Generative Information Systems Are Great If You Can Read

Adam Roegiest
Triangle Lab
Canada
adam@roegiest.com

Zuzana Pinkosova
zuzana.pinkosova@unimib.it
Università degli Studi di Milano Bicocca
Italy

ABSTRACT

Generative models, especially in information systems like ChatGPT and Bing Chat, have become increasingly integral to our daily lives. Their significance lies in their potential to revolutionize how we access, process, and generate information [44]. However, a gap exists in ensuring these systems are accessible to all, especially considering the literacy challenges faced by a significant portion of the population in (but not limited to) English-speaking countries. This paper aims to investigate the "readability" of generative information systems and their accessibility barriers, particularly for those with literacy challenges. Using popular instruction fine-tuning datasets, we found that this training data could produce systems that generate at a college level, potentially excluding a large demographic. Our research methods involved analyzing the responses of popular Large Language Models (LLMs) and examining potential biases in how they can be trained. The key message is the urgent need for inclusivity in systems incorporating generative models, such as those studied by the Information Retrieval (IR) community. Our findings indicate that current generative systems might not be accessible to individuals with cognitive and literacy challenges, emphasizing the importance of ensuring that advancements in this field benefit everyone. By situating our research within the sphere of information seeking and retrieval, we underscore the essential role of these technologies in augmenting accessibility and efficiency of information access, thereby broadening their reach and enhancing user engagement.

ACM Reference Format:
Adam Roegiest and Zuzana Pinkosova. 2024. Generative Information Systems Are Great If You Can Read. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10–14, 2024, Sheffield, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3627508.3638345

1 INTRODUCTION

In recent years, generative Large Language Models (LLMs) have notably impacted various aspects of our lives, from how we search (e.g., Bing Chat[1], Bard[2]) code (e.g., Github Copilot[3]) write (e.g., Jasper.ai[4], Copy.ai[5]) and even conduct the research (e.g., retrieval augmented generation [8, 49, 54], systematic reviews [111, 112], relevance assessment [40, 106]). At the same time, there is an increased emphasis on improving these systems' trustworthiness [33, 69, 88], reducing bias [41], and enhancing their usefulness for end users [9, 46].

Despite these advancements, generative information systems have seen notable failures [67] with some leading to concerning outcomes [29]. While not all negative outcomes will be so serious, they may still negatively affect large portions of society. Accordingly, this paper addresses the oft-neglected issue of accessibility in generative systems for individuals with lower literacy levels, framing it within the realm of information seeking and retrieval. The underlying hypothesis, subsequently supported in our findings, is that current development trends in IR may unintentionally focus mainly on users with standard linguistic abilities. This raises significant concerns about inclusivity and equitable access, underscoring the necessity for a more inclusive approach in the design and development of generative systems.

The Organisation for Economic Co-operation and Development (OECD)'s Programme for the International Assessment of Adult Competencies (PIAAC) reveals concerning literacy rates in countries including the United States of America (USA), Canada, the United Kingdom (UK), Ireland, and Australia [107], indicating that n average of 3-4% of the surveyed population could be classified as "functionally illiterate," with scores beneath Level 1 (Figure 1). Additionally, the percentage of participants who scored at or below Level 2 ranged from 39% in Australia to 55% in Ireland. According to Figure 1, individuals with scores at or beneath Level 2 usually face challenges when engaging with extensive texts or when required to perform detailed multi-step reasoning based on the text [107].[6].

Considering the UK Government's and WCAG's recommendations for elementary-level writing [99, 110], there is a clear necessity to design generative information systems that cater to these individuals. Such design would ensure individuals facing challenges in reading, like those with low literacy or cognitive impairments, are not excluded from the benefits these systems offer. [7] Here, accessibility encompasses two facets: formulating queries (or prompts) and comprehending the system's response. Either of these could pose difficulties for those with low literacy. In this paper, we choose to

CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom

ACM ISBN 979-8-4007-0434-5/24/03
https://doi.org/10.1145/3627508.3638345

[1] https://www.microsoft.com/en-us/edge/features/bing-chat
[2] https://bard.google.com/
[3] https://github.com/features/copilot
[4] https://www.jasper.ai/
[5] https://www.copy.ai/
[6] For additional context, in the USA, 52% of respondents scored at Level 2 or lower [91], corroborating other studies suggesting that the mean literacy level for American adults falls between the 6th and 8th grades [32, 64, 102]
[7] A more holistic solution would be to enhance the availability and quality of *lifelong* education for everyone, recognizing, however, that this calls for profound societal transformation.

concentrate predominantly on the comprehension aspect of system responses, given its comparative ease of analysis.

This work focuses on generated response comprehension, given its relative ease of analysis for those with low literacy. We begin by outlining criteria for text readability and explore its impact on users with low literacy and cognitive impairments (Section 2). Despite the PIAAC results revealing significant reading challenges in countries known for high research outputs, there are research gaps in the fields of IR and information science in understanding how these difficulties affect users' information retrieval abilities (Section 2.2). This oversight may be because many IR and IS researchers do not specialize in accessibility or cognitive sciences. In highlighting these accessibility issues, our goal is to further bring attention and consideration to the IR community. We argue that enhancing accessibility is not just ethically crucial but also essential for improving the effectiveness and reach of generative information systems. Moreover, we question whether genuine progress in the field can be achieved without ensuring universal accessibility.

This study focuses on the training methodologies of generative models and their baseline response generation, analyzing the readability of "ideal" responses from three prominent instruction fine-tuning datasets: Alpaca [104], Dolly [27], and self-instruct [113]. Our findings indicate that these responses typically cater to a collegiate audience or feature complex, technical language. However, these datasets may not fully represent the training data of well-known LLMs. We extended our analysis to five LLMs, including GPT-3.5-Turbo, GPT-4, PaLM2, Llama2-7b, and Falcon-7b, using the Alpaca dataset. All models, without specific prompts, consistently produced responses akin to college-level writing, with some showing tendencies towards lower grade-level responses but still maintaining a high level of sophistication. This highlights the importance of thoughtful integration of these models into information systems, considering their response characteristics and audience.[8]

Generative model training often involves reinforcement learning with human interaction feedback. Thus, it is essential to carefully examine the sources of this feedback and the methods by which it is collected. More specifically, providing a clear target audience during feedback solicitation may help to mitigate a likely bias towards the preferences of highly literate individuals. The readability and adaptability of these models should also consider users who communicate in variants of English such as dialects, creoles, or pidgins, rather than just the "standard" version as taught in schools. An illustrative case study involving Jamaican Patois revealed that models often default to standard English without specific prompts, potentially indicating a misalignment with the user's intent. Such disparity between the user's question and the model's answer can lead to a compromised user experience and may drive users away from interacting with such a system. Finally, based on the results presented in the paper, we discuss strategies and research for developing inclusive systems that cater to diverse users, including those with low literacy, ensuring equitable information access. This approach is vital for advancing the field of IR and achieving greater impact within the broader community.

[8]All experimental results and scripts are available at https://github.com/aroegies/llm-readability.

Level 1
Most of the tasks at this level require the respondent to read relatively short digital or print continuous, non-continuous, or mixed texts to locate a single piece of information that is identical to or synonymous with the information given in the question or directive. Some tasks, such as those involving non-continuous texts, may require the respondent to enter personal information onto a document. Little, if any, competing information is present. Some tasks may require simple cycling through more than one piece of information. Knowledge and skill in recognizing basic vocabulary, determining the meaning of sentences, and reading paragraphs of text are expected.
Level 2
At this level, the medium of texts may be digital or printed, and texts may comprise continuous, non-continuous, or mixed types. Tasks at this level require respondents to make matches between the text and information and may require paraphrasing or low-level inferences. Some competing pieces of information may be present. Some tasks require the respondent to: cycle through or integrate two or more pieces of information based on criteria; compare and contrast or reason about the information requested in the question; navigate within digital texts to access and identify information from various parts of a document.

Figure 1: Descriptions of Level 1 and Level 2 according to the OECD's PIAAC study and reproduced from Canada's PIAAC report [107].

2 BACKGROUND

Reading comprehension is the capacity to understand, apply, reflect upon, and engage with written content to achieve personal goals, enhance one's knowledge and skills, and actively contribute to societal activities [82]. Accordingly, it forms a foundation for how users interact with information systems and even more so with the incorporation of generative models into those systems. For the IR community, this aspect highlights the essential need for these sophisticated systems to be accessible and understandable to a broad spectrum of users, encompassing various literacy levels. In doing so, advancements in technology are not just innovative but also inclusive and user-friendly.

2.1 Readability Metrics and IR

Assessing readability in IR contexts is crucial, as it directly impacts how users interact with and comprehend information presented by generative models. The complexity of subjective readability perception, affected by a variety of factors such as language, legibility, and syntactic and semantic complexities [10, 59] poses a challenge across domains, including IR. Our study's focus on U.S. readability metrics is particularly relevant for the IR community due to: (i) the U.S. government's emphasis on plain language in official documents [1, 2, 79] and the trend of U.S. newspapers targeting a ninth-grade reading level [58], mirroring the need for clear information presentation in IR systems; (ii) the leadership of U.S.-based organizations in artificial intelligence (AI) development, with key contributions from Google's PaLM2/Bard [6], Facebook's Llama2 [109], and OpenAI's GPT models [81, 83], indicating the potential global impact of U.S. readability standards; and (iii) the extensive U.S. readability metrics aligned with its grade levels [26, 30, 43, 61, 75, 98], offering a robust framework for text clarity assessment.

Readability metrics mainly consist of two elements: the complexity of the words used and the length of the content [68]. These elements are weighted differently depending on the specific metric producing readability scores. Readability scores can often be translated into an estimate of the "years of education required to understand the content", which can provide a general idea of the reading grade level [28]. In this work, we are primarily concerned with directional accuracy (i.e., a rough estimate of how well generative models write). Therefore, we categorize responses into four buckets based on the estimated grade level or years of education: ≤ 6 years, $7-9$ years, $10-12$ years, >12 years (i.e., college-level writing), which mirrors that of WCAG [110]. To balance out the different possible ways to formulate the readability components, our work relies on three measures provided by the *py-readability-metrics* package.[9] We provide brief descriptions of these measures to inform subsequent discussions. It is important to mention that the library does not evaluate content shorter than 100 words.

2.1.1 Flesch-Kincaid[61]. The Flesch-Kincaid metric evolved from the earlier Flesch reading ease score [43], which has been used to regulate the readability of insurance policies in several U.S. states [1, 2]. The Flesch-Kincaid metric calculates readability using two components: the average words per sentence for length, and the average syllables per word for difficulty, with more emphasis placed on the latter. While the metric lacks a definitive upper limit for its score (i.e., years of education), in practice, exceedingly high scores are atypical unless the text is unusually lengthy or adversarial.

2.1.2 Coleman-Liau[26]. Developed by Coleman and Liau, this metric employs a straightforward approach to gauge word complexity by counting characters, rather than relying on syllable counts or other metrics. This method bears resemblance to the Automated Readability Index [98]. Coleman-Liau uses the average number of sentences per 100 words as the length component and the average number of characters per 100 words as its difficulty component, as opposed to the approach of the Automated Readability Index which uses aggregate counts of both.

2.1.3 Dale-Chall[30]. Edgar Dale and Jeanne Chall introduced a method to streamline the assessment of word complexity by utilizing a list of words recognized by a majority of fourth-grade students. The initial 1948 list contained 763 words familiar to 80% of these students [30], and a subsequent 1995 revision expanded this to 3,000 words [22]. The measure calculates the average words per sentence for length, while word difficulty is determined by the percentage of words in a sentence not present in the word list. However, it is important to note that due to the inherent nature of these lists, technical texts may receive higher scores, potentially leading to a bias in model responses to technical inquiries.

[9]https://github.com/cdimascio/py-readability-metrics

2.2 Reading Comprehension, Cognitive Impairment, and Literacy

Individuals with cognitive impairment[10] often face challenges in processing online text due to impacted cognitive functions like language, memory, attention, and executive functions. [73, 100, 105]. In terms of literacy, illiterate people often struggle to process difficult texts due to their limited reading abilities and lack of familiarity with words and concepts [114]. The lack of basic reading and writing skills can limit their ability to comprehend complex information and navigate through written texts [85]. Additionally, illiteracy can lead to functional and cognitive alterations, which may further hinder their cognitive processing abilities [86]. By considering the cognitive processing underpinning literacy limitations and cognitive impairment, we can better tailor responses from generative information systems. Readability metrics, such as those in Section 2.1, are widely recognized for predicting text difficulty, rooted in general text comprehension and processing [4]. Studies have explored the use of these metrics to address text complexity (e.g., Flesch-Kincaid and Coleman-Liau indices), word complexity (e.g., Dale-Chall formula), and text length [15]. While we focus on literacy in this work due to easily accessed measures, addressing all aspects of reading comprehension (e.g., impairments) is vital for the IR community. In doing so, the development of more inclusive and user-friendly systems, considering the full spectrum of user abilities and needs, can be facilitated. Addressing text complexity with readability metrics is a key step in aligning advanced language technologies with the varied cognitive needs of users.

2.3 IR, Cognitive Impairments, and Literacy

Recent advancements in generative LLMs have substantially improved natural language processing, leading to more coherent, contextually appropriate, and, to some extent, readable texts [62, 89]. These advances also enhance the model's adaptability to user needs in information-seeking tasks[77]. When integrated with IR systems [3, 8, 49, 54] and methodologies [40, 106, 111, 112], they hold the promise of elevating a user's information discovery experience. Such experience improvements, however, should be not be limited to those with neurotypical abilities, higher levels of education, or other factors that might disqualify a user from benefiting from these improvements. Such an approach aligns with one of the IR field's goals of effective and user-friendly information access.

Despite ongoing research on information-seeking within the IR context [17, 18, 52], the evaluation of generative systems in aiding individuals with learning disabilities remains under-explored, marking a novel area of contribution for this work. This gap becomes increasingly relevant with the shift towards conversational search and the integration of generative models in IR [3, 49, 54]. Cognitive impairments add complexity to the search process, affecting aspects such as keyword creation, search refinement tool usage, and information credibility assessment [24, 84]. Individuals with impairments face unique challenges that are tied to the tasks of reading and writing in addition to difficulty in rapid automatized naming and reduced short-term memory capacity [25, 63].

[10]We follow Berget and MacFarlane [17] and use cognitive impairment to include learning impairments (e.g., dyslexia,) as well as other conditions like autism spectrum disorder and aphasia.

These challenges could potentially influence their interaction and engagement with generative information systems, in particular, if the content accessibility and readability do not sufficiently support their needs (e.g., word choice and length, layout). Furthermore, the applicability of existing models to users with impairments might be limited [17, 18] as generative systems become commonplace. To address these issues, there is a need to apply readability indices and personalization in the design of generative system interfaces, ensuring content accessibility and alignment with user comprehension levels, thereby facilitating a more intuitive and effective information-seeking process.

Literacy significantly impacts the interaction with information, shaping competencies in navigation, evaluation, and utilization. The correlation between text readability and a user's literacy level presents a challenge that generative models must overcome to guarantee equitable information access for all users. Despite adaptive efforts by generative models, Murgia et al. [77] highlights continuing challenges in aligning individual literacy and readability standards [56]. In particular, Murgia et al. found that children often stumble upon understanding unfamiliar terms and specific lexicons within responses generated by systems like ChatGPT. Furthermore, it is important to highlight that despite efforts to adjust the content to age or grade-based literacy levels, the discrepancies in text comprehension and search aid required persist [51]. Neglecting these literacy challenges risks marginalizing users with lower literacy, exacerbating the digital divide. Therefore, this research emphasizes the importance of developing literacy-adapted systems in IR to bridge this gap and promote a fair digital environment. This approach continues work done by others [16, 18] but highlights the novel aspect that naïve incorporation of generative models into information systems will invariably yield disparity in usability and user experience for those that struggle with literacy.

2.4 Fairness, Transparency, Safety and Other Concepts

The primary focus of this paper is to shed light on accessibility concerns for generative models and the systems that incorporate them as they manifest for users with low literacy and cognitive impairments. While it is beyond is beyond the scope of this work to produce a summary of work, it is important to highlight the multitude of research that has been undertaken to understand and reduce bias [14, 31, 38, 45, 47, 116], improve helpfulness, safety and information quality [9, 11, 12, 37, 46, 65, 120], and increase transparency [7, 13, 50, 70, 76, 87, 93] across the IR, machine learning, and associated communities. Despite the research efforts, there is an often overlooked set of interconnected factors: the cognitive abilities, literacy and education level of users. If we truly wish to build fair, accountable, transparent, safe and *accessible* systems then it is imperative to address the biases and challenges faced by users with low literacy and cognitive impairments. By not doing so, we lessen the impact of all of the work we do by limiting its ability to impact and positively change the lives of people who currently may find systems to be inaccessible or outside of their abilities.

3 GENERATIVE LLM TRAINING DATA

The internet has served as a foundational platform for the pre-training of generative models [90]. This pre-training can span a wide range of writing levels, styles, and language use. However, it is reasonable to assume that many models have been pre-trained on curated collections, such as reputable news outlets and Wikipedia, which typically offer higher quality writing [81, 95, 117]. Further, this pre-training may teach models how to generate text but not necessarily how to appropriately generate a response. In contrast, instruction fine-tuning [34, 55, 115] has played a pivotal role in the broader adoption of these models, exemplified by the success of ChatGPT [80]. This is largely due to the reduced training overhead when building upon pre-trained foundation models. While the initial pre-training phase shapes a model's text generation capabilities, we argue that instruction fine-tuning provides a more targeted approach to modulating a model's writing style and level. In this section, we apply the readability metrics (Section 2.1) to assess instruction fine-tuning datasets. Our aim is to understand how these "benchmark" responses might influence the outputs of models trained on them, potentially posing challenges for users with lower literacy levels and/or cognitive impairments.

3.1 Datasets

This section provides an overview of the instruction fine-tuning datasets of interest and details any specific instructions used to generate the responses either by a generative model or a human (e.g., response length, writing style).

3.1.1 Self-instruct [113]. The self-instruct dataset comprises 82,612 instruction-response pairs. These were produced using a bootstrapping method with GPT-3 ("davinci"). Starting with a foundational set of 175 human-curated tasks, GPT-3 was employed to create additional instructions and subsequently generate outputs based on those instructions. Instead of resorting to prompt engineering, the dataset creators adjusted various hyperparameters of GPT-3 for each phase. For instance, during the instruction generation phase, GPT-3 was configured with a token limit of 1,024 and a temperature setting of 0.7. However, for producing outputs corresponding to these instructions, the token limit was set to 300 with a temperature of 0. Filtering was performed at various stages to ensure the diversity of instructions and their responses. Finally, the base model had the freedom to generate instructions and responses within these defined parameters but had no other constraints.

3.1.2 Alpaca [104]. The Standford Alpaca dataset comprises 52,002 instruction-response pairs, derived from the self-instruct dataset. This dataset differentiates itself by utilizing *text-davinci-003* rather than *davinci*. Additionally, it produces only one response per instruction and incorporates an initial prompt to guide the model's generation process. For this work, the prompt directs the model to ensure diversity in the instructions it creates, to keep the instruction to a maximum of two sentences, and to limit the response to less than 100 words. The prompt biases the model to attempt to produce short pieces of text despite being given an explicit maximum token length of 3,072 in the API call to OpenAI. Interestingly, this is the only dataset out of the three examined to constrain the responses to achieve "more readable" content.

	U.S. Grade Level			
Method	≤ 6	7 − 9	10 − 12	> 12
Flesch-Kincaid	714	1330	1316	1867
Coleman-Liau	639	1385	1377	1826
Dale-Chall	151	1124	848	3104

Table 1: Readability scores by U.S. grade-level bucket for the Alpaca instruction fine-tuning dataset[104] for the 5,227 instances out of 52,0002 that were long enough to be scored with an average length of 159.9 words.

	U.S. Grade Level			
Method	≤ 6	7 − 9	10 − 12	> 12
Flesch-Kincaid	258	648	787	773
Coleman-Liau	203	730	943	590
Dale-Chall	39	573	456	1398

Table 2: Readability scores by U.S. grade-level bucket for the Dolly instruction dataset[27] for the 2,466 instances out of 15,011 that were long enough to be scored with an average length of 234.4 words.

3.1.3 Dolly [27]. The Dolly dataset, *databricks-dolly-15k*, consists of 15,011 instructions generated by employees of Databricks in March and April 2023. Both the instruction and ideal response are generated by humans with no machine involvement. There may be additional context provided as input in the form of Wikipedia passages for certain instruction types (e.g., summarization and closed-book question answering). The authors note that while deliberate obscenities, intellectual property, or private individual information should be absent, the dataset might still reflect inherent Wikipedia biases and the personal inclinations of Databricks employees.

3.2 Dataset Readability

Across all three datasets (Tables 1, 2, and 3), we see that very few instances can be scored for readability due to the length requirement of the underlying library. While shorter texts may aid in readability, there is no guarantee that a short piece of text will be inherently more readable than a long one which we discuss in the following section.

For both the Alpaca and Dolly datasets, there is a consistent trend that all readability measures are skewed towards higher educational levels. This suggests that training on these datasets will likely yield a model that is capable of generating text at a college level. We acknowledge that just because a model could generate such text does not mean that it will. But given how well generative models have done on a variety of college-level tasks (e.g., bar exams [60, 74], GRE and SAT [81], MCAT [19]), the skew is not unexpected. The surprising aspect of this is that the instructions within these datasets do not explicitly specify any grade level for the response (e.g., "explain like I'm 5..."). Instead, the system is permitted to derive an appropriate response based on the different "ideal" responses to potentially similar types of instructions (e.g., "explain this", "tell me a story"). While this freedom allows the model to tailor a response according to its internal statistical model and any internal measures of "goodness" (e.g., creativity, accuracy), the same freedom means that a model may not naturally respond in an accessible manner (e.g., generating a college-level answer to a fifth-grader's question).

	U.S. Grade Level			
Method	≤ 6	7 − 9	10 − 12	> 12
Flesch-Kincaid	1097	547	211	283
Coleman-Liau	1171	554	269	144
Dale-Chall	226	800	392	720

Table 3: Readability scores by U.S. grade-level bucket for the self-instruct instruction dataset[113] for the 2,138 instances out of 82,612 that were long enough to be scored with an average length of 165.7 words.

For Dolly and Alpaca, the Dale-Chall metric reports higher grade-level responses than the other two measures, especially at the college level. While some of this is invariably influenced by the presence of technical or scientific words in the instructions (e.g., "Describe the process of osmosis" in Alpaca, "What is radioactive decay?" in Dolly), we argue that responses to such prompts should not consistently be of college-level complexity. However, their prevalence in the training data might predispose the model towards such outputs. Essentially, while the Dale-Chall metric might overestimate response difficulty due to its emphasis on words familiar to fourth-graders, it underscores the potential accessibility bias arising from the frequent use of challenging terms. This raises the possibility that models could benefit from datasets incorporating multiple "ideal" responses, each varying in complexity but retaining accuracy and other desired attributes.

The self-instruct dataset stands out due to its limited number of scorable responses and its skew towards lower grade levels compared to Alpaca and Dolly. This could be attributed to the utilization of GPT-3 (*davinci*), which is arguably "simpler" and less extensively trained than *text-davinci-003*. Additionally, it might not possess the same level of knowledge as an average Databricks employee. However, the Dale-Chall metric still categorizes the dataset as having an excessive use of "difficult" words in its responses, leading to a bias toward higher grade levels. Yet, this bias is less pronounced in the self-instruct dataset than in Alpaca and Dolly, especially when comparing the proportions within the ≤ 6 and > 12 grade level brackets. This provides further evidence that the "davinci" model may have tended to produce somewhat easier-to-understand responses which could easily be overlooked by a preference for newer, more advanced models.

3.3 "Short" Responses

Despite *py-readability-metrics* requiring 100 words to produce a reliable score, it is important to note that these shorter responses are still important. Therefore, the base components from the readability metrics in Section 2.1 are used to conduct a high-level analysis. From Table 4, it can be seen that a substantial segment of responses, irrespective of the dataset, consistently includes a sizable proportion of words not found in the Dale-Chall word list. Given that these responses typically range between 17-33 words, it is plausible to assume they may pose comprehension challenges for individuals

with stopwords				
Dataset	Word Count	Complex Words	Characters	Syllables
alpaca	33.94 (28.87)	0.39 (0.25)	4.95 (1.90)	1.66 (0.56)
Dolly	31.63 (26.12)	0.42 (0.24)	5.01 (1.06)	1.66 (0.41)
self-instruct	17.70 (21.40)	0.55 (0.32)	6.21 (2.70)	1.80 (0.59)
without stopwords				
	U.S. Grade Level			
Dataset	Word Count	Complex Words	Characters	Syllables
Alpaca	20.62 (17.24)	0.56 (0.26)	6.17 (2.05)	2.00 (0.64)
Dolly	19.76 (15.27)	0.59 (0.24)	6.08 (1.19)	1.95 (0.48)
self-instruct	10.77 (12.17)	0.64 (0.29)	7.24 (2.74)	2.02 (0.60)

Table 4: For the responses that were too short to be reliably scored, we report the average and standard deviation of the following four measures across the desired outputs with and without stopwords: average word count, average fraction of Dale-Chall complex words, average character count per word, average syllable count per word.

with literacy or cognitive challenges. Interestingly, words, on average, neither exhibit extensive length across datasets nor seem to contain many syllables. This suggests that the terminology, possibly unfamiliar to some readers, may also be relatively short or potentially specialized (e.g., code, domain-specific vernacular). However, when adjusting for stopwords (i.e., typically short words that could skew character and syllable count) by excluding them from our assessment, all average values (except for word count) noticeably increase. Accordingly, relying on the same signals as readability metrics when considering shorter model responses may underestimate reading difficulty due to the presence of stopwords and we remind system builders to be mindful of Mark Twain's sage advice to "[not] use a five dollar word when a fifty cent one will do" since short words do not guarantee comprehension.

4 GENERATIVE MODELS

In the previous section, we observed that when fine-tuned with instruction datasets, generative models have the potential to learn to produce college-level text, which might be challenging for those with lower literacy levels or cognitive impairment. The question remains, however, as to whether generative models exhibit these traits. Therefore, to provide an answer, this section turns towards five modern state-of-the-art models described below. Using the 52,002 Alpaca instructions without any added context (i.e., no engineered prompts), we aim to determine if these models generate text at a college-level proficiency.

4.1 Models

For this analysis, we have identified five models representative of the current state-of-the-art and anticipated for deployment in practical generative information systems. We focus on the following models that are primarily used for text generation[11]:

- *GPT-3.5-Turbo* [83]: Given that this model is the successor to the models used to generate the self-instruct and Alpaca datasets and considering its role in popularizing ChatGPT, we deemed it a suitable baseline due to its exposure to the general public. We used the August 2023 version ("gpt-3.5-turbo") using the default settings provided in the OpenAI API, such as a temperature of 1 and no token length limit.
- *GPT-4* [81]: GPT-4 is the subsequent successor to the above LLM and is OpenAI's general-purpose LLM that powers ChatGPT Plus and provides a basis for Bing Chat [12]. As this is arguably OpenAI's most powerful model,we included it to observe any changes in response generation over the "last" generation. We used the September 2023 version of the model ("gpt-4") via the OpenAI API with all default settings, such as a temperature of 1 and no token length constraints.
- *PaLM2* [6]: Google's generative model that helps to power the Bard retrieval augmented generation product [13]. Although it may not have attained the same level of popularity as OpenAI's models, due to its potential integration into prominent Google products we included it in our tests. We used the August 2023 version of the model ("chat-bison@0001") with all default settings via the Vertex API.
- *Llama2-7B* [109]: This is the open-source, commercially viable successor to Meta's Llama model [108]. We opted to assess this model given its potential as a robust open-source option suitable for production deployment. Llama2-7B contrasts with the larger models previously mentioned, possibly offering unique insights due to its more compact size. The version assessed was sourced from the HuggingFace in September 2023 ("meta-llama/Llama-2-7b-chat-hf") via their Inference Endpoints with an A10G instance.
- *Falcon-7B* [5]: Provided by the Technology Innovation Institute (TII), this model represents one of the first commercially viable and competitive alternatives to models produced by US-based corporations. As TII is not U.S.-based, we believe this may also affect the generation style due to potential organizational structure and culture differences (e.g., English is not necessarily a first language). We use the smaller version of the model to investigate how model size affects the results. We used the version of Falcon-7B available in September 2023 on HuggingFace ("tiiuae/falcon-7b-instruct") via their Inference Endpoints with an A10G instance.

	U.S. Grade Level			
Method	≤ 6	$7-9$	$10-12$	>12
Flesch-Kincaid	1450	4037	8476	10,676
Coleman-Liau	1217	2459	5506	15,457
Dale-Chall	10	1263	2665	20,701

Table 5: Readability scores by U.S. grade-level bucket for the Alpaca dataset [104] with outputs regenerated by GPT-3.5-Turbo[83] for the 24,639 instances out of 52,002 that were long enough to be scored with an average length of 341 words.

4.2 Results

4.2.1 GPT-3.5-Turbo. Prior experiences with GPT-3.5-Turbo led us to expect a certain degree of verbosity, the data presented in Table

[11] Specialized models, like those for code completion, will likely generate text differently but that is out of the scope of the current work

[12] https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

[13] https://ai.google/discover/palm2/

	U.S. Grade Level			
Method	≤ 6	$7-9$	$10-12$	>12
Flesch-Kincaid	2229	5339	7866	4936
Coleman-Liau	1668	3039	6496	9167
Dale-Chall	16	1505	2929	15,920

Table 6: Readability scores by U.S. grade-level bucket for the Alpaca dataset [104] with outputs regenerated by GPT-4[81] for the 20,370 instances out of 52,002 that were long enough to be scored with an average length of 303.7 tokens.

	U.S. Grade Level			
Method	≤ 6	$7-9$	$10-12$	>12
Flesch-Kincaid	1113	1093	1005	577
Coleman-Liau	915	1139	1056	678
Dale-Chall	573	1085	749	1381

Table 7: Readability scores by U.S. grade-level bucket for the Alpaca dataset [104] with outputs regenerated by PaLM2[6] for the 3,788 instances out of 52,002 that were long enough to be scored with an average length of 120.2 words.

	U.S. Grade Level			
Method	≤ 6	$7-9$	$10-12$	>12
Flesch-Kincaid	1089	1390	1203	1869
Coleman-Liau	1105	1195	1301	1950
Dale-Chall	114	909	643	3885

Table 8: Readability scores by U.S. grade-level bucket for the Alpaca dataset [104] with outputs regenerated by Falcon-7B [5] for the 5551 instances out of 52,002 that were long enough to be scored with an average length of 473.2 words.

5 indicates the model is more verbose than anticipated. With nearly half of the dataset can be scored and the majority aligning with higher grade levels, this highlights a need for guardrails tailored to a model's use (e.g., a prompt to produce easier to read responses). This tendency towards lengthier outputs, averaging 341 words, is interesting given that many of the responses for the training datasets were shorter. It is important to note that extended text poses challenges for those with limited literacy [92, 97] but also for those with impairments, such as dyslexia [20, 103].

That being said, the disparity between Dale-Chall and the other two metrics in the > 12 bucket also indicates that the vocabulary employed might be more technical or challenging, posing comprehension issues for those with linguistic challenges. Given both the verbosity and complexity of the words, it is advisable that this model should be carefully incorporated into information systems. It may be the case that injecting additional instructions to adapt responses for users with lower literacy levels is necessary.

4.2.2 GPT-4. As seen in Table 6, GPT-4 generates a higher number of lengthy responses compared to the fine-tuning data but these are less frequent and shorter than those from GPT-3.5-Turbo. Interestingly, there is a marked shift toward outputs aligned with fewer years of education in GPT-4, in comparison to GPT-3.5-Turbo with a much smaller ratio between the ≤ 6 years and > 12 years buckets for Flesch-Kincaid and Coleman-Liau metrics. However, GPT-4's performance on the Dale-Chall measure remains relatively consistent with its predecessor model, showing only a decrease in the number of scoreable responses. While some progress is noticeable concerning response and word length, there is a persistent inclination to use more "difficult" vocabulary. This bias could stem from the specialized nature of Alpaca instructions, which may often be technical or science-centric. This improvement comes at a substantial cost as GPT-4 took several days longer than GPT-3.5-Turbo and was approximately 25x more expensive ($\tilde{\$}20$ for GPT-3.5-Turbo to $\tilde{\$}500$ for GPT-4 in response generation). With that in mind, it may be more effective and efficient to attempt to improve GPT-3.5-Turbo's responses rather than GPT-4's responses.

4.2.3 PaLM2. In contrast to the GPT models, PaLM2 exhibits attributes more closely aligned with our observations from the fine-tuning data, as evidenced by Table 7. Approximately 7% of the responses generated by PaLM2 were scoreable. Moreover, in comparison with the Alpaca and Dolly datasets, PaLM2 demonstrates a notable skew towards responses indicative of fewer years of education. Interestingly, even the Dale-Chall metric for noticeably less when compared to the fine-tuning data, despite this metric's predisposition to register higher grade levels when confronted with technical terms. For context, the self-instruct data presented approximately $\tilde{3}$ times more instances in the > 12 bucket than in the ≤ 6 bucket according to the Dale-Chall metric. In contrast, PaLM2 displays a ratio of only about $\tilde{2}.4$ times more instances. While we cannot make any claims as to whether this behavior is intentional, due to PaLM2's closed nature, it is promising that the readability of responses may have been factored into the training of PaLM2 at some point[14]. Our interpretation of these findings, coupled with subsequent sections, suggests that PaLM2 might serve as a reasonable base model for incorporation into a generative information system without requiring as much consideration around prompt engineering or hyperparameters compared to other models.

4.2.4 Falcon-7B. Table 8 shows that Falcon-7B manages to produce a fairly balanced distribution for both Flesch-Kincaid and Coleman-Liau despite having an average length exceeding that of either GPT model[15]. The largest bucket remains the > 12 category, which is anticipated. On the other hand, the model's relative performance between the ≤ 6 and > 12 buckets is commendable. This implies that response length is not the sole determinant of reading difficulty, which mirrors our observations in Section 3.3. Despite Falcon-7B's smaller size compared to commercial models, its performance surpasses expectations, hinting that a more compact model might be preferable for optimizing accessibility. Further studies involving larger Falcon versions are necessary to determine and validate such a claim but that study is beyond the scope of this work.

4.2.5 Llama2-7B. Table 9 repeats the trend of GPT-3.5-Turbo and GPT-4 characterized by numerous long responses with a trend towards more years of education[16]. However, the ratio between lower

[14] As Bard does contain a "Simplify" button, we suspect that some thought may have been given to this topic but caution that a single button may not be a panacea.
[15] Upon reviewing the responses, we observed that the model sometimes repetitively generates the same set of words multiple times, which likely contributes to the overall length of the output.
[16] It is worth mentioning that, akin to Falcon-7B, repetition might also be influencing the length component of these results, potentially due to model size.

	U.S. Grade Level			
Method	≤ 6	7 − 9	10 − 12	> 12
Flesch-Kincaid	3385	6968	9222	10,771
Coleman-Liau	5842	5991	7798	10,715
Dale-Chall	286	3373	3975	22,712

Table 9: Readability scores by U.S. grade-level bucket for the Alpaca dataset [104] with outputs regenerated by Llama2-7B [109] for the 30,346 instances out of 52,002 that were long enough to be scored with an average length of 465.4 words.

GPT-3.5-Turbo

	U.S. Grade Level			
Method	≤ 6	7 − 9	10 − 12	> 12
Flesch-Kincaid	[1478, 1432]	[4067, 4166]	[8336, 8322]	[10,714, 10,722]
Coleman-Liau	[1236, 1223]	[2474, 2449]	[5519, 5547]	[15,434, 15,432]
Dale-Chall	[13, 9]	[1296, 1304]	[2636, 2700]	[20,708, 20,629]

PaLM2

	U.S. Grade Level			
Method	≤ 6	7 − 9	10 − 12	> 12
Flesch-Kincaid	[1112, 1125]	[1085, 1031]	[1000, 1021]	[566, 580]
Coleman-Liau	[918, 914]	[692, 1118]	[1093, 1006]	[692, 719]
Dale-Chall	[553, 448]	[1101, 1122]	[726, 723]	[1383, 1364]

Table 10: Readability scores by U.S. grade-level bucket for the Alpaca dataset [104] with outputs regenerated twice by GPT-3.5-Turbo[83] (24,663 and 24,642 instances) and by PaLM2 [6] (3763 and 3757 instances) that were long enough to be scored.

and higher education levels seems to be less skewed for Llama2-7B compared to either OpenAI model. The Llama2-7B and Falcon-7B results reinforce that the content of responses is just as crucial for readability as their length. Relying solely on length restrictions may not be sufficient to ensure comprehensible responses.

The findings of this section provide additional evidence suggesting that most ready-made generative models tend to generate responses akin to responses that a college-educated individual might have plausibly produced. This reinforces the argument that these models may not be universally suitable for broad applications without careful consideration regarding their intended audiences and user experience design. Based upon these findings, we discuss additional considerations in Section 7.

4.3 Readability Consistency for LLM Responses

Since we have not constrained response generation, we wish to determine if the initial results were a simple non-deterministic fluke. To test this, we regenerated the Alpaca responses twice more using GPT-3.5-Turbo and PaLM2, as they are both the most time and cost-effective models in previous experiments, using exactly the same configurations as the initial experiments (i.e., to test variability in response generation). The outcomes, presented in Table 10, reveal a remarkable consistency in the models' behavior. Although the specific scored responses and their associated grade-level categories varied, the overarching trends remained stable. This consistency, achieved without additional constraints, suggests that such response generation tendencies are intrinsic to the models rather than a single, atypical execution.

Using the PaLM2 results, we conducted further analysis of instructions with scored prompts and found 1,313 instructions are shared by all three runs, 1,400 are shared by two runs, and 4,569 are unique to one of the runs. The distribution of task words in these instructions is depicted in Table 1, which also highlights instructions that consistently yielded short responses across runs. Interestingly, the nature of the generative task does not significantly influence the category of response generated. When multiple models produce responses to the same instruction, approximately half the instances show variations in the evaluated grade level of one model's response compared to another, irrespective of the readability metric applied. This suggests that specific details of the task may have a more pronounced influence on readability than the task itself (e.g., tell a horror story versus a comedy story). Such observations underscore the idea that relying solely on the model's inherent behavior to consistently produce readable responses might be insufficient and proactive measures to ensure readability are essential (e.g., utilizing an inferred user reading level or user-selected reading level to modify the model prompt).

Shared by Three	Shared by Two	Unique	Too Short
Write (436)	Write (227)	Describe (526)	Generate (3899)
Generate (280)	Generate (160)	Explain (523)	Create (3213)
Create (152)	Explain (141)	Generate (348)	Describe (2431)
Compose (60)	Create (128)	Create (302)	Given (2101)
Explain (32)	Describe (84)	What (279)	What (1948)
Compare (26)	Compare (65)	Write (256)	Write (1839)
Summarize (25)	Summarize (53)	List (160)	Name (1786)
Tell (25)	What (41)	Name (159)	Identify (1373)
Given (23)	Analyze (32)	Compare (154)	Find (1363)
Come (22)	Compose (31)	How (132)	Explain (1321)

Table 11: Top 10 most common "instruction words" (i.e., the first word in the instruction) for three regenerations of the Alpaca dataset by PaLM2. Columns indicate the number of iterations that produced a readability scoreable response for identical prompts. A separate column displays prevalent instruction words from all responses that could not be scored.

5 REINFORCEMENT LEARNING FROM HUMAN FEEDBACK (RLHF)

OpenAI underscored the role of reinforcement learning from human feedback (RLHF) when introducing ChatGPT [80]. This training process fine-tunes model responses using reinforcement learning techniques, such as Proximal Policy Optimization [96], informed by human feedback mechanisms like ranked lists, numeric scores, or simple approval indicators like thumbs up/down[17] [23, 83, 101]. This process seeks to align model generation with human preference while also attempting to mitigate the introduction of undesired biases [21, 36, 41, 42, 48, 57, 71, 83, 118, 119]. While several RLHF datasets exist [9, 39, 46, 72, 78], there is can be a noticeable lack of comprehensiveness in their documentation, it it exists at all. For instance, a review of the repository at https://github.com/opendilab/awesome-RLHF#dataset (accessed Oct. 9, 2023) reveals that only a few, such as those cited above, provide comprehensive insights into their RLHF dataset creation. Furthermore, of the 124 "rlhf" entries on HuggingFace Dataset Hub, the majority seem to be adaptations of datasets from Anthropic [9, 46]. The press release for ChatGPT [80] provides a general summary of the process used but

[17] As seen in the ChatGPT interface.

no explicit details nor do subsequent publications reveal them [81]. This contrasts with OpenAI's earlier RLHF publications, which presented more detailed descriptions [78, 101].

The reluctance to provide comprehensive information about dataset creation is concerning, particularly given the rising demands for transparency in AI research [7, 13, 50, 76, 87, 93]. The absence of details is especially problematic for models that have gained widespread use, as the biases in these models may not be immediately clear to users. For example, a closer look at the supplementary material from Stiennon et al.[101], on the use of RLHF to enhance summarization, reveals that out of the 21 assessors who participated in their survey, 18 had achieved at least an undergraduate degree, and all had completed high school. Given this context, it is evident that these assessors likely possess high literacy levels, with preferences shaped by their education. To better understand and highlight potential biases arising from such assessors, extending Bender and Friedman's assessor attributes [13] would be beneficial to include more nuanced details. Such extensions could incorporate literacy or reading comprehension metrics, in addition to the existing native language and linguistic training attributes.

The attributes of assessors can substantially influence the evaluation process, especially in terms of how they interpret and act on given instructions. To illustrate, in the Stiennon et al. [101] study, assessors were tasked with evaluating a summary based on its **Coherence**. The guideline provided defined a coherent summary as one that is *"A summary is coherent if, when read by itself, it is easy to understand and free of English errors. A summary is not coherent if it is difficult to understand what the summary is trying to say. Generally, it is more important that the summary is understandable than it being free of grammar errors"*. This directive raises an intriguing point in relation to Bender et al.'s perspective [14], which postulates that coherence, within the realm of human interactions, is deeply rooted in the intent behind communication. Generative models, by their very nature, do not inherently possess this communicative intent. Thus, when assessors are guided to gauge the coherence of a computer-generated summary without an explicitly defined intent or target audience, they are inadvertently left to infer this intent themselves. We argue that, under such circumstances, assessors are inclined to adopt a stance that resonates with their own perspectives and experiences. Consequently, this introduces biases aligned with their personal attributes. In the context of this paper, such bias may manifest with a preference for summaries that those who possess a high degree of literacy find understandable.

In our analysis of the study conducted by Steinnon et al. [101], it is evident that even well-executed research can inadvertently allow unintended biases to manifest, despite the study's overall merits. Building upon the "model usage context" concept introduced by Gong et al. [50] for model deployment, we advocate for clearly specifying both the intended application of the model and its target demographic when formulating instructions for assessors in RLHF and related tasks. By incorporating this information, assessors can tailor their evaluations to more closely align with the model's intended user base, eliminating the ambiguity of deciphering an overarching intent. Nonetheless, it is imperative to approach this with caution to prevent the introduction or reinforcement of biases, especially those that may arise from overtly specifying attributes like literacy or intelligence. Integrating these enhanced instructions with the annotator attributes recommended by Bender and Friedman [13], and gathering other relevant metrics—such as literacy levels—in an ethical manner, can offer a richer understanding of system performance. This comprehensive approach can potentially lead to recommendations for enhancing both the accessibility and efficacy of the system.

6 BEYOND "STANDARD" ENGLISH

Our analysis predominantly centred on what might be referred to as "standard" English, the form taught in academic settings, due to its ease of readability scoring and interpretation. However, many other English variants have evolved around the world. These include pidgins, such as Nigerian pidgin, creoles, like Jamaican patois,dialects, for instance, Cajun English, and other linguistic categorizations that deviate from "standard" English, such as African American (Vernacular) English. For many, these linguistic forms are deeply embedded in their cultural identity.

These linguistic variations represent a diverse range of expressions, enhancing content relatability and comprehension for their speakers. They challenge the dominance of standard English in readability, offering greater cognitive and cultural accessibility. Ensuring these variants are appropriately integrated into generative systems could substantially improve access to information. These languages have dedicated resources, such as Wikipedia pages for Jamaican Patois[18] and Nigerian Pidgin, and are featured in media outlets like the BBC.[19] In research contexts, these languages provide accessible data sources, reflecting similar experiences to that of low literacy users. For speakers of these variants, native linguistic presentation enhances readability. In IR, systems capable of processing these variants could greatly improve information accessibility and relevance for diverse users. The response of a Nigerian pidgin speaker to a generative system responding in college-level English *may be* similar to the experience of an eighth-grade level reader, highlighting the importance of considering various linguistic backgrounds when building IR systems. This is crucial for IR systems, as ensuring they meet diverse linguistic needs can enhance user accessibility and experience. The adaptability of generative systems to different linguistic contexts is essential, aligning with concepts of "communicative intent" [14] discussed earlier. By accommodating these variants, generative models can better reflect diverse user groups, improving both accessibility and readability.

Despite the unpredictability of generative models, utilizing language variants like Jamaican Patois could provide valuable insights into improving their accessibility. To test this, we applied five generative models to 31 instructions translated into Jamaican Patois, aiming to improve user experience and develop broader accessibility solutions.[20] The queries should be viewed merely as a convenience sample, enabling a rapid preliminary assessment of model behavior. This exploratory test highlighted challenges in processing brief, non-standard English inputs, providing crucial insights for creating linguistically adaptable and effective IR systems.

Responses from models like PaLM2, Llama2-7B, and Falcon-7B were generally unremarkable, except Falcon-7B's failure to respond

[18] jam.wikipedia.org
[19] https://www.bbc.com/pidgin
[20] The switch to Dolly was necessitated by a need for more "natural" queries rather than the task-driven nature of Alpaca.

to 20 instructions and LLama2-7B's incoherent Patois replies. GPT-3.5-Turbo and GPT-4, however, coherently responded in Patois, with GPT-4 showing more consistency. This highlights the difficulty in tailoring IR system responses to diverse cultural and linguistic contexts without stereotypes or biases. The complexity lies in appropriately adapting responses, such as avoiding overly simplistic answers for an eighth-grade-level query, which could be seen as patronizing. The results reveal the importance of balancing readability with cultural sensitivity in IR systems and demonstrate challenges that models face with brief, non-standard English inputs, indicating potential user experience and accessibility issues.

7 LOOKING FORWARD

For the IR community, understanding and addressing the challenges posed by increasingly prevalent user-visible generative technologies is essential. This work contributes novel insights into these challenges, which are becoming more apparent to end-users as information systems evolve. The complexity of these challenges is significant, extending beyond the primary expertise of many in the IR community. Historically, addressing issues of fairness, helpfulness, and harmlessness in information systems may not have been a central focus. In the current landscape of IR, it is crucial to consider these aspects, emphasizing the importance of our work in this context. Collaborative efforts with experts from diverse disciplines have been crucial in both enhancing and developing IR systems. Extending these collaborations with experts in literacy and cognitive impairments are essential to create IR systems that are technologically advanced, equitable, and accessible.

While previous research has explored the intersection of search and impairments such as dyslexia (see Section 2.3), as well as methodologies for conducting user studies in these contexts [16], there seems to be a gap in the literature concerning literacy and its relationship with generative information systems. This gap highlights the importance of examining if current models of user behavior remain relevant when catering to users with low literacy and other impairments, similar to Berget et al.'s study in the context of dyslexia [18]. Furthermore, analyzing the proficiency of low-literacy users in navigating generative information systems for information tasks could offer valuable insights. These insights can guide enhancements in both the user interface and the system's internal architecture to optimize search outcomes.

Examining generative systems is essential for balancing accessibility and privacy. This involves minimizing the amount of context needed to be provided in a query, and ensuring that such support does not diminish those who use it (e.g., negatively impacting their self-esteem). Google's Bard already offers a "simplify" button, targeting reading levels between fourth and sixth grades for simplified outputs, while regular responses aim for seventh to twelfth grades, aligning with our PaLM2 findings. Simplification methods can vary, and in general, they should be tailored based on the needs of the particular user. For example, adjustments for general reading levels might not be the same as those tailored for users with dyslexia. Issues around "prompt tailoring" for readability and accessibility are complex and beyond this work. We avoided testing simple modifications (e.g., "Respond as if I can only read at ... level.") as we could not imagine someone, with a readability issue or other impairment, manually performing or selecting such modifications to their query without this option being carefully presented to them.

A counterargument to enhancing visual interfaces in generative information systems is to simply provide a voice assistant for users but this solution has its own limitations. Individuals with low literacy might still face challenges comprehending a spoken response if it uses vocabulary beyond their understanding. Furthermore, "listenability" gauges the ease of understanding spoken words, analogous to the readability metrics discussed in this paper [35, 53]. The concept of listenability within information science and retrieval appears to be relatively uncharted in our estimation, further research could offer more avenues for users to satisfy their information needs and potentially enhance accessibility in textual formats as a by-product.

Regardless of how future generative information systems deliver responses, it is essential to recognize and address the inherent biases when using humans or generative models as representatives during a system's design and feedback stages [40, 66, 106]. For example, when examining techniques for long-form question-answering, Nakano et al. selected contractors with the following in mind, "[d]ue to the challenging nature of the tasks, contractors were generally highly educated, usually with an undergraduate degree or higher," which indicates that such tasks may be beyond Level 2 in the PIAAC framework (Figure 1). Thus, when relying on highly educated individuals or models that replicate such expertise for intricate tasks, we need to be particularly careful in the design of feedback mechanisms, ensuring they do not skew results toward those "capable" of providing feedback in the desired manner.

Improving accessibility of generative information systems has both social and economic benefits. A recent Gallup study [94] argued that improving U.S. literacy levels could increase annual income by approximately $2.2 trillion, underscoring a large market for accessible systems. Such systems can boost productivity and income for low-literacy users in work environments, yielding commercial benefits for businesses focusing on accessibility. While not solving all societal literacy issues, this approach highlights the dual advantage of improved accessibility: better user experiences and competitive commercial gains. This makes addressing accessibility into a win-win scenario.

8 CONCLUSION

Our work addresses critical concerns about integrating generative models into IR systems, particularly highlighting potential accessibility issues for users with low literacy or reading impairments. Being aware of and understanding this gap in accessibility allows IR practitioners to more conscientiously design and develop systems to ensure that information access is not limited to a select group. We showed that the training data for generative models often leads to outputs akin to college-level writing, as demonstrated by our analysis of responses from five popular models using the Alpaca dataset. This tendency of models to mirror proficient human writing styles underscores the need for more inclusive training approaches in IR. We also discuss potential strategies to enhance accessibility in information systems and the challenges inherent in such efforts, underscoring the importance of this work in advancing equitable and user-friendly systems.

REFERENCES

[1] 1992. OFFICIAL ORDER of the COMMISSIONER OF INSURANCE of the STATE OF TEXAS: Adoption of Flesch Reading Ease Test (No. 92 - 0573). https://www.tdi.texas.gov/pubs/pc/pccpfaq.html.

[2] 2023. Florida Statutes Section 627.4145 - Readable Language In Insurance Policies. http://www.leg.state.fl.us/statutes/index.cfm?App_mode=Display_Statute&Search_String=&URL=0600-0699/0627/Sections/0627.4145.html.

[3] Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. 2023. Information Retrieval Meets Large Language Models: A Strategic Report from Chinese IR Community. *AI Open* 4 (2023), 80–90.

[4] Rodrigo AlarconLourdes Alarcon, Lourdes Moreno, and Paloma Martínez. 2020. Word-Sense disambiguation system for text readability. In *Proceedings of the 9th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*. 147–152.

[5] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance. (2023).

[6] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. PaLM2 Technical Report. (2023). arXiv:2305.10403

[7] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. 2019. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv:1808.07261

[8] Akari Asai, Matt Gardner, and Hannaneh Hajishirzi. 2022. Evidentiality-guided Generation for Knowledge-Intensive NLP Tasks. arXiv:2112.08688

[9] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862

[10] Alan Bailin and Ann Grafstein. 2016. *Readability: Text and context.* Springer.

[11] Aparna Balagopalan, Abigail Z. Jacobs, and Asia J. Biega. 2023. The Role of Relevance in Fair Ranking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*.

[12] Solon Barocas, Asia J. Biega, Benjamin Fish, Jundefineddrzej Niklas, and Luke Stark. 2020. When Not to Design, Build, or Deploy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*.

[13] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018).

[14] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*.

[15] Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review* 24 (2012), 63–88.

[16] Gerd Berget and Andrew MacFarlane. 2019. Experimental Methods in IIR: The Tension between Rigour and Ethics in Studies Involving Users with Dyslexia. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*.

[17] Gerd Berget and Andrew MacFarlane. 2020. What is known about the impact of impairments on information seeking and searching? *Journal of the Association for Information Science and Technology* 71, 5 (2020), 596–611.

[18] Gerd Berget, Andrew MacFarlane, and Nils Pharo. 2021. Modelling the information seeking and searching behaviour of users with impairments: Are existing models applicable? *Journal of Documentation* 77, 2 (2021).

[19] Vikas L Bommineni, Sanaea Bhagwagar, Daniel Balcarcel, Vishal Bommineni, Christos Davazitkos, and Donald Boyer. 2023. Performance of ChatGPT on the MCAT: The Road to Personalized and Equitable Premedical Learning. *medRxiv* (2023).

[20] Matthew D Carter, Marianna M Walker, Kevin O'Brien, and Monica S Hough. 2019. The effects of text length on reading abilities in accelerated reading tasks. *Speech, Language and Hearing* 22, 2 (2019), 111–121.

[21] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv:2307.15217

[22] J.S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula.* Brookline Books.

[23] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741

[24] Lynne Cole, Andrew MacFarlane, and George Buchanan. 2016. Does dyslexia present barriers to information literacy in an online environment? A pilot study. *Library and Information Research* 40, 123 (2016), 24–46.

[25] Lynne Cole, Andrew MacFarlane, and Stephann Makri. 2020. More than words: the impact of memory on how undergraduates with dyslexia interact with information. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 353–357.

[26] Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60 (1975).

[27] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM.* https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm

[28] Stephanie Cosentino, Jennifer Manly, and Dan Mungas. 2007. Do reading tests measure the same construct in multilingual and multiethnic elders? *Journal of the International Neuropsychological Society: JINS* 13, 2 (2007), 228.

[29] Ben Cost. 2023. Married father commits suicide after encouragement by AI chatbot: widow. https://nypost.com/2023/03/30/married-father-commits-suicide-after-encouragement-by-ai-chatbot-widow/.

[30] Edgar Dale and Jeanne S. Chall. 1948. A Formula for Predicting Readability. *Educational Research Bulletin* 27 (1948).

[31] Anubrata Das, Kunjan Mehta, and Matthew Lease. 2019. CobWeb: A Research Prototype for Exploring User Bias in Political Fact-Checking. arXiv:1907.03718

[32] Terry C. Davis and Michael S. Wolf. 2004. Health literacy: implications for family medicine. *Family Medicine* 36 (2004). Issue 8.

[33] Erik Derner and Kristina Batistič. 2023. Beyond the Safeguards: Exploring the Security Risks of ChatGPT. arXiv:2305.08005

[34] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv preprint arXiv:2305.14314* (2023).

[35] Teresa Hafer Donald L. Rubin and Kevin Arata. 2000. Reading and listening to oral-based versus literate-based discourse. *Communication Education* 49, 2 (2000).

[36] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. arXiv:2304.06767

[37] Shiri Dori-Hacohen and Scott A. Hale. 2022. Information Ecosystem Threats in Minoritized Communities: Challenges, Open Problems and Research Directions. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*.

[38] Shiri Dori-Hacohen, Roberto Montenegro, Fabricio Murai, Scott A. Hale, Keen Sung, Michela Blain, and Jennifer Edwards-Johnson. 2021. Fairness via AI: Bias Reduction in Medical Information. arXiv:2109.02202

[39] Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information. In *Proceedings of the 39th International Conference on Machine Learning*.

[40] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '23)*.

[41] Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. 2023. Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation. arXiv:2305.00955

[42] Emilio Ferrara. 2023. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. arXiv:2304.03738

[43] Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology* (1948).

[44] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. , 277–304 pages.

[45] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and Fairness in Large Language Models: A Survey. *arXiv preprint arXiv:2309.00770* (2023).

[46] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858

[47] Sourojit Ghosh and Aylin Caliskan. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. *arXiv preprint arXiv:2305.10510* (2023).

[48] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. arXiv:2209.14375

[49] Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, Rerank, Generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[50] Lina Gong, Jingxuan Zhang, Mingqiang Wei, Haoxiang Zhang, and Zhiqiu Huang. 2023. What Is the Intended Usage Context of This Model? An Exploratory Study of Pre-Trained Models on Various Model Repositories. *ACM Trans. Softw. Eng. Methodol.* 32, 3 (may 2023).

[51] Michael Green. 2021. Why Don't You Act Your Age?: Recognizing the Stereotypical 8-12 Year Old Searcher by Their Search Behavior. (2021).

[52] Laurence Habib, Gerd Berget, Frode Eika Sandnes, Nick Sanderson, Philippe Kahn, Siri Fagernes, and Ali Olcay. 2012. Dyslexic students in higher education and virtual learning environments: An exploratory study. *Journal of Computer Assisted Learning* 28, 6 (2012), 574–584.

[53] Kenneth A. Harwood. 1955. I. Listenability and readability. *Speech Monographs* 22, 1 (1955).

[54] Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2023. FiD-Light: Efficient and Effective Retrieval-Augmented Text Generation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*.

[55] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations*.

[56] Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models. *arXiv preprint arXiv:2309.05454* (2023).

[57] Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2023. Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias. arXiv:2308.00225

[58] Jerry L. Johns and Thomas E. Wheat. 1984. Newspaper Readability: Two Crucial Factors. *Journal of Reading* 27, 5 (1984).

[59] Lorna Kane, Joe Carthy, and John Dunnion. 2006. Readability applied to information retrieval. In *European Conference on Information Retrieval*. Springer, 523–526.

[60] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. GPT-4 Passes the Bar Exam. https://ssrn.com/abstract=4389233. (15 3 2023).

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

[62] Anton Korinek. 2023. Exploring the impact of language models on cognitive automation with David Autor, ChatGPT, and Claude. (2023).

[63] Birgit Kvikne and Gerd Berget. 2021. In search of trustworthy information: a qualitative study of the search behavior of people with dyslexia in Norway. *Universal Access in the Information Society* 20 (2021), 1–12.

[64] Amelia Lake. 2022. Hidden in Plain Sight: The Secret Epidemic of Illiteracy in the United States. https://www.yalehrj.org/post/hidden-in-plain-sight-the-secret-epidemic-of-illiteracy-in-the-united-states. *Yale Human Rights Journal* (2022).

[65] Ida Larsen-Ledet, Bhaskar Mitra, and Siân Lindley. 2022. Ethical and Social Considerations in Automatic Expert Identification and People Recommendation in Organizational Knowledge Management Systems. arXiv:2209.03819

[66] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. arXiv:2309.00267

[67] Peter Lee. 2016. Learning from Tay's introduction. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

[68] Timo Lenzner. 2014. Are readability formulas valid tools for assessing survey question difficulty? *Sociological Methods & Research* 43, 4 (2014), 677–698.

[69] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.* 55, 9 (2023).

[70] Ruohan Li, Jianxiang Li, Bhaskar Mitra, Fernando Diaz, and Asia J. Biega. 2022. Exposing Query Identification for Search Transparency. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*.

[71] Gabrielle Kaili-May Liu. 2023. Perspectives on the Social Impacts of Reinforcement Learning with Human Feedback. arXiv:2303.02891

[72] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training Socially Aligned Language Models in Simulated Human Society.

[73] Andrew MacFarlane, A Albrair, CR Marshall, and George Buchanan. 2012. Phonological working memory impacts on information searching: An investigation of dyslexia. In *Proceedings of the 4th Information Interaction in Context Symposium*. 27–34.

[74] Eric Martínez. 2023. Re-Evaluating GPT-4's Bar Exam Performance. https://ssrn.com/abstract=4441311. (18 5 2023).

[75] G. Harry McLaughlin. 1969. SMOG Grading-a New Readability Formula. *Journal of Reading* 12, 8 (1969).

[76] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*.

[77] Emiliana Murgia, Maria Soledad Pera, Monica Landoni, and Theo Huibers. 2023. Children on ChatGPT Readability in an Educational Context: Myth or Opportunity?. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 311–316.

[78] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback.

[79] Administration of William J. Clinton. 1998. Memorandum on Plain Language in Government Writing.

[80] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

[81] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774

[82] Organization for Economic Cooperation and Development (OECD). 1999. *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Author, Paris.

[83] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155

[84] Srishti Palani, Adam Fourney, Shane Williams, Kevin Larson, Irina Spiridonova, and Meredith Ringel Morris. 2020. An eye tracking study of web search by people with and without dyslexia. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 729–738.

[85] Yaneth del Rosario Palo Villegas, Andrea Elena Pomareda Vera, María Elena Rojas Zegarra, and M Dolores Calero. 2020. Effectiveness of the "Mente Sana [Healthy Mind]" cognitive training program for older illiterate adults with mild cognitive impairment. *Geriatrics* 5, 2 (2020), 34.

[86] Karl Magnus Petersson, Alexandra Reis, and Martin Ingvar. 2001. Cognitive processing in literate and illiterate subjects: A review of some recent behavioral and functional neuroimaging data. *Scandinavian journal of psychology* 42, 3 (2001), 251–267.

[87] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*.

[88] Huachuan Qiu, Shuai Zhang, Anqi Li, Hongliang He, and Zhenzhong Lan. 2023. Latent Jailbreak: A Benchmark for Evaluating Text Safety and Output Robustness of Large Language Models. arXiv:2307.08487

[89] Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019. Better language models and their implications. *OpenAI blog* 1, 2 (2019).

[90] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020).

[91] Bobby D. Rampey, Madeline Goodman Robert Finnegan and, Leyla Mohadjer, Tom Krenzke, Jacquie Hogan, and Stephen Provasnik. 2016. Skills of U.S. Unemployed, Young, and Older Adults in Sharper Focus: Results from the Program for the International Assessment of Adult Competencies (PIAAC) 2012/2014:

First Look. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2016039rev.
[92] L Revell. 1994. Understanding, identifying, and teaching the low-literacy patient.. In *Seminars in Perioperative Nursing*, Vol. 3. 168–171.
[93] Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Comput. Surv.* 55, 10 (2023).
[94] Johnathan Rothwell. 2020. Assessing the Economic Gains of Eradicating Illiteracy Nationally and Regionally in the United States. https://www.barbarabush.org/wp-content/uploads/2020/09/BBFoundation_GainsFromEradicatingIlliteracy_9_8.pdf.
[95] Kevin Schaul, Szu Yu Chen, and Nitasha Tiku. 2023. Inside the secret list of websites that make AI like ChatGPT sound smart. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/. (19 4 2023).
[96] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347
[97] Marilyn Schultz and H Lippman. 2002. Low literacy skills needn't hinder care. *RN* 65, 4 (2002), 45–48.
[98] RJ Senter and Edgar A Smith. 1967. *Automated readability index.* Technical Report. Technical report, DTIC document.
[99] Government Digital Service. 2023. Content design: planning, writing and managing content. https://www.gov.uk/guidance/content-design/writing-for-gov-uk.
[100] Marcela Lima Silagi, Vivian Urbanejo Romero, Maira Okada de Oliveira, Eduardo Sturzeneker Trés, Sonia Maria Dozzi Brucki, Márcia Radanovic, and Leticia Lessa Mansur. 2021. Inference comprehension from reading in individuals with mild cognitive impairment. *Acta Neurologica Belgica* 121 (2021), 879–887.
[101] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, Vol. 33.
[102] Lauren M. Stossel, Nora Segar, Peter Gliatto, Robert Fallar, and Reena Karani. 2012. Readability of Patient Education Materials Available at the Point of Care. *Journal of General Internal Medicine* 27 (9 2012). Issue 9.
[103] H Lee Swanson. 2012. Adults with reading disabilities: Converting a meta-analysis to practice. *Journal of Learning disabilities* 45, 1 (2012), 17–30.
[104] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.
[105] Binod Thapa-Chhetry and Tia Keck. 2019. A Chrome app for improving reading comprehension of health information online for individuals with low health literacy. In *2019 IEEE/ACM 1st International Workshop on Software Engineering for Healthcare (SEH)*. IEEE, 57–64.
[106] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large Language Models Can Accurately Predict Searcher Preferences. (September 2023). https://www.microsoft.com/en-us/research/publication/large-language-models-can-accurately-predict-searcher-preferences/
[107] Tourism and the Centre for Education Statistics Division. 2013. Skills in Canada: First Results from the Programme for the International Assessmentof Adult Competencies (PIAAC). http://www.cmec.ca/Publications/Lists/Publications/Attachments/315/Canadian-PIAAC-Report.EN.pdf.
[108] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971
[109] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]
[110] W3C. [n. d.]. Understanding WCAG 2.0: Reading Level: Understanding SC 3.1.5. https://www.w3.org/TR/UNDERSTANDING-WCAG20/meaning-supplements.html.
[111] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*.
[112] Shuai Wang, Harrisen Scells, Martin Potthast, Bevan Koopman, and Guido Zuccon. 2023. Generating Natural Language Queries for More Effective Systematic Review Screening Prioritisation. arXiv:2309.05238
[113] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-Instruct: Aligning Language Model with Self Generated Instructions.
[114] Willian Massami Watanabe. 2010. Facilita: reading assistance to the functionally illiterate. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*. 1–2.
[115] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.
[116] Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025* (2023).
[117] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564
[118] Yachao Zhao, Bo Wang, Dongming Zhao, Kun Huang, Yan Wang, Ruifang He, and Yuexian Hou. 2023. Mind vs. Mouth: On Measuring Re-judge Inconsistency of Social Bias in Large Language Models. arXiv:2308.12578
[119] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. arXiv:2305.11206
[120] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043