google SERP`, the variance remained relatively unchanged.

In the case of the ratio of abstract words, only Google SERP observed a significant increase after the query was reformulated (median change: 0.06). The ratio's distribution had a lesser variance for both Google SERP and Bing SERP after query reformulation as seen in Figure 5.60b.

(a) Ratio of concrete words (Distributions were not significantly different for Google SERP)

(b) Ratio of abstract words (Distributions were not significantly different for Bing SERP)

Figure 5.60: Analysis of text concreteness based on responses collected from Google SERP and Bing SERP for original query and reformulated query.

5.2.3. Effect of query reformulation on RR

Since the RBO scores were based on the URLs of the first 10 responses of each search engine, the low RBO scores indicate that both Google and Bing ranked very different resources in their top 10 responses for the original ASD group queries and their reformulated counterparts respectively. Since we already observed a significant change in the accessibility indicators for the SERP responses, we inspected the change in the accessibility indicators for the textual content in the RR responses as well.

Text structure. The number of sentences dropped by a huge margin for Google RR (median change: -18.52) while the number increased by a small margin for Bing RR (median change: 3.21) after the query was reformulated although the difference was significant for Google SERP but not for Bing RR. The distribution of the number of sentences also had a lesser variance for both Google RR and Bing RR as seen in Figure 5.61a.

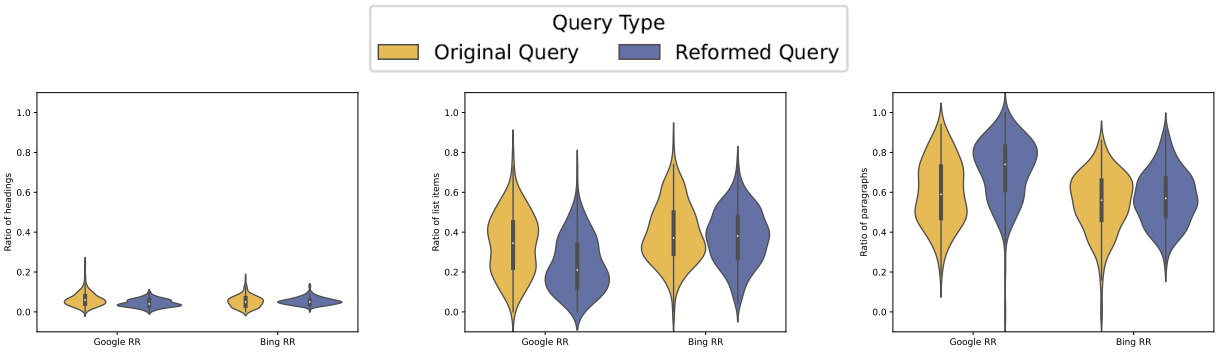
Both Google RR and Bing RR observed a slight but significant increase in the average length of sentences after query reformulation. However, the increase was more in Bing RR (median change: 1.3) than in Google RR (median change: 1.08). The distributions for the indicator also had a lesser variance for both Google RR and Bing RR responses after the query was reformulated as seen in Figure 5.61b.

(a) Number of sentences (Distributions were not significantly different for `Bing RR`)

(b) Average sentence length

Figure 5.61: Analysis of text structure (sentence-level) based on responses collected from `Google RR` and `Bing RR` for original query and reformulated query.

`Google RR` observed a significant increase in the ratio of paragraphs (median change: 0.15) but a significant decrease in the ratio of headings (median change: -0.02) as well as in the ratio of list items (median change: -0.13) after the query was reformulated. `Bing RR` also observed an increase in the ratio of paragraphs (median change: 0.01) and a decrease in the ratio of list items (median change: -0.01) but the ratio of headings remained unchanged. The differences in all three ratios were by a much smaller margin for `Bing RR` compared to `Google RR`, and all the differences were also found to be not significant.



(a) Ratio of headings

(b) Ratio of list items

(c) Ratio of paragraphs

Figure 5.62: Analysis of text structure (full body) based on responses collected from `Google RR` and `Bing RR` for original queries and reformulated queries (Distributions for none of the three ratios were not significantly different for `Bing RR` responses).

Both `Google RR` and `Bing RR` observed a significant increase in the average paragraph length but `Bing RR` had a much larger increase (median change: 9.46) compared to `Google RR` (median change: 6.34).

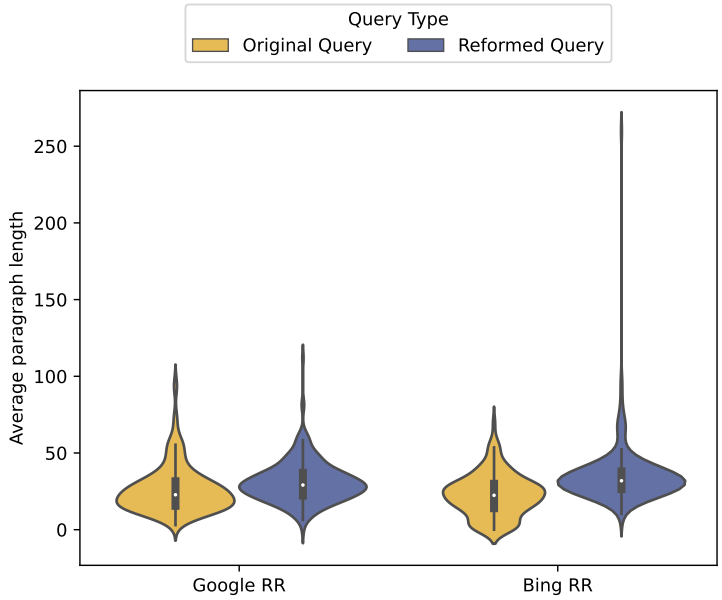


Figure 5.63: Average paragraph length in `Google RR` and `Bing RR` for original queries and reformulated queries (An outlier has been removed from `Google RR` and `Bing RR` for visualisation purposes).

Text readability. We observed a significant increase in the Flesch reading ease score for `Google RR` (median change: 12.46) after the query was reformulated with many of the responses scoring above the minimum threshold of 65 as seen in Figure 5.64a indicating that the query reformulation led to many `Google RR` responses to become readable for autistic users. `Bing RR` responses on the other hand fell in their scores (median change: -1.02) although the difference was not significant.

`Google RR` also showed an improvement in its Coleman-Liau readability index values after the query was reformulated with a significant fall in the index value by 2.57. However, despite the drop, most `Google RR` responses did not score below the maximum threshold of 8 and hence remained unreadable by autistic users even after the query was reformulated. `Bing RR`'s index value also decreased although only by 0.08 and the decrease was also not significant. However, most responses for both `Google RR` and `Bing RR` remained above the maximum threshold index value of 8 as seen in Figure 5.64b.

(a) Flesch reading ease score

(b) Coleman-Liau Readability index

Figure 5.64: Analysis of text readability based on responses collected from `Google RR` and `Bing RR` for original queries and reformulated queries (Distributions for both text readability indicators were not significantly different for `Bing RR`).

Text concreteness. The average concreteness dropped significantly for `Google RR` (median change: -0.01) but increased in `Bing RR` (median change: 0.01) although the difference in `Bing RR` was not significant. The ratio of concrete words increased significantly for `Google RR` (median change: 0.01) but remained unchanged for `Bing RR`. In the case of the ratio of abstract words, the ratio increased significantly for both `Google RR` (median change: 0.12) and `Bing RR` (median change: 0.08).

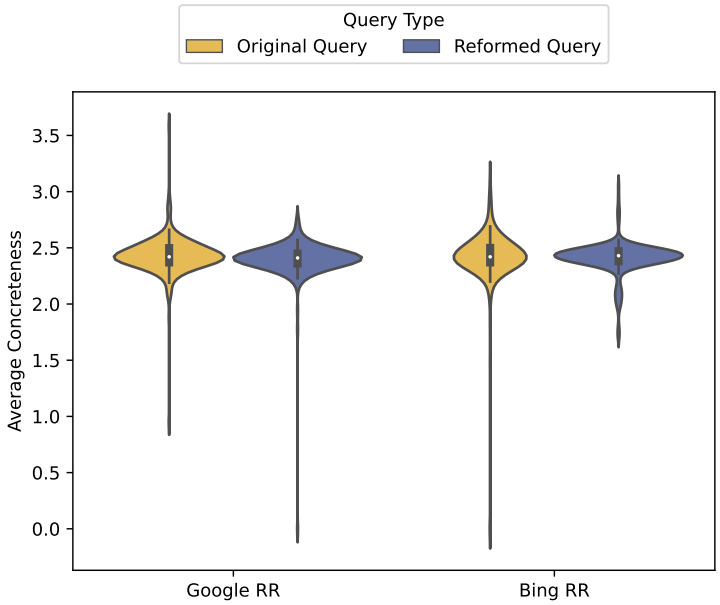
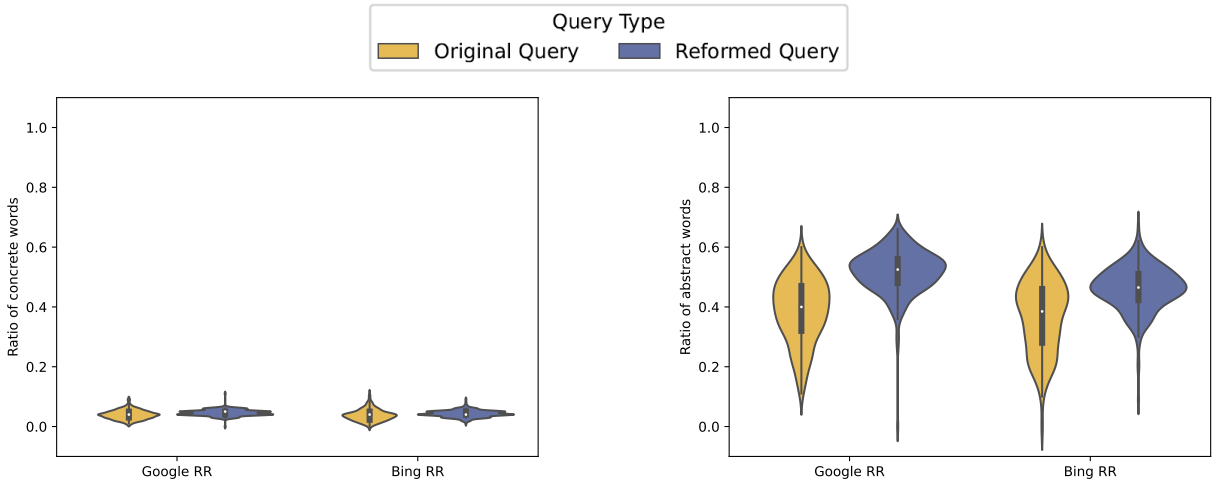


Figure 5.65: Average concreteness of `Google RR` and `Bing RR` for original queries and reformulated queries (Distributions were not significantly different for `Bing RR`).



(a) Ratio of concrete words

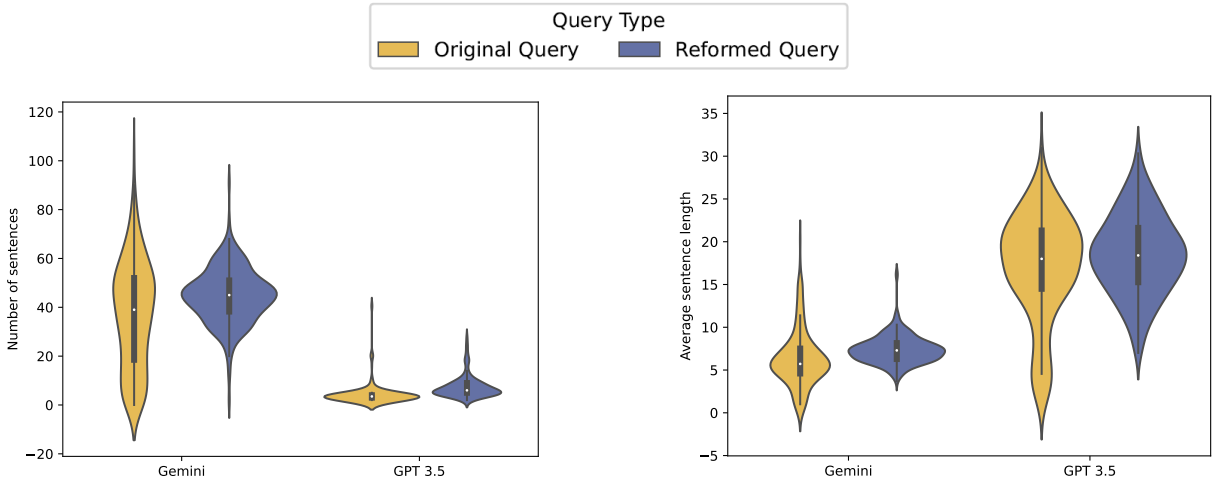
(b) Ratio of abstract words

Figure 5.66: Analysis of text concreteness based on responses collected from `Google RR` and `Bing RR` for original queries and reformulated queries.

5.2.4. Effect of prompt reformulation on `Chatbot`

Both the chatbots generated different responses for the original prompt and reformulated prompt as inferred from the low ROUGE-L scores. Thus we investigated the effect of prompt reformulation on the accessibility of the responses generated by the chatbots.

Text structure. The number of sentences increased significantly for both `Gemini` (median change: 6) and `GPT 3.5` (median change: 2.5) when the prompt was reformulated. The distribution also had a lesser variance for `Gemini` responses after the prompt was reformulated as can be seen from Figure 5.67a. The average length of the sentences also increased for both `Gemini` (median change: 1.58) and `GPT 3.5` (median change: 0.4) although the difference was significant only for `Gemini` and not for `GPT 3.5`.



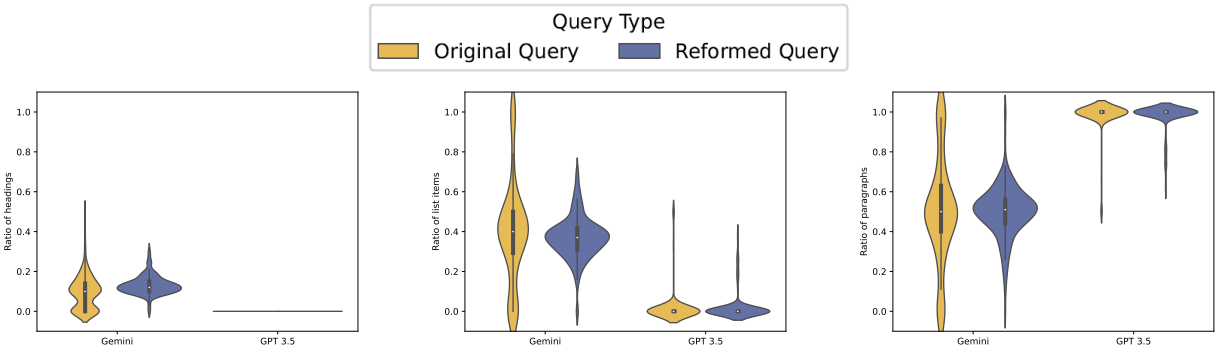
(a) Number of sentences

(b) Average sentence length (Distributions were not significantly different for `GPT 3.5`)

Figure 5.67: Analysis of text structure (sentence-level) based on responses collected from `Gemini` and `GPT 3.5` for original prompts and reformulated prompts.

The ratio of headings and the ratio of paragraphs increased for the responses generated by `Gemini` (median change: 0.02 and 0.01 respectively) while the ratio of list items decreased (median change: -0.03).

However only the increase in the ratio of headings and the ratio of paragraphs was significant, the decrease in the ratio of list items was not significant. All three distributions were observed to have a lesser variance after the prompt was reformulated in Figures 5.68a - 5.68c. On the other hand, we did not observe any change in any of the three ratios for `GPT 3.5` responses.



(a) Ratio of headings **(b)** Ratio of list items **(c)** Ratio of paragraphs

Figure 5.68: Analysis of text structure (full body) based on responses collected from `Gemini` and `GPT 3.5` for original prompts and reformulated prompts (Distributions for three ratios were not significantly different for `GPT 3.5`).

The average paragraph length increased significantly for `Gemini` (median change: 13.8) but decreased significantly for `GPT 3.5` (median change: -0.16) after prompt reformulation. The distribution also had a lesser variance for `GPT 3.5` after the prompt was reformulated as seen in Figure 5.69.

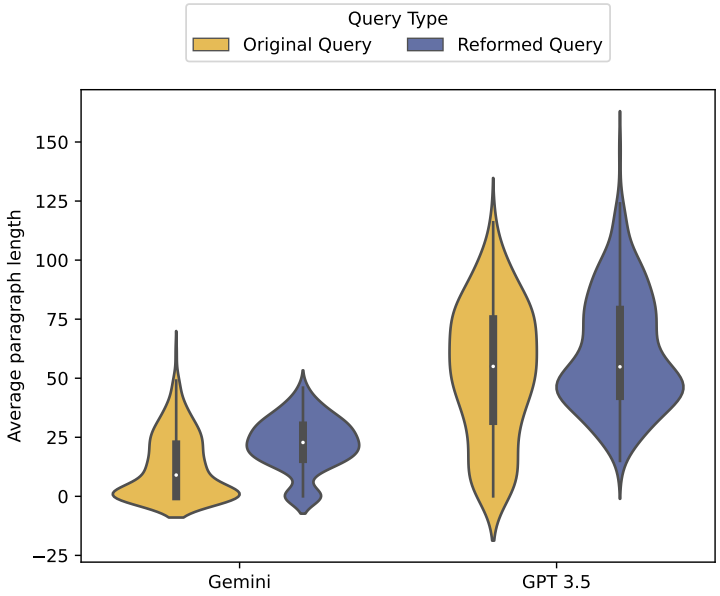
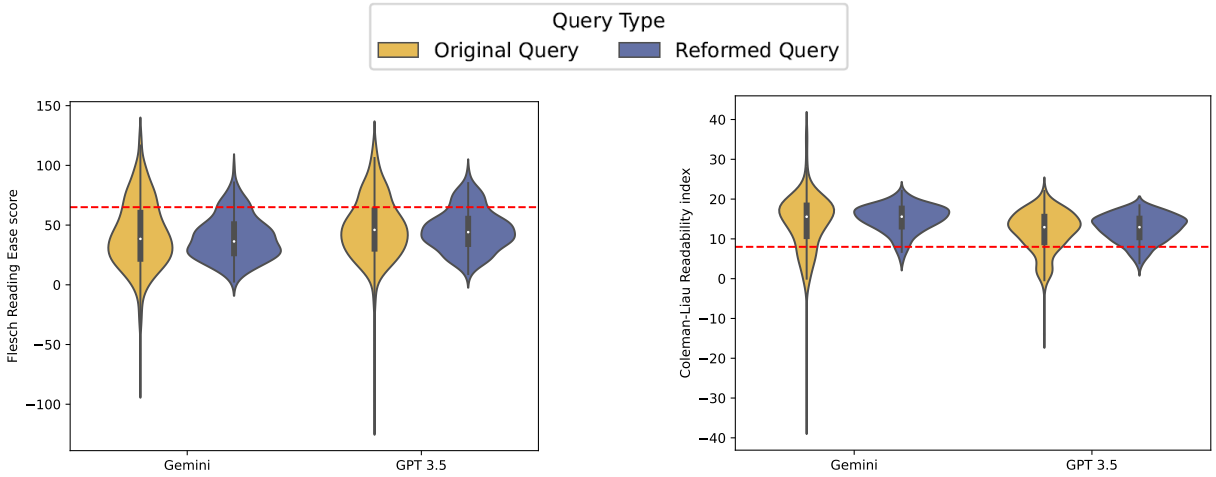


Figure 5.69: Average paragraph length in `Gemini` and `GPT 3.5` responses for original prompts and reformulated prompts.

Text readability. Both `Gemini` and `GPT 3.5` observed a decrease in Flesch Reading Ease scores (median change: -2.15 and -1.88 respectively) when the prompt was reformulated, although the decrease was not significant for either `Gemini` or `GPT 3.5`. The distributions also had a lesser variance for both `Gemini` and `GPT 3.5` as seen in Figure 5.70a.

There was an increase in the Coleman-Liau readability index value for `Gemini` (median change: 0.01) but a decrease for `GPT 3.5` (median change: -0.03). Despite the decrease in the index value in `GPT 3.5`,

most of the responses still scored above the maximum threshold of 8 as seen in Figure 5.70b and thus remained unreadable even after prompt reformulation. Furthermore, the change in the readability index for both the response groups was also not significant. The distributions however had a lesser variance for both `Gemini` and `GPT 3.5` as seen in Figure 5.70b.



(a) Flesch reading ease score

(b) Coleman-Liau Readability index (An outlier has been removed from `Gemini` and `GPT 3.5` response for visualisation purposes)

Figure 5.70: Analysis of text readability based on responses collected from `Gemini` and `GPT 3.5` for original prompts and reformulated prompts (Distributions were not significantly different for `Gemini` and `GPT 3.5` for both text readability indicators).

Text concreteness. The average concreteness decreased significantly for `Gemini` (median change: -0.04) but increased for `GPT 3.5` (median change: 0.01) when the prompt was reformulated. However, the difference was found to be not significant for `GPT 3.5`. The distribution also had a lesser variance for both `Gemini` and `GPT 3.5` respectively after the prompt was reformulated as seen in Figure 5.71.

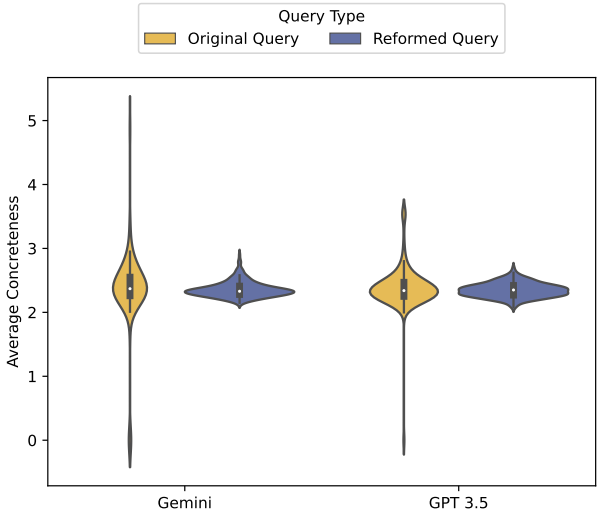


Figure 5.71: Average concreteness of `Gemini` and `GPT 3.5` responses for original prompts and reformulated prompts (Distributions were not significantly different for `GPT 3.5`).

There was no change observed in the ratio of concrete words for both `Gemini` or `GPT 3.5` but the ratio of abstract words increased for both `Gemini` (median change: 0.04) and `GPT 3.5` (median change: 0.01).

However, the increase was significant only for `Gemini` and not for `GPT 3.5`. The distributions had a lesser variance for both `Gemini` and `GPT 3.5` after the prompt was reformulated as seen in Figures 5.30a and 5.30b.

(a) Ratio of concrete words (Distributions were not significantly different for `Gemini` and `GPT 3.5`)

(b) Ratio of abstract words (Distributions were not significantly different for `GPT 3.5`)

Figure 5.72: Analysis of text concreteness based on responses collected from `Gemini` and `GPT 3.5` for original prompts and reformulated prompts.

6

Discussion

In this chapter, we answer our research questions based on the results reported in Chapter 5. The outcomes of our experiments revealed interesting trends in the structure, readability, and concreteness of the textual content in the responses generated by Google, Bing, Gemini, and ChatGPT. Each IST had its strengths and limitations when accommodating the needs of autistic users which we also elaborate upon here.

6.1. Comparing accessibility of IST responses for autistic users

Our first question was whether SEs and LLM-based chatbots produced responses whose textual content catered to the needs of autistic users. To answer this question we conducted several empirical investigation (described in Section 3.4) to compare the IST responses on three main aspects of accessibility of textual content for autistic users - (1) structure, (2) readability, and (3) concreteness of the text.

Our first comparison was a pairwise comparison of responses generated by three response group types - `SERP`. `RR`, and `Chatbot` for queries typically asked by the general public vs. the queries more characteristic to autistic users. We then investigated how the accessibility of each of the aforementioned response group types differed from one another. To further analyse the differences we performed a pairwise comparison of the accessibility indicators between the response groups under each response group type. We also investigated the effect of the n-gram length and domain-specificity of the query/prompt on the accessibility indicator scores of each response group individually.

The results of all these experiments were reported in Section 5.1. In this section, we elaborate further on the inferences and implications of the observed results. At the end of this section, we compile our overall inferences and suggest a few ways IST responses could be made more accessible for autistic users.

Comparison between responses generated for Control and ASD group queries. In our experiments comparing the accessibility indicator scores of the response group types for control and ASD group queries, we found that in all response group types responses generated for control group queries were significantly easier to read than the responses generated for ASD group queries. The `SERP` and `RR` responses were also significantly shorter for control group queries than for ASD group queries. We observed an opposite trend in `Chatbot` responses with longer responses for control group queries but the difference was not statistically significant.

All response group types also generated more concrete responses for control group queries than for ASD group queries with the responses for control group queries scoring a significantly higher average concreteness and ratio of concrete words and a significantly lower ratio of abstract words. However, the total ratio of concrete or abstract words was significantly lower for responses generated for control group queries than the responses generated for ASD group queries. We computed the text concreteness indicators based on the dataset compiled by Brysbaert et al [13] which consists of over 40000 commonly known English lemmas. Given that the responses generated for ASD group queries had more words that had a concreteness rating, we can infer that although the responses generated for control group queries by all three response group types were more concrete on average, the responses generated by them for the ASD group queries consisted of more common English words.

The lower reading ease and lower concreteness of the responses generated for ASD group queries indicate that the ISTs are generating more accessible responses for queries asked by the general public rather than for queries characteristic to autistic users. Autistic users already struggle with searching for information on the web and from our results it appears that due to the nature of their queries, they are being exposed to responses that are less suitable for them from the get-go which could be the reason why autistic people state that they do not prefer searching for information online in the first place as reported in the online survey conducted by Gillespie et al. [31]. Despite this unexpected and unfortunate outcome, the fact that ISTs are responding differently to ASD group queries, we can infer that the queries posed by autistic users are unique enough for them to be flagged by the IST's algorithm. The algorithms used by SEs and LLM-based chatbots to generate their responses only need to be improved in terms of generating responses that use simpler and more concrete language which would make the responses better suited to autistic users, especially for queries that are generally asked by autistic users.

To understand how the algorithm of each IST could be improved to present information to autistic users in an accessible manner, we need to uncover how each IST is falling short in the accessibility of their responses when generating them for autistic users' queries. Therefore we compared and contrasted the accessibility indicator scores of the IST responses specifically for ASD group queries to highlight the strengths and limitations of the responses produced by each IST in the context of accessibility to autistic users.

General comparison of IST responses for ASD group queries. Comparing the `SERP`, `RR`, and `Chatbot` responses for ASD group queries revealed that `SERP` had much lesser textual content than `RR` and `Chatbot` responses. This is to be expected, as the purpose of the SERP is to give the user only a short gist of the actual content within each ranked resource. `RR` was observed to have the most textual content as expected however the text was structured in a manner that aligns with the needs of autistic users. `Google SERP`, in particular, had significantly lesser textual content than the responses collected from `Bing SERP`, and the content was more often structured as lists and short paragraphs which made the responses more suitable for autistic users. Hence, the structure of `Google SERP` responses is better suited for autistic users, however, we cannot make a similar conclusive distinction between the structure of `Google RR` responses and `Bing RR` responses as all the accessibility indicators for the structure of the responses had similar distributions for `Google RR` and `Bing RR`.

In the context of text readability, as per our results, most of the `SERP` and `RR` responses were suitable for 10^{th} graders and above while autistic users prefer texts that would be readable by 8^{th} graders or lower. Upon comparing the Google and Bing responses, we found that more `Google SERP` responses met the readability criteria to be suitable for autistic users than `Bing SERP` responses. We however did not observe a similar distinction in readability scores between `Google RR` and `Bing RR` responses.

Putting together the results from the pairwise comparisons for the text structure and text readability indicator scores for response groups under `SERP` and `RR`, it becomes apparent that Google and Bing retrieve responses with similar structure and readability, but Google SERP provides snippets that are more concise and easier to read than Bing SERP.

Upon comparing the text structure and readability indicator scores for `Gemini` and `GPT 3.5`, we found that the indicator scores weren't consistently better for one response group over the other as we observed in the case of `Google SERP` and `Bing SERP`. Therefore it is not so simple to make general remarks on the response structure and readability of `Chatbot` responses. This is because, in our pairwise comparisons, we observed that while both `Gemini` and `GPT 3.5` had a higher variance in their indicator score distributions compared to the response groups under `SERP` and `RR`, `Gemini` and `GPT 3.5` varied in different ways. While `Gemini` responses were inconsistent in the way the content was divided into headings, list items, and paragraphs; `GPT 3.5` responses varied in the average length of the sentences and paragraphs. `GPT 3.5` responses were also only slightly more readable than `Gemini` responses, but most responses in both the response groups were only suitable for college students or higher which is even less suitable for autistic users than the `SERP` and `RR` responses.

The pairwise comparisons under each response group type revealed significant differences in the scores for text structure and readability indicators, however, in the case of text concreteness, all response groups were similarly concrete on average. The distributions of average concreteness for each response group under each response group type were not even significantly different. Thus practically all of them

had similar concreteness around 2.4 which would be categorised as abstract as per our categorisation rule. This implies that all ISTs generated mostly abstract responses for queries typically asked by which is not suitable for autistic users.

Through a deeper analysis of the indicators for text concreteness, we found that the scale of the ratio of concrete words was much smaller than the scale of the ratio of abstract words for all response groups. This indicates that the abstract words had a much larger contribution to the average concreteness of the responses than the concrete words. This could be due to our strict categorisation rule that a word would be considered concrete only if it has a concreteness rating $>$4. Another interesting highlight was that the ratio of abstract words was observed to be much higher for `Chatbot` responses compared to `SERP` and `RR` responses. Given that the ratio of concrete words in `Chatbot` responses was observed to be the same as that of `SERP` and `RR` responses, but the higher ratio of abstract words in `Chatbot` responses implies that the `Chatbot` consisted of more words with a concreteness rating. As discussed previously, since the concreteness indicators were computed using a dataset consisting of over 40000 commonly known English lemmas [13], we can infer that the `Chatbot` responses consisted of more commonly known words than SE responses. Accessibility guidelines suggest using fewer jargon and metaphors and using vocabulary that an autistic user is familiar with [12, 1], given that the `Chatbot` responses had more common words than SE responses, the responses although less readable may still be more suitable for autistic users.

Despite the distinct trade-offs between SE and Chatbot responses, one thing common for `SERP`, `RR`, and `Chatbot` response group types, as well as for the response groups under each response group type, was the large variances in the indicator scores. While the significant and often substantial differences between the response groups under each response group type could explain the variance observed in the distributions for each response group type during general comparisons (as reported in Section 5.1.2), we observed large variances in the indicator distributions for each response group as well, especially in `Gemini` and `GPT 3.5`.

Control and ASD group queries were already found to lead ISTs into responding differently as reported in Section 5.1.1. Hence we investigated whether different kinds of ASD group queries lead to different responses from the ISTs which could explain the variances in the distributions of each response group's accessibility indicator scores. Specifically, we focused on the potential effect of the length and domain-specificity of the queries and prompts generally asked by autistic users on the accessibility indicator scores of the response groups.

Comparison of response groups across different n-gram lengths of ASD group queries/prompts. Upon analysing the results of our investigation on the effect of n-gram length of ASD group queries and prompts on the accessibility indicator scores of the response groups, we found that response groups under `SERP` and `RR` were much more sensitive to the changing query length than the response groups under `Chatbot`. Furthermore, we noticed that the text readability indicator scores of `Google SERP` and `Google RR` were impacted differently than `Bing SERP` and `Bing RR` implying that for the SEs, the impact of query length is unique to the IST itself, at least in terms of the readability of the responses generated for ASD group queries.

In the case of `Google SERP` and `Google RR`, while the scores of each accessibility indicator fluctuated differently for different query lengths, we observed that unigram queries in general led to responses whose text structure was suitable for autistic users. The sentences and paragraphs were shorter on average for responses generated by `Google SERP` and `Google RR` for unigram queries compared to the responses generated by them for other query lengths. The responses generated by them for unigram queries also had a higher ratio of headings and list items and a lower ratio of paragraphs compared to the responses for other query lengths. However, the responses generated for unigram queries were harder to read (i.e. lower Flesch reading score and higher Coleman-Liau readability index) than the responses generated for other query lengths.

The average concreteness of the responses generated by `Google SERP` and `Google RR` remained consistent across the different query lengths as did the ratio of concrete words, however too short (unigram queries) or too long queries ($\geq$ 4-gram queries) resulted in responses with a lower ratio of abstract words in `Google SERP` and `Google RR`. The consistent ratio of concrete words but fall in the ratio of abstract words imply that very short or very long queries led to responses that consisted of less number of common

English words which would make them less suitable for autistic users as discussed previously in other sections of this chapter.

Bing SERP and Bing RR responses observed fluctuations in their accessibility indicator scores in a similar fashion to Google SERP and Google RR responses except for the fluctuations in the text readability indicators. The unigram queries led to responses that were structured to be more suitable for autistic users compared to responses generated by them for other query lengths. However unlike the responses generated for Google SERP and Google RR, the responses generated by Bing SERP and Bing RR for unigram queries were easier to read (i.e. higher Flesch reading ease score and lower Coleman-Liau readability index) than responses generated by them for other query lengths.

Like the responses in Google SERP and Google RR, the average concreteness and the ratio of concrete words of Bing SERP and Bing RR remained consistent regardless of query length. Unigram queries and queries $\geq$ 4-gram led to responses with a lower ratio of abstract words, which implies that even for Bing SERP and Bing RR, very short or very long queries led to responses with less number of common English words which would make the responses less suitable for autistic users.

Analysing the effect of prompt length on the accessibility indicator scores for GPT 3.5 and Gemini revealed that the two response groups were largely unaffected by the different prompt lengths. The only differences we found were that longer prompts led to longer sentences on average and the responses became more difficult to read (i.e. Flesch reading ease score dropped and Coleman-Liau readability index rose) for both the response groups. We noticed fluctuations in the average paragraph length of the responses. Still, we did not observe any trend in GPT 3.5 and the differences in Gemini were found to be not significant.

Overall each IST responds differently to varying query/prompt lengths. The fluctuations in the responses in each IST with varying query/prompt length create a very inconsistent online environment for autistic users when they are searching for information. Given their preference for routine and repeated behaviour [28], the observed fluctuations would pose a hindrance to an autistic user's web search experience.

Hence we also investigate how the ISTs respond to queries/prompts directly related to autism compared to more general queries and prompts.

Comparison of response groups over domain-specificity of ASD group queries/prompts. In our experiments investigating the impact of the domain-specificity of the ASD group queries/prompts, we found that a lot of the differences in the accessibility indicator scores observed were often not significant, unlike our observed differences across varying n-gram lengths of ASD group queries and prompts.

We did not observe many significant differences in the responses generated by Google SERP for domain-specific and general queries, however, we did observe trends which indicated that the structure of the responses generated for domain-specific queries was slightly better suited to autistic users. Google RR also showed trends of improvement in the way the responses were structured (i.e. lower ratio of paragraphs and higher ratio of headings and list items) but the responses were more text-heavy (i.e. more sentences and longer sentences and paragraphs). The concreteness of Google SERP and Google RR was similar for responses they generated for the two types of queries but domain-specific queries did lead to responses that consisted of more common English words. However, this difference did not hold true in reality.

In the case of Bing SERP and Bing RR, we observed that the structure of responses generated by Bing SERP for domain-specific queries was overall less suitable for autistic users but for Bing RR the responses generated for domain-specific queries although more text-heavy, were structured more suitably for autistic users. While the difference in structure did not hold true for Bing SERP, the differences was found to be significant for Bing RR. Like in Google SERP and Bing SERP, while the concreteness was not affected by the domain-specificity of the queries, the responses generated by Bing SERP and Bing RR for domain-specific queries had more common English words than for responses generated for general queries. However, once again, the difference was not significant.

For Gemini and GPT 3.5, the responses generated for domain-specific prompts were significantly less concrete on average and the structure of the responses generated by Gemini for domain-specific prompts were also less suited for autistic users as the responses consisted of longer sentences and paragraphs. Gemini responses for domain-specific prompts also consisted of more paragraphs than list items compared

to responses for general prompts, although the differences in the text structure indicators did not hold true in practice. GPT 3.5 on the other hand remain largely unaffected by the domain-specificity of the prompt.

Despite the lack of many significant differences between the responses generated by each response group for domain-specific and general queries, there was one important significant difference. For all response groups, the responses generated for domain-specific queries were significantly less easy to read (i.e. lower Flesch reading ease scores and higher Coleman-Liau readability index) than responses generated for general queries. The only exception to this was the difference in Flesch reading ease score for Google SERP wherein the difference was found to be not significant. From our general comparisons discussed previously, we already know that ASD group queries lead to ISTs generating responses that are not easy to read for autistic users. The results of our investigation into the impact of query/prompt domain-specificity on the IST responses suggest that the issue might be exacerbated for queries that are directly related to autism and hence the ISTs may be unintentionally hindering autistic people from learning about their own disorder.

Overall takeaways and suggested improvements. In an ideal scenario, the "perfect" IST would take advantage of the benefits offered by SEs and Chatbots, and provide autistic users with responses that cater to all their needs. Until such an IST is developed, existing ISTs can be improved to tackle the limitations that we bring to light through our results.

For instance, from our results, we know that SEs are good at structuring their responses in a way that is easily parseable by autistic users. Their responses are also comparatively easier to read (although not yet suitable for autistic users) but the textual content is more abstract and uses less common English words which makes it difficult to read for autistic users. However, LLM-based chatbots have the advantage of simpler navigation compared to traditional SEs. Accessibility guidelines for autistic users suggest online interfaces to offer consistent and simple navigations to make them "autism-friendly" [12, 66]. LLM-based chatbots simplify the web search process by generating a cohesive response for a user's prompt. On the other hand, in a traditional SE, the user has to explore a list of responses and determine their relevance to their information need.

This presents a trade-off scenario wherein either the autistic user uses a traditional SE and puts in the extra effort to rank the relevance of a list of resources that are easier to parse and easier to read, or the autistic user opts for the simpler navigation of a Chatbot but receives responses in huge paragraphs written in complex English sentences which are difficult to parse as well as understand.

From the study conducted by Yechiam et al. [96] we know that autistic users do tend to scroll more on the SERP compared to non-autistic users, which suggests an investigative behaviour in autistic users when looking for information on a traditional SE. Hence we could assume that autistic users may not mind the extra effort of browsing through a SERP to find relevant results for their query, and hence the strengths of the LLM-based chatbots could be implemented in the SEs response generation to improve the accessibility of SE responses.

The SEs could make use of LLMs to summarise the web documents, as experimented in [19] and [17], to frame the web snippets presented on their SERP to at least overcome the inconsistency in the concreteness of the textual information provided on the SERP. The readability of responses collected from all ISTs fails to meet the needs of autistic users thus efforts need to be made to improve the overall readability of content on the web. This can be achieved in two ways - (1) web browsers having a default feature that simplifies the textual content provided by the online ISTs, similar to the web browser extensions developed by [8] and [57] to simplify textual content in web resources and online medical texts respectively, and (2) a general push on creating web content that makes use of simpler language.

Our results from the investigation of the nature of the query on the accessibility of the IST responses present other issues that necessitate modifications to the algorithms that each IST uses to generate its responses.

Firstly, our results suggest that each IST responds differently to different query/prompt lengths. We know that web users are generally more prone to using shorter queries [79], and hence may not be affected by the fluctuations in the way the IST responds to longer queries/prompts compared to the shorter ones. However, we do not know if autistic users are also consistent in their query lengths like the general public. Hence, noting the observed negative impact of longer queries and prompts on the accessibility of IST

responses, it would be better for developers to adopt a precautionary approach and improve the algorithms to be more consistent in their response generation regardless of the length of the query/prompt to facilitate a more accessible and comfortable web search experience for autistic users.

Studies have found autism to significantly affect the quality of life of autistic people [48, 52]. Therefore autistic users require all the resources they can get to understand their disorder so that they can lead their lives as comfortably as possible. However our results show that not only do ISTs respond to queries typically asked by autistic users in a way that is not accessible to autistic users, the issue is further exacerbated for queries that are directly related to their disorder. Thereby ISTs are unintentionally barring autistic people from crucial information which could help their learn about their disorder and ways to cope with their daily struggles, and hence developers must look into ways to modify the algorithm to rectify the issue at the earliest.

6.2. Investigating the effect of the query/prompt reformulation on accessibility of IST responses

As discussed previously in Section 3.1, query reformulation and prompt engineering have been studied extensively on their effect on the responses generated by SEs and LLM-based chatbots [75, 55, 73, 90]. Thus we also investigate whether reformulating the query or prompt could lead to the ISTs responding differently to the ASD group queries. The queries and prompts were reformulated to explicitly state that the user is autistic using tactics described in Section 3.1.

In our investigation on the impact of query/prompt reformulation on the IST responses, we first computed the similarity between the responses generated by an IST for an original query/prompt and the responses generated by it for the reformulated query/prompt. The low similarity scores indicated that reformulating the query led to very different responses from all the ISTs. Hence we investigated whether the difference in the responses before and after query/prompt reformulation also led to a change in the accessibility indicator scores of each response group. From our results reported in Sections 5.2.2, 5.2.3, and 5.2.4, we know that there were significant changes in the indicator scores for all response groups. In this section, we elaborate on the inferences we draw from the reported results.

Effect on `SERP`. We observed significant differences in the accessibility indicator scores of both `Google SERP` and `Bing SERP` although the difference in the indicator scores due to query reformulation was more pronounced in `Google SERP` than in `Bing SERP`.

Upon reformulating the query to explicitly state that the user is autistic, the number of sentences increased significantly for both `Google SERP` and `Bing SERP`, although the sentences themselves became significantly shorter for both the response groups. However the responses generated by both the response groups were structured more as long chunks of text instead of short paragraphs or lists of bullet points which made them more difficult to parse for autistic users. Thus in general query reformulation led to both the response groups to generate responses that were structured in a way that is less accessible to autistic users.

Although we observed a decline in the accessibility of the structure of the responses generated by `Google SERP` and `Bing SERP` upon reformulating the query, we also observed a substantial improvement in the readability and concreteness of the responses generated by the two response groups. The responses generated by `Google SERP` in particular saw a dramatic increase in the Flesch reading ease score and a drop in the Coleman-Liau readability index at similar scale. For the first time in our entire study, we finally observed a response group generating responses that were largely readable by autistic users. We observed improvements in the readability of `Bing SERP` responses as well however unfortunately neither the increase in Flesch reading ease score nor the decrease in Coleman-Liau readability index was found to hold true in practice.

The average concreteness of the responses also increased significantly for both `Google SERP` and `Bing SERP` responses upon reformulating the queries. The ratio of abstract words remained unchanged for both the response groups but the ratio of abstract words increased significantly for `Google SERP` which implies that the total ratio of words categorised as abstract or concrete increased significantly for `Google SERP` responses. As discussed previously, our concreteness indicator scores were computed using the dataset compiled by Brysbaert et al. [13] which consists of over 40000 commonly known English lemmas.

Since the responses generated by Google SERP for the reformulated queries consisted of more words present in the dataset, we can infer that reformulating the query led to Google SERP responses consisting of more common English terms. We however did not observe such a change in the responses generated by Bing SERP after the query was reformulated.

Except for the less accessible structure of the responses, we observed that query reformulation led to a mostly significant improvement in both Google SERP and Bing SERP responses. Our query reformulation tactic was particularly successful in nudging Google SERP to generate responses that would actually be easy to read for autistic users.

Effect on RR. In the case the response groups under RR, the impact of query reformulation was not the same on the responses generated by Google RR and Bing RR. However much like our observations while comparing the effect of query reformulation on Google SERP and Bing SERP, the impact of our reformulations was more prominent in the responses generated by Google RR compared to Bing RR.

Google RR responses had lesser number of sentences which were longer on average for the reformulated queries compared to the responses it generated for the original queries. We also observed an increase in the ratio of the paragraphs and decrease in the ratio of headings and ratio of list items in Google RR when the query was reformulated. Our results collectively indicate that upon reformulating the query, the Google RR responses became more text-heavy and were more often arranged in large blocks of text instead of small paragraphs and lists thereby making the responses more difficult for an autistic user to parse through easily. In Bing RR we only observed a significant increase in the average length of the sentences and paragraphs which indicates that the responses generated by it for reformulated queries were much more text-heavy than the responses generated for the original queries. The structure of the Bing RR responses also consisted of more paragraphs upon query reformulation, however the increase was found to not hold true in practice.

The responses generated by Google RR were much easier to read after the query was reformulated (i.e. the Flesch reading ease score increased and the Coleman-Liau readability index decreased) with many responses surpassing the minimum threshold of 65 on the Flesch reading ease score making them easily readable for autistic users. The change in the readability of Bing RR responses was a bit more mixed, with a decrease in the Flesch reading ease score indicating a decline in readability but also a decrease in the Coleman-Liau readability index which indicates an improvement in the readability. Nevertheless the differences observed in the readability indicator scores for Bing RR after the query was reformulated were found to not hold in practice and regardless of the change, most responses remained unreadable by autistic users even after the query was reformulated.

Query reformulation also had a negative impact on the average concreteness of Google RR. We observed an increase in the average concreteness of Bing RR responses but the difference was found to be not significant. The ratio of concrete words increased significantly for Google RR responses while the ratio remained unchanged for Bing RR. The ratio of abstract words however increased significantly for both the response groups, hence we observed an overall increase in the number of words that were found in the dataset compiled by Brysbaert et al. [13] which in turn indicates an increase in the number of common English words in the responses generated by both the response groups after the query was reformulated.

Putting all our observations together we observed a significant improvement in the readability of Google RR responses but other than the responses becoming easier to read for autistic users, the structure and concreteness of the Google RR responses did not support the needs of autistic users and the issue was worsened after the query was reformulated. On the other hand, our results show that the accessibility of the responses generated by Bing RR remained largely unaffected by our query reformulation tactic.

Effect on Chatbot. Upon reformulating the prompt to state that the user is autistic, we observed significant differences mostly in Gemini responses but not in GPT 3.5.

The responses generated by Gemini for the reformulated prompts had more sentences which were also longer on average compared to the responses generated by the response group for the original prompt. There was also a significant increase in the ratio of headings and ratio of paragraphs while the ratio of list items remained practically unchanged. The paragraphs were longer on average for Gemini after the prompt was reformulated. Hence we observe that for Gemini, prompt reformulation led to the response

group generating responses structured in a way that is less suitable for autistic users. The structure of GPT 3.5 responses remained largely unaffected by the prompt reformulation other than the responses for reformulated prompts having a larger number of sentences and longer paragraphs on average which also indicate that prompt reformulation led to a decline in the accessibility of GPT 3.5 responses for autistic users in the context of the responses' structure.

We observed a decreased in Flesch reading ease scores for both the response groups after prompt reformulation but the Coleman-Liau readability index increased for Gemini but decreased for GPT 3.5. However all the differences in the readability indicator scores were found to be not significant. Hence in practice, we cannot make any reliable inferences on the impact of prompt reformulation on the readability of the responses generated by the two response groups.

The average concreteness on the other hand decreased significantly for Gemini but increased for GPT 3.5 although like most other indicator scores, the difference in GPT 3.5 was found to be not significant. The ratio of concrete words remained unaffected by prompt reformulation for both the response groups but the ratio of abstract words increased significantly for Gemini responses. The overall increase in the number of words of the Gemini responses which were categorised as concrete or abstract (using the concreteness ratings from the dataset compiled by Brysbaert et al. [13]) indicates that reformulating the prompt nudged Gemini to generate responses consisting of more common English words compared to the responses generated by it for the original prompts. We observed a similar increase in the ratio of abstract words for GPT 3.5 responses after prompt reformulation but the difference did not hold true in practice.

Unlike in SERP and RR responses, the impact of prompt reformulation on the accesibility of the responses generated by Chatbot was not very prominent. However we did notice that the variance reduced drastically for all indicator score distributions for both Gemini and GPT 3.5 which suggests that reformulating the prompt to indicate that the user is autistic nudges the two chatbots to generate responses in a particular style which caused the accessibility indicator scores to converge closer to the central values compared to the score distributions for the responses generated by the chatbots for the original prompts.

Overall takeaways from the observed impact of query and prompt reformulation. From our observations, we noticed that reformulating the query or prompt impacted the accessibility of the responses of each IST in a different way.

Google's algorithm appears to be the most sensitive to the query reformulation tactic used by us as seen from the significant differences in accessibility indicators for responses generated by both Google SERP and Google RR. The drastic improvement in the readability of Google responses was a surprising achievement. This is because until now none of our experiments yielded responses from any of the ISTs that would be considered readable by autistic users as per the readability criteria suggested by Yaneva et al. [94].

Our query reformulation tactic did not have as strong of an impact on Bing's algorithm. While we observed significant differences in the indicator scores of Bing SERP upon reformulating the queries, Bing RR was not affected significantly in terms of the accessibility of the responses to autistic users. Furthermore, even the significant differences in Bing SERP were not as drastic as the differences we observed in Google SERP.

There is no one way of reformulating a query, and from our results, we can see that not every SE responds the same way to a query reformulation as the other SE. Noting the positive impact on Google's responses, we must investigate the impact of other query reformulation tactics to ascertain if we can achieve the same improvement in the accessibility of responses generated by Bing as we observed in Google's responses.

In the context of the two chatbots, the impact of prompt reformulation was a lot more subtle compared to the impact of query reformulation on the SEs. While we did not observe a significant difference in many of the accessibility indicators, especially for GPT 3.5, we noticed that most of the indicator scores had lesser variance in their distributions after the prompt was reformulated. This indicates that although our prompt reformulation may not have nudged the chatbots to generate more accessible responses, it did lead to the chatbots adopting a more consistent style of phrasing their responses at least in the context of accessibility. This presents an unexpected solution to reduce the variability in the linguistic style of

responses generated by `Chatbot`s which would make the responses more predictable and hence more suitable for autistic users.

While query/prompt reformulation presents an optimistic solution to improving the accessibility of each IST's responses for autistic users, we must also take into account that Autism is under-diagnosed in several countries [62] therefore a lot of autistic users may not be aware they are on the spectrum and thus will never explicitly mention that they are autistic. Our reformulation tactics reveal that the ISTs can be nudged to produce more accessible responses and from the results discussed in Section 6.1, we know that ISTs respond differently to ASD group queries despite the lack of any explicit signals that the user is autistic. Hence we must look into implicit signals present in queries characteristic to autistic users that can be used to nudge the ISTs to generate responses that are more accessible to them.

7

Conclusion

The mass digitization of books and other sources of knowledge has led to the dissemination of information on a global scale. People with a steady internet connection can easily access information about their topics of interest without spending money or time to travel to the physical location where the information source exists [38]. With the Internet serving as a large knowledge base, online search systems like SEs and large-language model (LLM) powered chatbots have become popular means to access the information available on the web. With over 5.4 billion internet users across the globe [63], it is not possible to cater to every user's needs on a personal level. This leads to generalising the needs of the user base which often leads to the minority communities i.e. non-traditional users, not being represented appropriately. For our study, we focus on one such group of non-traditional users - Autistic users.

Autism Spectrum Disorder (ASD) is a neurological developmental disorder that affects the way the person behaves and interacts with their surroundings. Their attention issues, struggles with language and visual comprehension, stimulus sensitivity, etc. affect their social lives which in turn affects their quality of life [48, 52]. Online communication mediums have been found to help aid autistic people with their issues with socialising, with most autistic users enjoying making connections and sharing their interests on social media platforms [31, 87]. However, autistic users do not hold the same interest in online mediums when it comes to searching for information [31]. Furthermore, controlled experiments have revealed that autistic users are also not efficient in processing the information presented to them on the web [24, 25, 94]. Based on the results of such experiments, researchers have been compiling an extensive list of guidelines to improve the accessibility of web content for autistic users.

The accessibility guidelines have been used to develop assistive technologies for autistic users however they are yet to tackle autistic users' struggles with searching for information on the web. Thus autistic users have to still rely on online ISTs to find information on topics they need to learn about. However, whether these online tools are meeting their needs has not been investigated extensively. Therefore for our study, we focused on the online ISTs that autistic users interact with when searching for information and investigated the accessibility of the IST responses for autistic users.

Through our investigation, we aimed to answer two questions - (1) whether searching for information more accessible for an autistic user on a traditional search engine or through an LLM-based chatbot, and (2) whether reformulating the query/prompt to explicitly state that the user is autistic yields more accessible IST responses for autistic users.

To answer our questions, we focused on two types of online ISTs - the traditional SEs and the recently popular LLM-based chatbots. To collect responses from them, we required queries typically asked by autistic users. Due to the lack of standard datasets for the same, we resorted to generating synthetic queries by extracting common key phrases from Reddit posts created by users in subreddits dedicated to autistic users and their struggles. Our framework for generating the synthetic queries was inspired by frameworks used in similar studies for people with mental health disorders [22, 56] and we created a set of 250 synthetic queries, i.e. ASD group queries, to probe four popular ISTs - Google, Bing, Gemini, and ChatGPT 3.5. For better contextualization of our experiments, we also randomly sampled 250 queries from the Yahoo! Search Query Tiny sample dataset [91] and created a set of queries to represent the information needs of the general public i.e. the control group queries.

To investigate the accessibility of the responses generated by these four ISTs, we simulated the last 4 stages of the ISP model proposed by Kuhlthau [42] namely - Exploration, Formulation, Collection, and Presentation. We relied on the ISP stage to SE functionality mapping proposed by Milton and Pera[56] to collect responses from SEs that a user would typically come across at a particular stage of their ISP for any given query, i.e. we collected the first 10 web snippets presented on the SERP for a query to represent the Exploration and Formulation stages, and the first 10 web resources (RR) linked to those first 10 web snippets respectively to represent the Collection and Presentation stages. In chatbots, the stages are all condensed to a single functionality which is the back-and-forth conversation between the user and the chatbot, so we only collected the first response generated by them for a given prompt.

Due to the limited scope of this study, we focused only on the textual content of the collected responses. According to widely accepted web content accessibility guidelines for autistic users, for a text to be accessible to an autistic person, it should be concise and divided into small paragraphs or lists preceded by relevant headings [12]. The text should also be written in simple English [95] and not consist of niche jargon, metaphors and abstract concepts [12]. Based on these guidelines, we created quantifiable indicators, which we call accessibility indicators, to quantitatively compare the accessibility of IST responses on three major categories - (1) text structure, (2) text readability, and (3) text concreteness of the response.

To conduct our quantitative comparisons we categorised the collected responses into three response group types based on which functionality of the IST the response is collected from - (1) `SERP` for responses collected from SERP, (2) `RR` for responses collected from RR, and (3) `Chatbot` for responses collected from the LLM-based chatbots. The response group types were further divided into two response groups each to represent the responses collected from the given functionality of a particular IST. For `SERP`, we had `Google SERP` for responses collected from the Google SERP, and `Google RR`. Similarly for `RR`, we had `Google RR` and `Bing RR`. For `Chatbot` the two response groups were `Gemini` for responses generated by Gemini and `GPT 3.5` for responses generated by ChatGPT 3.5.

Comparing the accessibility indicator scores of the responses generated by each response group type for ASD group queries and control group queries revealed that all ISTs respond differently for ASD group queries compared to control group queries. Surprisingly the responses collected from each response group type for the control group queries were found to be significantly more accessible to autistic users than the responses collected for ASD group queries. Further investigation into how each response group type responds to the ASD group queries revealed that SEs and chatbot responses both have their strengths and limitations in the context of accessibility to autistic users. SE responses are better structured and comparatively easier to read (although the readability was found to not be suitable for autistic users) while chatbot responses were much more concrete compared to SE responses. Based on our results, we suggested making use of the LLM used by the chatbots to generate their responses to generate more concrete snippets for the SERP and paraphrase the content on RR to make the responses from each SE functionality more concrete.

Noting the large variance in the distributions of the accessibility indicator scores of each response group, we investigated whether the nature of the ASD group queries had an impact on the accessibility of the IST responses. We focused on two aspects of the query/prompt - (1) n-gram length, and (2) domain-specificity of the responses i.e. whether the query was directly related to autism or not.

We found that SE responses are more sensitive to the query length than the LLM-based chatbots. The only significant impact we observed of the prompt length on the accessibility of chatbot responses was that longer prompts led to a significant decline in the readability of the responses and the responses also became longer on average. In the case of the two SEs, unigram queries led to relatively different responses from the other query lengths. Unigram queries led to responses that were structured in a way that is more suitable for autistic users. However, the responses generated by Google for unigram queries were the most difficult to read while for Bing the responses for unigram queries were the easiest to read.

We do not have sufficient information on the query formulation behaviours of autistic users, and hence noting the fluctuations in the accessibility of both SE and chatbot responses over the query/prompt length, developers should adopt a more precautionary approach and modify the IST algorithms such that they provide more consistent responses across different query/prompt lengths, at least in the context of accessibility.

The impact of domain-specificity was different for every response group and many of the differences in the accessibility indicator scores of the response groups were found to not hold true in practice. However,

one significant difference between the responses generated for domain-specific and general queries was the significantly lower readability of the responses collected from all response groups for domain-specific queries. This suggests a trend of ISTs unintentionally barring autistic users from accessing and learning about their disorder.

Query reformulation and prompt engineering have been studied extensively on their effect on the responses generated by SEs and LLM-based chatbots [75, 55, 73, 90]. Our investigation of the impact of reformulating the ASD group queries and prompts to explicitly state that the user is autistic led to more prominent changes in the accessibility of SE responses than in the chatbot responses. Our reformulation tactic was particularly successful in nudging Google's algorithm to generate responses that are accessible to autistic users with many responses meeting the readability criteria to be suitable for autistic users. We observed similar improvements in the accessibility of Bing responses however the differences in the indicator scores were either of a much smaller scale or they did not hold in practice. Further investigation into the impact of different query reformulation strategies on each SE's responses may lead to more positive results in the future.

In the case of chatbots, while we did not observe many significant differences in the accessibility indicator scores after the prompt was reformulated, we did observe that the indicator scores had a smaller variance in their distributions after the prompt was reformulated. This suggests that while our prompt reformulation strategy may not have been successful in nudging the chatbots to generate more accessible responses, it led to the chatbots using a more consistent style of phrasing their responses at least in the context of accessibility.

While our query and prompt reformulation strategies presented optimistic solutions to improving the accessibility of IST responses, we need to consider the fact that ASD is largely undiagnosed which means that many autistic users in the world are not aware of their diagnosis and thus would never reformulate their queries to explicitly state that they are autistic. However, from our results, we know that the ISTs can be nudged to generate more accessible responses and from our previous results we know that ISTs respond differently to ASD group queries compared to control group queries. Thus future investigation must look into how ISTs can be implicitly nudged to generate more accessible responses by analysing the nature of the queries typically formulated by autistic users.

Our study had certain limitations which could be used as opportunities for future work. Firstly, we focused mainly on the textual component of the responses provided by online ISTs. Studies have revealed that autistic people benefit greatly from visual stimuli during information search as long as the stimulus is relevant to the information presented [95, 70]. Therefore, future work could implement similar investigation on online ISTs but instead focus on the visual components of the responses. The more advanced versions of the popular LLM-based chatbots now also support multimodal responses but how this multimodality impacts the accessibility of the responses to autistic users remains to be answered in future studies. We also had to make several assumptions when simulating the ISP for autistic users, the most important assumption being that the ISP model proposed by Kuhlthau [42] can also be extended to include the search behaviours of autistic people. Studies conducted after the completion of this work may propose a model that is far better suited to autistic users but our study can then serve as a framework that can then be adjusted to simulate the new model of the information search process.

References

[1] URL: https://wave.webaim.org/cognitive.

[2] W. W. A. I. (WAI). Ian, data entry clerk with autism, June 2024. URL: https://www.w3.org/WAI/people-use-web/user-stories/story-two/.

[3] G. Allen, A. Milton, K. L. Wright, J. A. Fails, C. Kennington, and M. S. Pera. Supercalifragilisticexpialidocious: why using the "right" readability formula in children's web search matters. In *European Conference on Information Retrieval*, pages 3–18. Springer, 2022.

[4] V. Arlington, A. P. Association, et al. Diagnostic and statistical manual of mental disorders. *American Psychiatric Association*, 5:612–613, 2013.

[5] P. Benford and P. Standen. The internet: a comfortable communication medium for people with asperger syndrome (as) and high functioning autism (hfa)? *Journal of Assistive Technologies*, 3(2):44–53, 2009.

[6] D. Bilal and J. Gwizdka. Children's query types and reformulations in google search. *Information Processing & Management*, 54(6):1022–1041, 2018.

[7] E. Billstedt, C. Gillberg, and C. Gillberg. Autism after adolescence: population-based 13-to 22-year follow-up study of 120 individuals with autism diagnosed in childhood. *Journal of autism and developmental disorders*, 35:351–360, 2005.

[8] J. Bingel, G. Paetzold, and A. Søgaard. Lexi: a tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, 2018.

[9] S. Bölte, K. D. Bartl-Pokorny, U. Jonsson, S. Berggren, D. Zhang, E. Kostrzewa, T. Falck-Ytter, C. Einspieler, F. B. Pokorny, E. J. Jones, et al. How can clinicians detect and treat autism early? methodological trends of technology use in research. *Acta paediatrica*, 105(2):137–144, 2016.

[10] A. Bosseler and D. W. Massaro. Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. *Journal of autism and developmental disorders*, 33:653–672, 2003.

[11] K. Bottema-Beutel, S. K. Kapp, J. N. Lester, N. J. Sasson, and B. N. Hand. Avoiding ableist language: suggestions for autism researchers. *Autism in adulthood*, 2021.

[12] T. C. P. Britto and E. B. Pizzolato. Towards web accessibility guidelines of interaction and interface ... Apr. 2016. URL: https://www.researchgate.net/publication/301552021_Towards_Web_Accessibility_Guidelines_of_Interaction_and_Interface_Design_for_People_with_Autism_Spectrum_Disorder.

[13] M. Brysbaert, A. B. Warriner, and V. Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911, 2014.

[14] M. C. Buzzi, G. Paolini, C. Senette, M. Buzzi, and M. T. Paratore. Designing an accessible web app to teach piano to students with autism. In *Proceedings of the 13th Biannual Conference of the Italian SIGCHI Chapter: Designing the next Interaction*, CHItaly '19, Padova, Italy. Association for Computing Machinery, 2019. ISBN: 9781450371902. DOI: 10.1145/3351995.3352037. URL: https://doi.org/10.1145/3351995.3352037.

[15] E. Cary, A. Rao, E. S. M. Matsuba, and N. Russo. Barriers to an autistic identity: how rrbs may contribute to the underdiagnosis of females. *Research in Autism Spectrum Disorders*, 109:102275, 2023.

[16] CDC. About autism spectrum disorder, May 2024. URL: https://www.cdc.gov/autism/about/index.html#cdc_disease_basics_overview-overview.

[17] Y. Chang, K. Lo, T. Goyal, and M. Iyyer. Booookscore: a systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*, 2023.

[18] J. Chen, J. Mao, Y. Liu, F. Zhang, M. Zhang, and S. Ma. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of the web conference 2021*, pages 743–755, 2021.

[19] G. Chhikara, A. Sharma, V. Gurucharan, K. Ghosh, and A. Chakraborty. Lamsum: a novel framework for extractive summarization of user generated content using llms. *arXiv preprint arXiv:2406.15809*, 2024.

[20] M. Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975.

[21] A. Dattolo and F. L. Luccio. A review of websites and mobile applications for people with autism spectrum disorders: towards shared guidelines. *Smart Objects and Technologies for Social Good*:264–273, July 2017. DOI: 10.1007/978-3-319-61949-1_28.

[22] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2098–2110, San Jose, California, USA. Association for Computing Machinery, 2016. ISBN: 9781450333627. DOI: 10.1145/2858036.2858207. URL: https://doi.org/10.1145/2858036.2858207.

[23] B. Driver and V. Chester. The recognition and diagnosis of autism in women and girls: a literature review. *Advances in Autism*, 7(3):194–207, 2021.

[24] S. Eraslan, V. Yaneva, Y. Yesilada, and S. Harper. Do web users with autism experience barriers when searching for information within web pages? In *Proceedings of the 14th International Web for All Conference*, W4A '17, Perth, Western Australia, Australia. Association for Computing Machinery, 2017. ISBN: 9781450349000. DOI: 10.1145/3058555.3058566. URL: https://doi.org/10.1145/3058555.3058566.

[25] S. Eraslan, V. Yaneva, Y. Yeşilada, and S. Harper. Web users with autism: eye tracking evidence for differences. *Behaviour & Information Technology*, 38:678–700, 2018. URL: https://api.semanticscholar.org/CorpusID:70245307.

[26] S. Eraslan, Y. Yesilada, V. Yaneva, and L. A. Ha. "keep it simple!": an eye-tracking study for exploring complexity and distinguishability of web pages for people with autism. *Universal Access in the Information Society*, 20(1):69–84, 2020. DOI: 10.1007/s10209-020-00708-9.

[27] R. Flesch. How to write plain english, July 2016. URL: https://web.archive.org/web/20160712094308/http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml.

[28] C. for Disease Control and Prevention. Signs and symptoms of autism spectrum disorder, Jan. 2024. URL: https://www.cdc.gov/autism/signs-symptoms/index.html.

[29] T. W. Frazier, E. A. Youngstrom, L. Speer, R. Embacher, P. Law, J. Constantino, R. L. Findling, A. Y. Hardan, and C. Eng. Validation of proposed dsm-5 criteria for autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(1):28–40, 2012.

[30] U. Frith and M. Snowling. Reading for meaning and reading for sound in autistic and dyslexic children. *British journal of developmental psychology*, 1(4):329–342, 1983.

[31] K. Gillespie-Lynch, S. K. Kapp, C. Shane-Simpson, D. S. Smith, and T. Hutman. Intersections between the autism spectrum and the internet: perceived benefits and preferred functions of computer-mediated communication. *Intellectual and developmental Disabilities*, 52(6):456–469, 2014.

[32] F. Happé and U. Frith. The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *Journal of autism and developmental disorders*, 36:5–25, 2006.

[33] E. M. Hassrick, L. G. Holmes, C. Sosnowy, J. Walton, and K. Carley. Benefits and risks: a systematic review of information and communication technology use by autistic people. *Autism in Adulthood*, 3(1):72–84, 2021.

[34] P. L. Howard and F. Sedgewick. 'anything but the phone!': communication mode preferences in the autism community. *Autism*, 25(8):2265–2278, 2021.

[35] P. Howlin and P. Moss. Adults with autism spectrum disorders. *The Canadian Journal of Psychiatry*, 57(5):275–283, 2012.

[36] T. Karunaratne and A. Adesina. Is it the new google: impact of chatgpt on students' information search habits, 2023.

[37] L. Kenny, C. Hattersley, B. Molins, C. Buckley, C. Povey, and E. Pellicano. Which terms should be used to describe autism? perspectives from the uk autism community. *Autism*, 20(4):442–462, 2016.

[38] S. Khan, S. Khan, and M. Aftab. Digitization and its impact on economy. *International Journal of Digital Library Services*, 5(2):138–149, 2015.

[39] J. Kim, J. Lee, E. Park, and J. Han. A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, 10(1):11846, 2020.

[40] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, 1975.

[41] N. Koteyko. Understanding autistic adults' use of social media. *Proceedings of the ACM on Human-Computer Interaction*, 2023.

[42] C. C. Kuhlthau. Inside the search process: information seeking from the user's perspective. *J. Am. Soc. Inf. Sci.*, 42:361–371, 1991. URL: `https://api.semanticscholar.org/CorpusID:14416802`.

[43] M.-C. Lai and S. Baron-Cohen. Identifying the lost generation of adults with autism spectrum conditions. *The Lancet Psychiatry*, 2(11):1013–1027, 2015.

[44] A. Larco, E. Diaz, C. Yanez, and S. Luján-Mora. Autism and web-based learning: review and evaluation of web apps. *Trends and Advances in Information Systems and Technologies: Volume 2 6*:1434–1443, 2018.

[45] M. Liao, H. Duan, and G. Wang. Application of machine learning techniques to detect the children with autism spectrum disorder. *Journal of Healthcare Engineering*, 2022(1):9340027, 2022.

[46] C.-Y. Lin. Rouge: a package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[47] C.-S. Lin, S.-H. Chang, W.-Y. Liou, and Y.-S. Tsai. The development of a multimedia online language assessment tool for young children with autism. *Research in developmental disabilities*, 34(10):3553–3565, 2013.

[48] L.-Y. Lin and P.-C. Huang. Quality of life and its related factors for adults with autism spectrum disorder. *Disability and rehabilitation*, 41(8):896–903, 2019.

[49] S. Lock. What is ai chatbot phenomenon chatgpt and could it replace humans?, Dec. 2022. URL: `https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans`.

[50] J. Locke, E. H. Ishijima, C. Kasari, and N. London. Loneliness, friendship quality and the social networks of adolescents with high-functioning autism in an inclusive school setting. *Journal of research in special educational needs*, 10(2):74–81, 2010.

[51] C. Lord and S. L. Bishop. Recent advances in autism research as reflected in dsm-5 criteria for autism spectrum disorder. *Annual review of clinical psychology*, 11(1):53–70, 2015.

[52] D. Mason, H. McConachie, D. Garland, A. Petrou, J. Rodgers, and J. R. Parr. Predictors of quality of life for autistic adults. *Autism Research*, 11(8):1138–1147, 2018.

[53] O. Matthews, S. Eraslan, V. Yaneva, A. Davies, Y. Yesilada, M. Vigo, and S. Harper. Combining trending scan paths with arousal to model visual behaviour on the web. *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 2019. DOI: 10.1145/3320435.3320446.

[54] M. O. Mazurek. Social media use among adults with autism spectrum disorders. *Computers in Human Behavior*, 29(4):1709–1714, 2013.

[55] B. Meskó. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of medical Internet research*, 25:e50638, 2023.

[56] A. Milton and M. S. Pera. Into the unknown: exploration of search engines' responses to users with depression and anxiety. *ACM Trans. Web*, 17(4), July 2023. ISSN: 1559-1131. DOI: 10.1145/3580283. URL: https://doi.org/10.1145/3580283.

[57] N. Nakhaie Ahooie. *Enhancing access to medical literature through an LLM-based browser extension*. Master's thesis, N. Nakhaie Ahooie, 2024.

[58] J. Nyrenius, J. Eberhard, M. Ghaziuddin, C. Gillberg, and E. Billstedt. The 'lost generation'in adult psychiatry: psychiatric, neurodevelopmental and sociodemographic characteristics of psychiatric patients with autism unrecognised in childhood. *BJPsych open*, 9(3):e89, 2023.

[59] I. M. O'Connor and P. D. Klein. Exploration of strategies for facilitating the reading comprehension of high-functioning students with autism spectrum disorders. *Journal of autism and developmental disorders*, 34:115–127, 2004.

[60] E. O'Nions, I. Petersen, J. E. Buckman, R. Charlton, C. Cooper, A. Corbett, F. Happé, J. Manthorpe, M. Richards, R. Saunders, and et al. Autism in england: assessing underdiagnosis in a population-based cohort study of prospectively collected primary care data. *The Lancet Regional Health - Europe*, 29:100626, June 2023. DOI: 10.1016/j.lanepe.2023.100626.

[61] B. Oliveira and C. Teixeira Lopes. The evolution of web search user interfaces - an archaeological analysis of google search engine result pages. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, CHIIR '23, pages 55–68, Austin, TX, USA. Association for Computing Machinery, 2023. ISBN: 9798400700354. DOI: 10.1145/3576840.3578320. URL: https://doi.org/10.1145/3576840.3578320.

[62] W. H. Organisation. Autism, Mar. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders.

[63] A. Petrosyan. Internet and social media users in the world 2024, May 2024. URL: https://www.statista.com/statistics/617136/digital-population-worldwide/#:~:text=As%20of%20April%202024%2C%20there,population%2C%20were%20social%20media%20users..

[64] B. O. Ploog. Stimulus overselectivity four decades later: a review of the literature and its implications for current research in autism spectrum disorder. *Journal of autism and developmental disorders*, 40:1332–1349, 2010.

[65] L. Quoniam, F. Balme, H. Rostaing, E. Giraud, and J.-M. Dou. Bibliometric law used for information retrieval. *Scientometrics*, 41:83–91, Aug. 1998. DOI: 10.1007/BF02457969.

[66] D. M. Raymaker, S. K. Kapp, K. E. McDonald, M. Weiner, E. Ashkenazy, and C. Nicolaidis. Development of the aaspire web accessibility guidelines for autistic web users. *Autism in Adulthood*, 1(2):146–157, 2019.

[67] Reddit. Data api terms, Apr. 2023. URL: https://www.redditinc.com/policies/data-api-terms.

[68] Reddit. Reddit 10-q quarterly report, May 2024. URL: https://investor.redditinc.com/financials/sec-filings/sec-filings-details/default.aspx?FilingId=17519460.

[69] J. Redish. Readability formulas have even more limitations than klare discusses. *ACM J. Comput. Doc.*, 24(3):132–137, Aug. 2000. ISSN: 1527-6805. DOI: 10.1145/344599.344637. URL: https://doi.org/10.1145/344599.344637.

[70] M. Rezae, N. Chen, D. McMeekin, T. Tan, A. Krishna, and H. Lee. The evaluation of a mobile user interface for people on the autism spectrum: an eye movement study. *International Journal of Human-Computer Studies*, 142:102462, 2020. DOI: 10.1016/j.ijhcs.2020.102462.

[71] C. Risi. Google search backtracks on recent ui changes, Jan. 2020. URL: https://www.criticalhit.net/technology/google-search-backtracks-on-recent-ui-changes/.

[72] S. Rubio-Martín, M. T. García-Ordás, M. Bayón-Gutiérrez, N. Prieto-Fernández, and J. A. Benítez-Andrades. Early detection of autism spectrum disorder through ai-powered analysis of social media texts. In *2023 IEEE 36th International symposium on computer-based medical systems (CBMS)*, pages 235–240. IEEE, 2023.

[73] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, pages 881–890, 2010.

[74] A. Sarica, A. Quattrone, and A. Quattrone. Introducing the rank-biased overlap as similarity measure for feature importance in explainable machine learning: a case study on parkinson's disease. In *International Conference on Brain Informatics*, pages 129–139. Springer, 2022.

[75] D. C. Schmidt, J. Spencer-Smith, Q. Fu, and J. White. Cataloging prompt patterns to enhance the discipline of prompt engineering. *URL: https://www. dre. vanderbilt. edu/~ schmidt/PDF/ADA_Europe_Position_Paper. pdf [accessed 2023-09-25]*, 2023.

[76] D. Sharma, R. Shukla, A. K. Giri, and S. Kumar. A brief review on search engine optimization. *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*:687–692, 2019. URL: https://api.semanticscholar.org/CorpusID:199059492.

[77] P. T. Shattuck, G. I. Orsmond, M. Wagner, and B. P. Cooper. Participation in social activities among adolescents with an autism spectrum disorder. *PloS one*, 6(11):e27176, 2011.

[78] D.-Y. Song, S. Y. Kim, G. Bong, J. M. Kim, and H. J. Yoo. The use of artificial intelligence in screening and diagnosis of autism spectrum disorder: a literature review. *Journal of the Korean Academy of Child and Adolescent Psychiatry*, 30(4):145, 2019.

[79] A. Spink and B. J. Jansen. A study of web search trends. *Webology*, 1(2):4, 2004.

[80] O. I. Talantseva, R. S. Romanova, E. M. Shurdova, T. A. Dolgorukova, P. S. Sologub, O. S. Titova, D. F. Kleeva, and E. L. Grigorenko. The global prevalence of autism spectrum disorder: a three-level meta-analysis. *Frontiers in psychiatry*, 14:1071181, 2023.

[81] N. Thin, N. Hung, S. Venkatesh, and D. Phung. Estimating support scores of autism communities in large-scale web information systems:347–355, 2017.

[82] A. L. Uitdenbogerd, M. Spichkova, and M. Alzahrani. Web-based search: how do animated user interface elements affect autistic and non-autistic users? *arXiv preprint arXiv:2211.11993*, 2022.

[83] UNESCO. Right to information. URL: https://www.unesco.org/en/right-information.

[84] G. I. van Schalkwyk, C. E. Marin, M. Ortiz, M. Rolison, Z. Qayyum, J. C. McPartland, E. R. Lebowitz, F. R. Volkmar, and W. K. Silverman. Social media use, friendship quality, and the moderating role of anxiety in adolescents with autism spectrum disorder. *Journal of autism and developmental disorders*, 47:2805–2813, 2017.

[85] W3C. Vision for w3c, Apr. 2024. URL: https://www.w3.org/TR/w3c-vision/.

[86] G. Wan, X. Kong, B. Sun, S. Yu, Y. Tu, J. Park, C. Lang, M. Koh, Z. Wei, Z. Feng, et al. Applying eye tracking to identify autism spectrum disorder in children. *Journal of autism and developmental disorders*, 49:209–215, 2019.

[87] T. Wang, M. Garfield, P. Wisniewski, and X. Page. Benefits and challenges for social media users on the autism spectrum. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, pages 419–424, 2020.

[88] D. L. Williams, N. J. Minshew, and G. Goldstein. Further understanding of complex information processing in verbal adolescents and adults with autism spectrum disorders. *Autism*, 19(7):859–867, 2015.

[89] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang. A brief overview of chatgpt: the history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023. DOI: 10.1109/JAS.2023.123618.

[90] Z. Xu, X. Luo, J. Yu, and W. Xu. Mining web search engines for query suggestion. *Concurrency and Computation: Practice and Experience*, 23(10):1101–1113, 2011.

[91] Yahoo! search query tiny sample, 2009. URL: https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=43.

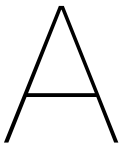
[92] V. Yaneva, S. Eraslan, Y. Yesilada, R. Mitkov, et al. Detecting high-functioning autism in adults using eye tracking and machine learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(6):1254–1261, 2020.

[93] V. Yaneva and R. Evans. Six good predictors of autistic text comprehension. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 697–706, 2015.

[94] V. Yaneva, L. A. Ha, S. Eraslan, and Y. Yesilada. Adults with high-functioning autism process web pages with similar accuracy but higher cognitive effort compared to controls. In *Proceedings of the 16th International Web for All Conference*, W4A '19, San Francisco, CA, USA. Association for Computing Machinery, 2019. ISBN: 9781450367165. DOI: 10.1145/3315002.3317563. URL: https://doi.org/10.1145/3315002.3317563.

[95] V. Yaneva, I. Temnikova, and R. Mitkov. Accessible texts for autism: an eye-tracking study. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '15, pages 49–57, Lisbon, Portugal. Association for Computing Machinery, 2015. ISBN: 9781450334006. DOI: 10.1145/2700648.2809852. URL: https://doi.org/10.1145/2700648.2809852.

[96] E. Yechiam and E. Yom-Tov. Unique internet search strategies of individuals with self-stated autism: quantitative analysis of search engine users' investigative behaviors. *Journal of Medical Internet Research*, 23(7):e23829, 2021.



List of subreddits

1. r/autism
2. r/aspergers
3. r/AutismInWomen
4. r/AutismTranslated
5. r/AutisticAdults
6. r/AutisticPride
7. r/aspergirls
8. r/AskAutism
9. r/autism_controversial
10. r/AutismCertified
11. r/TrueEvilAutism
12. r/AutisticPeeps
13. r/SpicyAutism
14. r/austismlevel2and3
15. r/AutisticLiberation
16. r/Autism_Pride

B

Frequency distribution of extracted key phrases from Reddit posts

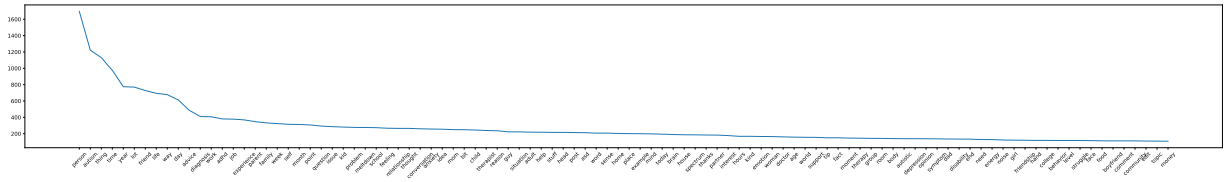


Figure B.1: Frequency distribution of 100 most common unigrams

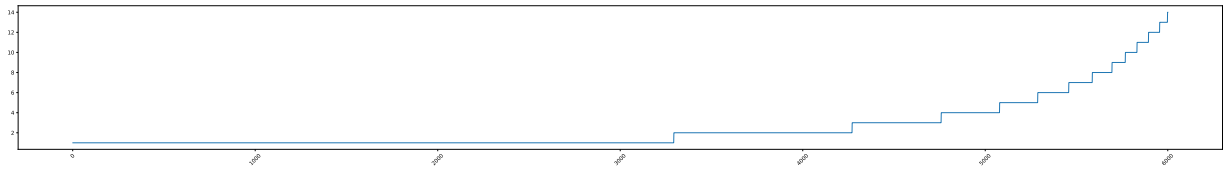


Figure B.2: Frequency distribution of unigrams in ascending order of frequency

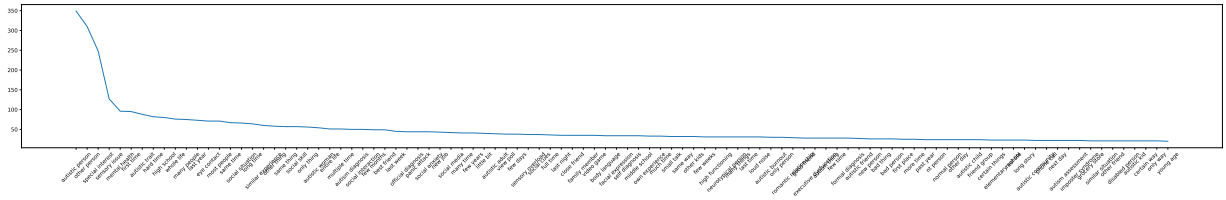


Figure B.3: Frequency distribution of 100 most common bigrams

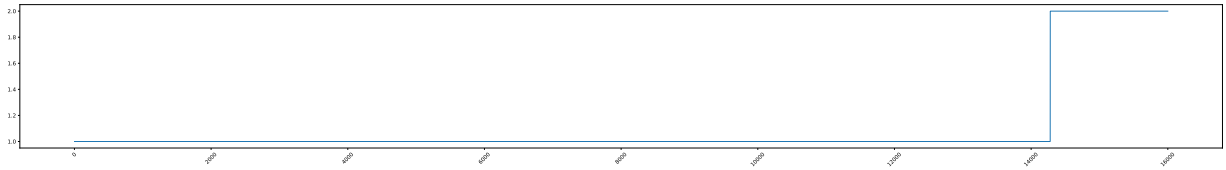


Figure B.4: Frequency distribution of bigrams in ascending order of frequency

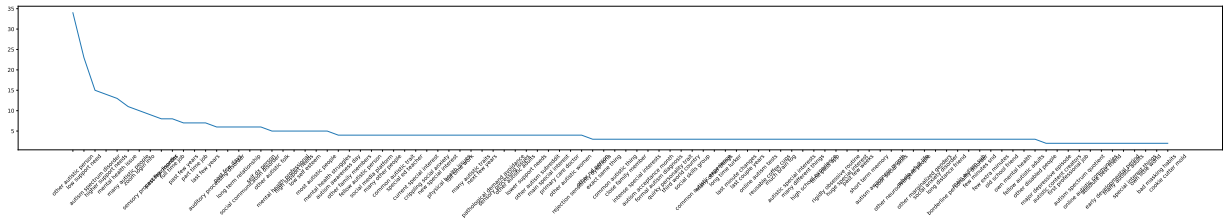


Figure B.5: Frequency distribution of 100 most common trigrams

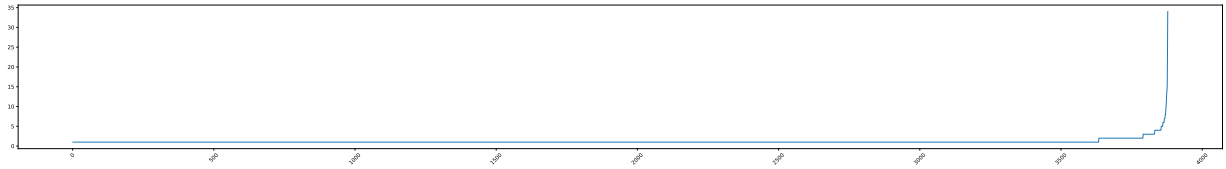


Figure B.6: Frequency distribution of trigrams in ascending order of frequency

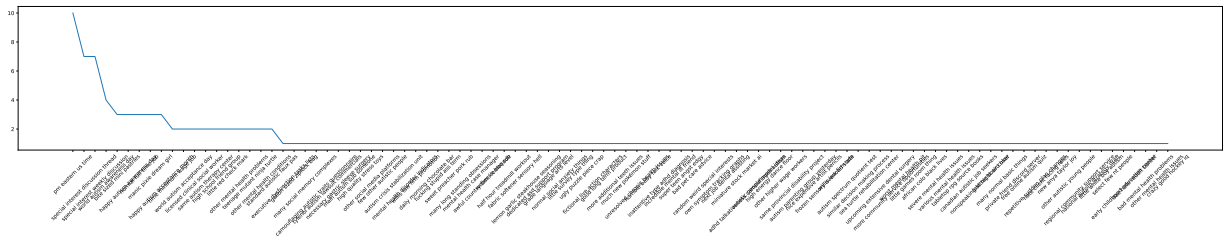


Figure B.7: Frequency distribution of 100 most common quadgrams

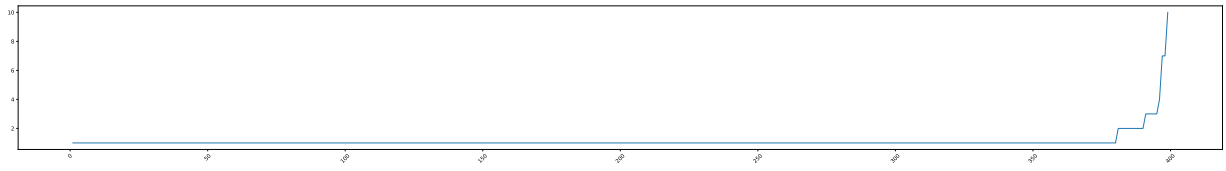


Figure B.8: Frequency distribution of quadgrams in ascending order of frequency

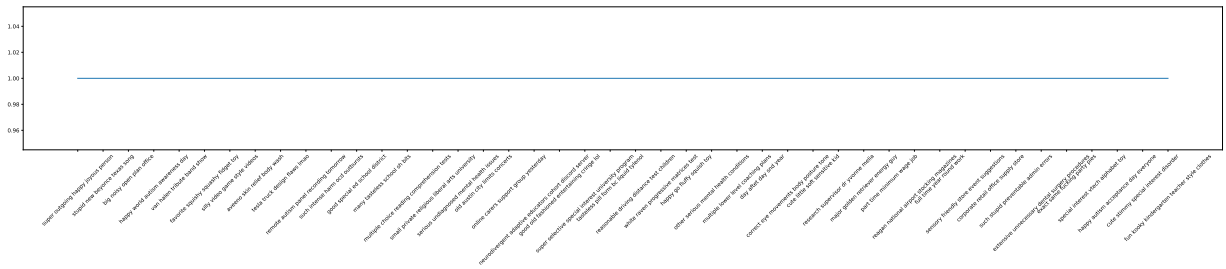


Figure B.9: Frequency distribution of 100 most common ngrams (n>4)

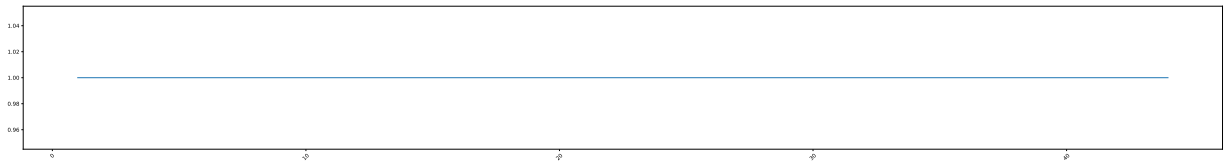


Figure B.10: Frequency distribution of ngrams (n>4) in ascending order of frequency

C

Quantitative comparison of accessibility of response group types for control and ASD group queries

Accessibility Indicator	SERP		RR		Chatbot	
	Control	**ASD**	**Control**	**ASD**	**Control**	**ASD**
# sentences	4.22	4.13	119.06	123.64	5	5
Avg sentence length	9.97	10.35 *	4.01	4.88 *	10.5	9.44
Ratio of headings	0.27	0.27	0.06	0.05	0	0 *
Ratio of list items	0	0 *	0.33	0.36 *	0	0
Ratio of paragraphs	0.71	0.72 *	0.58	0.57	1	1
Avg paragraph length	17.75	19.25 *	19.49	22.56 *	27	25
Flesch reading ease score	64.55	59.32 *	62.52	59.06 *	52.29	42.41 *
Coleman-Liau readability index	11.19	11.36	11.38	11.88 *	12.53	13.86 *
Avg concreteness	2.5	2.37 *	2.5	2.42 *	2.49	2.36 *
Ratio of concrete words	0.05	0.04 *	0.04	0.04 *	0.06	0.04 *
Ratio of abstract words	0.35	0.49 *	0.26	0.4 *	0.53	0.61 *

Table C.1: Median values of accessibility indicators for SERP, RR, Chatbot for Control and ASD group queries.

* indicates that the distribution for the accessibility indicator was found to be significantly different ($p < 0.05$) for control and ASD group queries for a response group type using the Mann-Whitney U test for statistical significance.

D

Quantitative comparison of accessibility of response group for ASD group queries

Accessibility Indicator	SERP		RR		Chatbot	
	Google SERP	Bing SERP	Google RR	Bing RR	Gemini	GPT 3.5
# sentences	4	4.29 *	111.8	137.29 *	39	3.5 *
Avg sentence length	9.23	12.17 *	5.04	4.65 *	5.72	18 *
Ratio of headings	0.28	0.26 *	0.06	0.05 *	0.1	0 *
Ratio of list items	0	0 *	0.34	0.37 *	0.4	0 *
Ratio of paragraphs	0.72	0.73 *	0.59	0.56 *	0.5	1 *
Avg paragraph length	16.73	23.06 *	22.78	22.44	9	55 *
Flesch reading ease score	61.76	58.08 *	58.73	59.38	38.45	45.98 *
Coleman-Liau readability index	10.96	11.86 *	11.91	11.88	15.57	12.97 *
Avg concreteness	2.4	2.34 *	2.42	2.42	2.37	2.34
Ratio of concrete words	0.04	0.04	0.04	0.04	0.04	0.05 *
Ratio of abstract words	0.44	0.55 *	0.4	0.38	0.5	0.7 *

Table D.1: Median values of accessibility indicators for different response groups.

* indicates that the distribution for the accessibility indicator was found to be significantly different ($p < 0.05$) for the two response groups under the response group type using the Mann-Whitney U test for statistical significance.

E

Quantitative comparison of accessibility of response groups for different query & prompt lengths

Accessibility Indicator	Response Group	1-gram	2-gram	3-gram	4-gram	>4 -gram
# sentences	Google SERP *	3.56	4.27	4.14	4.62	5.2
	Google RR	117.25	108.78	107.77	109	151.78
	Bing SERP *	3.58	4.22	4.83	5.24	5.4
	Bing RR *	100.75	150	139.32	155.58	121.16
	Gemini	36	40	46	33.5	47
	GPT 3.5 *	3	3	4	4	4
Avg sentence length	Google SERP *	9.32	9.12	9.29	9	6.99
	Google RR *	4.09	5.37	5.49	5.04	4.31
	Bing SERP	12.12	12.4	11.78	12.82	11.91
	Bing RR *	2.92	4.89	5.43	5.02	5.2
	Gemini *	5.51	5.36	6.3	6.07	5.78
	GPT 3.5 *	16.42	18.16	19.29	21	21.71
Ratio of headings	Google SERP *	0.32	0.26	0.27	0.24	0.22
	Google RR *	0.07	0.05	0.05	0.04	0.04
	Bing SERP *	0.29	0.26	0.22	0.22	0.22
	Bing RR *	0.03	0.05	0.06	0.05	0.06
	Gemini *	0.1	0.09	0.11	0.15	0.16
	GPT 3.5	0	0	0	0	0
Ratio of list items	Google SERP	0	0	0	0	0
	Google RR	0.39	0.3	0.36	0.41	0.29
	Bing SERP *	0	0	0	0.04	0.03
	Bing RR *	0.48	0.32	0.34	0.42	0.35
	Gemini	0.38	0.41	0.4	0.36	0.35
	GPT 3.5 *	0	0	0	0	0
Ratio of paragraphs	Google SERP *	0.68	0.74	0.72	0.72	0.78
	Google RR *	0.55	0.65	0.6	0.54	0.64
	Bing SERP *	0.71	0.73	0.76	0.74	0.75
	Bing RR *	0.49	0.62	0.6	0.51	0.6
	Gemini	0.5	0.48	0.5	0.45	0.52
	GPT 3.5 *	1	1	1	1	1

Avg paragraph length	Google SERP *	15.12	17.66	17.56	18.07	16.72
	Google RR *	16.52	26.82	27.37	22.02	27.36
	Bing SERP *	21.16	23.19	24.88	24.24	23.86
	Bing RR *	13.66	26.12	26.71	27.47	24.04
	Gemini	10.8	4.96	8.96	6.92	11.42
	GPT 3.5 *	52.34	48.5	67.5	47.34	59.25
Flesch reading ease score	Google SERP *	56.2	68.38	64.66	77.13	59.55
	Google RR	56.91	62.48	56.92	65.19	53.1
	Bing SERP	58.34	61.11	54.83	60.56	56.44
	Bing RR	61.3	60.66	54.74	59.52	58.8
	Gemini	42.53	40.2	33.46	40.07	26.67
	GPT 3.5 *	50.11	50.26	37.42	43.34	39.99
Coleman-Liau readability index	Google SERP *	12.2	9.34	10.09	9.2	10.46
	Google RR	12.36	11.2	12.12	11.32	12.68
	Bing SERP	11.39	11.42	12.58	12.61	12.97
	Bing RR *	11.68	11.1	12.7	12.76	12.1
	Gemini *	13.74	15.28	16.91	15.1	17.24
	GPT 3.5 *	11.28	11.32	14.6	13.5	13.78
Avg concreteness	Google SERP	2.39	2.43	2.38	2.39	2.44
	Google RR *	2.44	2.44	2.4	2.41	2.47
	Bing SERP	2.3	2.32	2.37	2.37	2.33
	Bing RR	2.43	2.42	2.42	2.4	2.34
	Gemini	2.35	2.37	2.41	2.32	2.38
	GPT 3.5	2.34	2.36	2.34	2.34	2.29
Ratio of concrete words	Google SERP *	0.04	0.05	0.03	0.03	0.02
	Google RR *	0.03	0.04	0.04	0.04	0.04
	Bing SERP	0.04	0.04	0.04	0.03	0.04
	Bing RR *	0.03	0.04	0.04	0.04	0.03
	Gemini	0.04	0.03	0.04	0.04	0.03
	GPT 3.5 *	0.06	0.05	0.04	0.04	0.04
Ratio of abstract words	Google SERP *	0.42	0.45	0.47	0.38	0.36
	Google RR *	0.31	0.44	0.45	0.4	0.34
	Bing SERP *	0.56	0.56	0.56	0.45	0.46
	Bing RR *	0.26	0.42	0.44	0.4	0.38
	Gemini	0.5	0.5	0.5	0.48	0.5
	GPT 3.5	0.7	0.69	0.72	0.68	0.7

Table E.1: Median values of accessibility indicators for different n-gram lengths of queries/prompts for all response groups.

* indicates that the distribution for the accessibility indicator was found to be significantly different ($p < 0.05$) for the response group's responses for different ngram lengths of the query/prompt using the Kruskal-Wallis H test for statistical significance.

F

Quantitative comparison of accessibility of response groups for domain-specific and general queries & prompts

Accessibility Indicator	Response Group	Domain-specific Query/Prompt	General Query/Prompt
# sentences	Google SERP	3.71	4.11
	Google RR	114	111.6
	Bing SERP	4.62	4.2
	Bing RR	144	135.12
	Gemini	38	39
	GPT 3.5 *	4	3
Avg sentence length	Google SERP	9.27	9.23
	Google RR	5.27	4.97
	Bing SERP *	12.52	12.09
	Bing RR *	5.12	4.41
	Gemini *	6.92	5.57
	GPT 3.5	18	18.25
Ratio of headings	Google SERP	0.29	0.28
	Google RR	0.06	0.06
	Bing SERP	0.24	0.26
	Bing RR *	0.06	0.05
	Gemini	0.11	0.1
	GPT 3.5	0	0
Ratio of list items	Google SERP	0	0
	Google RR *	0.48	0.32
	Bing SERP	0	0
	Bing RR *	0.46	0.36
	Gemini	0.38	0.4
	GPT 3.5	0	0
Ratio of paragraphs	Google SERP	0.71	0.72
	Google RR *	0.47	0.62
	Bing SERP	0.74	0.72
	Bing RR *	0.48	0.57
	Gemini	0.5	0.5
	GPT 3.5	1	1

Avg paragraph length	Google SERP	16.75	16.7
	Google RR *	29.78	21.47
	Bing SERP *	23.56	22.71
	Bing RR *	27.72	20.03
	Gemini	10	8.75
	GPT 3.5	47.5	56
Flesch reading ease score	Google SERP	58.86	63.91
	Google RR *	55.67	60.23
	Bing SERP *	50	60.18
	Bing RR *	52.68	61.27
	Gemini *	30.02	41.97
	GPT 3.5 *	31.89	47.79
Coleman-Liau readability index	Google SERP *	12.41	10.74
	Google RR *	12.9	11.5
	Bing SERP *	13.84	11.31
	Bing RR *	13.22	11.58
	Gemini *	17.38	14.83
	GPT 3.5 *	15.48	11.96
Avg concreteness	Google SERP	2.37	2.41
	Google RR *	2.39	2.43
	Bing SERP	2.33	2.34
	Bing RR *	2.38	2.43
	Gemini *	2.3	2.39
	GPT 3.5 *	2.29	2.37
Ratio of concrete words	Google SERP *	0.03	0.04
	Google RR *	0.03	0.04
	Bing SERP	0.04	0.04
	Bing RR	0.03	0.04
	Gemini	0.03	0.04
	GPT 3.5 *	0.04	0.05
Ratio of abstract words	Google SERP	0.46	0.43
	Google RR	0.42	0.39
	Bing SERP	0.55	0.55
	Bing RR	0.4	0.37
	Gemini	0.52	0.5
	GPT 3.5	0.69	0.71

Table F.1: Median values of accessibility indicators for domain-specific and general queries for all response groups.

* indicates that the distribution for the accessibility indicator was found to be significantly different ($p < 0.05$) for the response group's responses for the domain-specific and general queries/prompts using the Mann-Whitney U test for statistical significance.

G

Quantitative comparison of accessibility of response groups for original and reformulated queries & prompts

Accessibility Indicator	Response Group	Original Query/Prompt	Reformulated Query/Prompt
# sentences	Google SERP *	4	5
	Bing SERP *	4.29	5.11
	Google RR *	111.8	93.285
	Bing RR	137.29	140.5
	Gemini *	39	45
	GPT 3.5 *	3.5	6
Avg sentence length	Google SERP *	9.23	8.535
	Bing SERP *	12.17	11.445
	Google RR *	5.04	6.12
	Bing RR *	4.65	5.95
	Gemini *	5.715	7.295
	GPT 3.5	18	18.4
Ratio of headings	Google SERP *	0.28	0.22
	Bing SERP *	0.26	0.21
	Google RR *	0.06	0.04
	Bing RR	0.05	0.05
	Gemini *	0.1	0.12
	GPT 3.5	0	0
Ratio of list items	Google SERP *	0	0
	Bing SERP *	0	0.02
	Google RR *	0.345	0.21
	Bing RR	0.37	0.38
	Gemini *	0.4	0.37
	GPT 3.5	0	0
Ratio of paragraphs	Google SERP *	0.72	0.78
	Bing SERP *	0.73	0.76
	Google RR *	0.59	0.74
	Bing RR	0.56	0.57
	Gemini	0.5	0.51
	GPT 3.5	1	1

Avg paragraph length	Google SERP *	16.725	19.795
	Bing SERP *	23.06	24.9
	Google RR *	22.78	29.12
	Bing RR *	22.435	31.9
	Gemini *	9	22.8
	GPT 3.5 *	55	54.835
Flesch reading ease score	Google SERP *	61.765	78.135
	Bing SERP	58.08	60.12
	Google RR *	58.73	71.19
	Bing RR	59.375	58.35
	Gemini	38.45	36.3
	GPT 3.5	45.98	44.105
Coleman-Liau readability index	Google SERP *	10.995	6.71
	Bing SERP	11.86	11.555
	Google RR *	11.91	9.335
	Bing RR	11.88	11.8
	Gemini	15.565	15.58
	GPT 3.5	12.97	12.94
Avg concreteness	Google SERP *	2.4	2.54
	Bing SERP *	2.34	2.42
	Google RR *	2.42	2.41
	Bing RR	2.42	2.43
	Gemini *	2.37	2.33
	GPT 3.5	2.34	2.35
Ratio of concrete words	Google SERP	0.04	0.04
	Bing SERP *	0.04	0.04
	Google RR *	0.04	0.05
	Bing RR *	0.04	0.04
	Gemini	0.035	0.04
	GPT 3.5	0.05	0.05
Ratio of abstract words	Google SERP *	0.44	0.5
	Bing SERP	0.55	0.55
	Google RR *	0.4	0.525
	Bing RR *	0.385	0.465
	Gemini *	0.5	0.54
	GPT 3.5	0.7	0.71

Table G.1: Median values of accessibility indicators for responses collected from different response groups for original and reformed query/prompt.

* indicates that the distribution for the accessibility indicator was found to be significantly different ($p < 0.05$) for the response group's responses for the reformulated and original queries/prompts using the Mann-Whitney U test for statistical significance.