The 36th Canadian Conference on Artificial Intelligence
DOI: 0

Canadian AI 2023
5—9 June 2023, Montréal

Don't Worry Accountants, ChatGPT Won't Be Taking Your Job...Yet

Stacey Taylor[†,*], Vlado Keselj[†]
[†] Faculty of Computer Science, Dalhousie University

Abstract

ChatGPT has demonstrated the ability to generate plausible human-like text and research is underway to evaluate and benchmark its current performance in various domains. The research we present here provides a preliminary benchmark on ChatGPT's ability to emulate the style and information presented in financial statement note disclosures. Using text from Canada's major banks ($n = 5$) over the period of 2019–2021, we query ChatGPT to generate two required note disclosures and compare its text against the note disclosures written by the banks in their corporate annual reports. We find that the similarity between ChatGPT's text and the human-authored text is very low, but also find that ChatGPT's text is significantly more readable for one of the two disclosures ($p < 0.05$).

Keywords: ChatGPT, Machine Learning, Financial Statements, Similarity, Stylometry, Readability

1. Introduction

The release of ChatGPT has raised both interest and concern over its human-like text generation. Given its utility and range of generation abilities, this type of tool would be very useful in the creation of long regulatory documents such as the annual report for public companies, which is, on average, 186 pages [1]. While ChatGPT is widely trained and has shown its abilities to emulate styles and provide realistic answers to requests (with some error), this tool is still in its infancy. As such, we anticipate that the efficacy of this tool will evolve over time. This progression is of interest to both the scientific and professional communities. The purpose of this research is to examine and benchmark ChatGPT's current ability to generate the notes to the financial statements for public companies. This is important for several reasons. Textual financial statement data is scarce and the utility of ChatGPT to augment financial data should be evaluated. Also, ChatGPT's ability to emulate could be used to produce fraudulent or incorrect financial statement data, which may be difficult to detect in time-sensitive situations. Therefore, the type of research and evaluation that we have presented in this paper can help address these concerns.

To conduct this research, we chose two note disclosures from the Annual Report for ChatGPT to generate for Canada's five major banks over the period of 2019–2021. We find that when ChatGPT's text is compared to that of human authors, the similarity is very low, with the highest at 0.42 for disclosure 1, and 0.32 for disclosure 2. We also find that not all texts are easily identifiable as either human-authored or Artificial Intelligence (AI)-authored. This was unexpected, but also revealed that there is more "boilerplate" text in the disclosure notes than expected. Finally, results also show that readability is significantly improved at a statistical significance level of $p < 0.05$ for disclosure 1 when using ChatGPT. This also opens up an interesting avenue of research on how AI tools like ChatGPT can be used to help improve readability in order to make complicated and difficult financial text more accessible to a wider population.

[*] stacey.taylor@dal.ca

The rest of our paper is organized as follows: Section 2 provides a brief background on financial note disclosures; Section 3 addresses related work; Section 4 outlines the methodology; Section 5 provides the results and discussion; and Section 6 gives the conclusion, future work, and limitations.

2. Background

In Canada, public companies are required to use the International Financial Reporting Standards (IFRS) when preparing interim and annual financial statements [2]. A critical part to the financial statements are the *note disclosures*, often referred to simply as "notes". These disclosures provide critical additional information about items recognized in the financial statements [3, 4], as well as those that are not [5]. Disclosures may be required by the Generally Accepted Accounting Principles (GAAP) or other regulations, while other disclosures may be provided by management to facilitate user understanding [5]; there is no "one size fits all" approach to disclosures.

Per International Accounting Standard (IAS) 1, "[t]he notes must present information about the basis of preparation of the financial statements and the specific accounting policies used, disclose any information required by IFRS that is not presented elsewhere in the financial statements and, provide additional information that is not presented elsewhere in the financial statements but is relevant to an understanding of any of them" [6].

Two note disclosures were selected for this research: "basis for preparation" and "subordinated debt". Both are required note disclosures under IFRS, specifically IAS 1 and IFRS 7 [6, 7]. They are referred to as "disclosure 1" and "disclosure 2" in the paper.

3. Related Work

ChatGPT is a fine-tuned iteration of the large language model GPT-3.5, trained using both supervised and reinforcement learning [8]. It was released as a chatbot by OpenAI on Nov 30, 2022. As it has only been several months since ChatGPT's release, evaluation methods are currently being explored and are relatively new. For that reason, we have looked to existing (and older) literature for evaluation of human-authored texts as well as emerging literature focused on ChatGPT.

Cosine and Jaccard similarity measures are well-known and often used to evaluate text similarity. For text clustering, Huang found that Jaccard produced more pure clustering than Cosine [9]. Qurashi *et al.* found that Cosine similarity provided better semantic analysis than Jaccard, pointing out that while Jaccard is a popular tool, it is a lexical tool that does not perform well for semantic analysis [10]. Singh *et al.* used Burrows' Delta, Kilgariff's Chi-Square, and Mendenhall's method, and found that, when presented with unlabelled texts, the best results for identifying the most likely author were returned using Burrows' Delta [11]. Smeuninx *et al.*'s work evaluated CEO letters from the yearly annual report and compared them against sustainability reports. Using the Flesch Reading Ease Score, they found that the CEO letters were easier to read at a statistical significance level of $p < 0.001$ [12].

Emerging research has also been done on evaluating ChatGPT in various domains. Ventayen compared ChatGPT generated responses with that of pre-existing human-authored research papers, using the paper's title for the query [13]. The generated texts were then passed through *Turnitin*, a widely used plagiarism software, and found to pass Pangasinan State University (where the research was conducted) similarity thresholds [13]. Frieder *et al.* evaluated ChatGPT's math capabilities by querying ChatGPT with 728 different prompts.

Questions ranged from elementary math problems to those tackled by olympiads [14]. ChatGPT's answers were then evaluated by human experts. Results show that, even with ChatGPT's extensive training, its capabilities are far below those of the average grad student. Frieder *et al.* also point out that while there is evidence that ChatGPT understands the question, it often fails to return the correct answer, sparking the ironic realization that a student would do better cheating off of a peer than using ChatGPT [14].

4. Methodology

The Annual Reports were gathered from the System for Electronic Document Analysis and Retrieval (SEDAR)[1] for three years over the period of 2019 – 2021 for Canada's five major banks: Royal Bank of Canada (RBC), Canadian Imperial Bank of Commerce (CIBC), Bank of Nova Scotia (BNS), TD Bank (TD), and Bank of Montreal (BMO). Using the listing of the notes to the financial statements, we selected two financial statement note disclosures that the five banks had in common. We took this approach as many disclosures are specific to the company's financial statements and may be different to those of another company. The note disclosures that were selected are: (1) Basis of Preparation and (2) Subordinated Debt. For reproducibility purposes, Table 1 provides where each of these note disclosures can be found in each report. We also note that while disclosures may have a slightly different name between companies (e.g. *General Information* (RBC [16]) versus *Basis for Preparation* (CIBC [17])), the intent and information provided in the disclosures are substantially the same. This naming difference is also clearly outlined in Table 1.

Disclosure Name	Bank	Year (Page)
1: General Information	RBC	2019 (125), 2020 (132), 2021 (138)
1: Basis for Preparation	CIBC	2019 (108), 2020 (114), 2021 (122)
1: Statement of Compliance	BNS	2019 (148), 2020 (160), 2021 (158)
1: Nature of Operations	TD	2019 (132), 2020 (137), 2021 (139)
1: Basis of Presentation	BMO	2019 (142), 2020 (150), 2021 (151)
2: Subordinated Debentures	RBC	2019 (190), 2020 (200), 2021 (205)
2: Subordinated Indebtedness	CIBC	2019 (159), 2020 (162), 2021 (168)
2: Subordinated Debentures	BNS	2019 (204), 2020 (213), 2021 (212)
2: Subordinated Notes & Debentures	TD	2019 (188), 2020 (192), 2021 (188)
2: Subordinated Debt	BMO	2019 (176), 2020 (183), 2021 (616)

Table 1. Disclosure References by Bank and Year

The following questions were given to ChatGPT for generation:

For disclosure 1: "Write the <name of the disclosure> note disclosure for <insert bank name here> for the <insert year> Annual Report." As discussed above, there is some variability in the name of this disclosure. For example, the Royal Bank uses "General Information" while CIBC uses "Basis of Preparation". To account for this variability in naming convention, the proper names (and casings) found in the bank's annual report were used. We did this to ensure that we queried ChatGPT using the most accurate information for the question in relation to each bank.

As an example, the question for RBC for 2019 was: "Write the *General Information* note disclosure for the Royal Bank of Canada for the 2019 Annual Report", whereas the question for CIBC for 2019 was "Write the *Basis of Preparation* note disclosure for the Canadian Imperial Bank of Commerce for the 2019 Annual Report."

For disclosure 2: "Write the note disclosure for <name of the note disclosure> for <insert bank name here> <insert year here> Annual Report, including a table of the

[1] SEDAR is the official regulatory filing repository for the Canadian Securities Administrators [15]

debentures and any relevant footnotes." We formulated the question this way as the note disclosures in the annual report contain a table and footnotes. We tried asking the question without mentioning the table and the footnotes, and ChatGPT would not include these elements in its generation without specific direction to do so.

Like disclosure 1, there is some variability in the naming convention such as "Subordinated Notes and Debentures" or "Subordinated Debt", for example. And so we followed our approach from disclosure 1 and used the name of the disclosure (and its casing) as it was presented in the annual report.

As an example, the question for TD Bank for 2019 was "Write the note disclosure for *Subordinated Notes and Debentures* for TD Bank's 2019 Annual Report, including a table of the debentures and any relevant footnotes.", whereas the question for BMO for 2019 was "Write the note disclosure for *Subordinated debt* for the Bank of Montreal's 2019 Annual Report, including a table of the debentures and any relevant footnotes."[2]

While querying ChatGPT for the tables may seem extraneous given that they were ultimately removed (discussed below), we needed the tables to be generated in order to capture the footnotes. Also, we wanted to ensure that we used the exact same protocol for the human-authored text and the AI-generated text by ChatGPT in order to address any concerns of bias in the question posed to ChatGPT as well as any concern that we were not comparing truly "like" text.

The text of the note disclosures were then extracted from each report to be used for evaluation purposes. Tables for disclosure 2 were removed from both the bank's annual report text, as well as ChatGPT's text. Any footnotes, however, were retained, as any text outside of the tables should be included. In the interests of full disclosure and responsible use of AI, it should be noted that ChatGPT was used to generate the note disclosures for comparative purposes only and that text was considered and used as a "data source" in the context of this research. ChatGPT did not aid in the formulation of the research problem, the research itself, or the writing of the paper.

To assess the original and generated texts, two measures of similarity were used (Cosine and Jaccard); a stylometry analysis was conducted using Burrows' Delta, and readability was evaluated using the Flesch Reading Ease Score.

4.1. Cosine Similarity

We used Cosine similarity as it is the standard similarity measure for Information Retrieval. It is based on the word array vector representations of the texts and the cosine value of these two vectors [9, 18]. An important aspect of Cosine similarity that is particularly useful in our case is that text length is not a limiting factor and is therefore an appropriate measure for similarity of texts of different lengths [9]. The Cosine similarity formula between vectors x and y is given in Eq. 4.1.

$$Cosine(x, y) = \frac{x \cdot y}{|x| \cdot |y|} \tag{4.1}$$

4.2. Jaccard Similarity

Jaccard similarity is another well-known measure that evaluates the similarity of two sets, U and V. Like Cosine similarity between vectors with positive components (i.e., in the first quadrant), the range of Jaccard similarity is between 0 and 1, where 0 indicates that there is no overlap, and 1 indicates there is complete overlap. Therefore, if the result is closer to

[2] If the paper is accepted, we will provide a URL to a github repository that contains the dataset that includes the original annual reports from the banks, the extracted disclosure text, as well as the ChatGPT generated text.

0, is it said to be dissimilar, whereas results closer to 1 indicate a high similarity between the sets [18, 19]. The formula for the Jaccard similarity is

$$Jaccard(U, V) = \frac{|U \cap V|}{|U \cup V|} \tag{4.2}$$

4.3. Burrows' Delta

Following the work of Laramée, we also used a stylometry tool to evaluate if we could detect if the text was human-authored or AI-generated (ChatGPT), using Burrows' Delta [20, 21]. Using a masked text author, features are equally weighted to identify the likely author [20]. Disclosures were grouped into disclosure 1 and disclosure 2 (as per Table 1), and any indication of the text author was removed from the documents, including the document names, which are named as "`disc1_text1`", "`disc1_text2`", ... (for disclosure 1) and "`disc2_text1`", "`disc2_text2`", ... (for disclosure 2). For each disclosure, there were a total of 30 texts — 15 that were human-authored and 15 that were AI-generated using ChatGPT. 8 texts from each were randomly selected as the training set and labelled as either "human" or "GPT". The remaining 14 documents were then labelled as "disputed" and made available for testing. Burrows' Delta was used to identify the likely author of the disputed documents. We randomly selected 4 test documents from the disputed label (documents 22, 11, 1, and 30) for each disclosure and calculated the Delta for each.

$$\Delta_c = \sum_i \frac{\left|Z_{c(i)} - Z_{t(i)}\right|}{n} \tag{4.3}$$

where c is the disputed document, i is the test document, $Z_{c(i)}$ is the Z-score of word frequencies i in the disputed document, $Z_{t(i)}$ is the Z-score of the word frequencies of the test document, and n is the number of unique words [20].

4.4. Flesch Reading Ease Score

To calculate the readability of the documents, the spaCy Readability package was used [22]. This package provides the readability scores for Flesch-Kincaid Grade Level, Flesch(-Kincaid) Reading Ease, and Dale-Chall. The *Grade Level* test focuses on the grade level needed in order to read the text. The purpose of our analysis is to examine how difficult the financial disclosures are for adults. Therefore, grade level is not within the scope of our research. *Dale-Chall* incorporates the percentage of difficult words in a sentence, along with the average length of the sentence [23]. As financial disclosures are not general text, we believe that this measure would be biased, given the complexity of the subject matter.

Therefore, we elected to use *Flesch Reading Ease Score*, which measures the ease of reading a text passage. The scoring ranges from 0 to 100 [24]. Texts are determined to be harder to read as the score drops. Our focus was on texts which scored below 60, as that indicates that the text is no longer *easy* to read [24]. Texts that are below 50 are assessed as "university level", and those below 30 are considered only readable by university graduates [24]. Given that some very well-known and successful entrepreneurs have no university degree (e.g. Bill Gates, Mark Zuckerberg, Steve Jobs, and Richard Branson [25]), we will expand this interpretation to indicate that scores below 50 are business entry-level, and scores below 30 are business professional-level, where professionals have several years of experience.

The equation for the Flesch Reading Ease (FRE) score is as follows:

$$\text{FRE} = 206.835 - 1.1015 \cdot \frac{\text{total words}}{\text{total sentences}} - 84.6 \cdot \frac{\text{total syllables}}{\text{total words}} \tag{4.4}$$

5. Results and Discussion

5.1. Cosine Similarity and Jaccard Similarity

The results of the Cosine Similarity and Jaccard Similarity are found in Table 2 and Table 3, respectively. As expected, the similarity between the human-authored text and the AI-generated text by ChatGPT are very low using both measures.

Bank, Year	Human Vs. GPT	Human YoY	ChatGPT YoY	Bank, Year	Human Vs. GPT	Human YoY	ChatGPT YoY
Disclosure 1				**Disclosure 2**			
RBC, 2019	0.23	N/A	N/A	RBC, 2019	0.30	N/A	N/A
RBC, 2020	0.23	0.97	0.88	RBC, 2020	0.31	0.95	0.97
RBC, 2021	0.19	0.96	0.91	RBC, 2021	0.31	0.92	0.97
CIBC, 2019	0.31	N/A	N/A	CIBC, 2019	0.29	N/A	N/A
CIBC, 2020	0.28	0.87	0.93	CIBC, 2020	0.28	0.94	0.99
CIBC, 2021	0.27	0.94	0.96	CIBC, 2021	0.29	0.94	0.99
BNS, 2019	0.42	N/A	N/A	BNS, 2019	0.29	N/A	N/A
BNS, 2020	0.42	0.95	0.99	BNS, 2020	0.32	0.79	0.99
BNS, 2021	0.42	0.95	0.99	BNS, 2021	0.30	0.99	0.97
TD, 2019	0.40	N/A	N/A	TD, 2019	0.30	N/A	N/A
TD, 2020	0.35	0.95	0.99	TD, 2020	0.32	0.79	0.99
TD, 2021	0.35	0.99	0.99	TD 2021	0.26	0.95	0.99
BMO, 2019	0.39	N/A	N/A	BMO, 2019	0.25	N/A	N/A
BMO, 2020	0.40	0.98	0.99	BMO, 2020	0.25	0.94	0.99
BMO, 2021	0.40	0.99	0.99	BMO, 2021	0.26	0.95	0.99
Mean	0.34	0.96	0.96	**Mean**	0.29	0.90	0.98
Median	0.35	0.96	0.99	**Median**	0.29	0.94	0.99
Std Dev	0.08	0.03	0.04	**Std Dev**	0.02	0.10	0.01

Table 2. Results for Cosine Similarity

The results indicate that the similarity never reaches 0.50 for either disclosure under both measures. The highest similarity is 0.42 and 0.32 for disclosures 1 and 2, respectively, using Cosine similarity. The results for Jaccard are even lower, with the highest similarity at 0.20 for disclosure 1 and 0.15 for disclosure 2.

This strongly supports that ChatGPT is not yet able to capture the bank's *voice* when writing these disclosures, even for a straightforward disclosure such as *The Basis of Preparation.* We do note, though, that similarity scores for disclosure 1 are higher than those for disclosure 2, indicating that ChatGPT's performance is better for disclosure 1 than 2.

Another interesting finding from these results is that, like ChatGPT, the human-authored text *also* takes a very "boilerplate approach" for both disclosures as indicated by the high degree of similarity year-over-year (YoY). This raises an important question — how many times does a text have to be used before it is considered boilerplate? Current thinking suggests that text need only be used a few times before it is considered "boilerplate". Given that annual reports are very long, the analysis that is done in our work here opens up a new avenue of research in examining important questions like "Are investors more or less likely to miss key information if disclosures take a boilerplate approach?" or, "Is there a better way to communicate the changes from year-to-year of a (mostly) boilerplate report to investors?"

Also, interest has been growing in this area in the past five years, with research looking at boilerplate detection itself, but also targeted removal of boilerplate from either documents or web corpora [26–28]. Given what we have shown in our results, removal of boilerplate could have unintended consequences, resulting in the removal of large swaths of text from regulatory documents or corpora.

Bank, Year	Human Vs. GPT	Human YoY	ChatGPT YoY	Bank, Year	Human Vs. GPT	Human YoY	ChatGPT YoY
Disclosure 1				**Disclosure 2**			
RBC, 2019	0.16	N/A	N/A	RBC, 2019	0.14	N/A	N/A
RBC, 2020	0.15	0.95	0.74	RBC, 2020	0.13	0.89	0.95
RBC, 2021	0.15	0.92	0.81	RBC, 2021	0.14	0.84	0.94
CIBC, 2019	0.15	N/A	N/A	CIBC, 2019	0.13	N/A	N/A
CIBC, 2020	0.14	0.67	0.84	CIBC, 2020	0.13	0.84	0.97
CIBC, 2021	0.13	0.76	0.91	CIBC, 2021	0.12	0.85	0.97
BNS, 2019	0.20	N/A	N/A	BNS, 2019	0.13	N/A	N/A
BNS, 2020	0.20	0.86	0.98	BNS, 2020	0.13	0.90	0.94
BNS, 2021	0.20	0.88	0.98	BNS, 2021	0.14	0.96	0.95
TD, 2019	0.19	N/A	N/A	TD, 2019	0.14	N/A	N/A
TD, 2020	0.17	0.86	0.97	TD, 2020	0.13	0.59	0.96
TD, 2021	0.17	0.92	0.99	TD 2021	0.15	0.39	0.96
BMO, 2019	0.17	N/A	N/A	BMO, 2019	0.11	N/A	N/A
BMO, 2020	0.18	0.86	0.98	BMO, 2020	0.11	0.87	0.96
BMO, 2021	0.18	0.97	0.98	BMO, 2021	0.11	0.87	0.96
Mean	0.17	0.86	0.92	**Mean**	0.13	0.80	0.95
Median	0.17	0.87	0.98	**Median**	0.13	0.86	0.96
Std Dev	0.02	0.09	0.09	**Std Dev**	0.01	0.17	0.01

Table 3. Results for Jaccard Similarity

5.2. **Burrows' Delta**

The results of Burrows' Delta are found in Table 4. The lowest score between human, GPT, and the disputed category indicates the likely author; these have been bolded in Table 4. If the lowest score points to the disputed category, it means that the Delta is having trouble distinguishing between human and GPT.

Disclosure	Test Document	Human	GPT	Disputed
Disclosure 1	22	3.65	**2.90**	3.11
Disclosure 1	11	2.42	2.97	**2.36**
Disclosure 1	1	**2.33**	3.55	3.04
Disclosure 1	30	3.88	**2.82**	3.14
Disclosure 2	22	1.59	**0.19**	1.43
Disclosure 2	11	1.79	2.11	**1.33**
Disclosure 2	1	**1.24**	2.36	1.54
Disclosure 2	30	2.32	**1.35**	2.10

Table 4. The Results of Burrows' Delta

Documents were blinded when given to the Burrows' Delta calculation. For reference, documents 1–15 are human-authored and documents 16–30 are AI-generated (ChatGPT). Therefore, when we reviewed the four test cases (8 documents total for each disclosure) Burrows' Delta was able to correctly identify 6 documents — for documents 22, 1, and 30 for both disclosure 1 and disclosure 2.

The results did identify an interesting test case, however, for test document 11. This document is more similar to the stylometry found in the disputed documents, which is a mix of both human and AI authored texts. We found this very interesting and using our master key list of all documents and authors, we determined that the text for document 11 for both disclosure 1 and disclosure 2 were authored by TD Bank. So, we extended our testing to evaluate all of TD's texts. Using the same blinded protocol, we calculated Burrows' Delta for documents 11 (included in the original test set), as well as documents 10 and 12 for

Doc - Human	Disclosure 1	Disclosure 2	Doc - GPT	Disclosure 1	Disclosure 2
text1	23.53	47.81	text16	32.70	41.14
text2	26.52	51.81	text17	33.63	41.60
text3	22.87	52.17	text18	33.63	40.05
text4	19.01	43.98	text19	19.84	41.21
text5	22.25	43.70	text20	20.60	41.21
text6	25.95	46.37	text21	21.54	41.21
text7	39.65	34.56	text22	21.78	40.93
text8	**39.95**	32.31	text23	21.78	43.01
text9	**39.95**	31.89	text24	21.78	43.01
text10	15.50	**29.33**	text25	25.53	48.43
text11	**14.00**	42.28	text26	25.53	48.43
text12	**14.00**	35.38	text27	25.53	48.43
text13	21.02	46.84	text28	24.84	**56.50**
text14	21.58	45.46	text29	24.84	**56.60**
text15	21.58	47.41	text30	24.84	**56.50**
Mean	24.49	42.08	**Mean**	25.22	45.88
Median	22.25	43.89	**Median**	24.84	43.01
Std Dev	8.79	7.46	**Std Dev**	4.60	6.21

Table 5. Results for Flesch Reading Ease

both disclosures. The results were the same — these documents were more similar to the disputed category. We did notice that the Cosine results show an incredibly high similarity YoY for these disclosures for TD Bank (range of 0.94 to 0.99), suggesting a "boilerplate" approach, which is also found in the AI-generated disclosures of ChatGPT. This may be why the Delta cannot fully distinguish TD's disclosures as either human or GPT.

5.3. **Flesch Reading Ease**

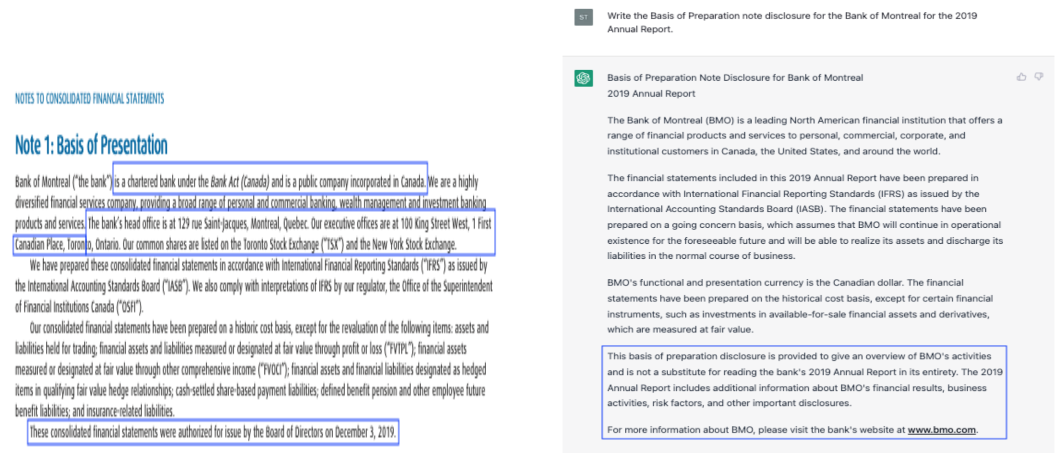
The results from the Flesch Reading Ease evaluation can be found in Table 5. As discussed in the methodology section, we are interested in texts that are below 60 as those texts are considered difficult to read. Overall, we find that the ChatGPT texts are easier to read as those texts have a mean of 42.08 and 45.88 for disclosures 1 and 2 respectively, whereas the human-authored texts have a mean of 24.29 and 25.22.

This has important implications for the readability of the annual report. While ChatGPT's text is currently lacking in similarity for both disclosures, the significant increase in the readability is noteworthy. Keeping in mind that readability below 30 is considered to be university grad/professional-level, the majority of the human-authored texts fall in that category. ChatGPT's text is more accessible and measured at university level/business entry-level, and get very close to the main threshold of 60, which is the very beginning of the "difficult to read" category.

We also compared the statistical significance of the standard deviations using the F-test. The difference of standard deviations between the human-authored and ChatGPT-generated texts for disclosure 1 is statistically significant with a p-value of 0.02 and an F-statistic of 3.66, at a significance level of 0.05. The difference for Disclosure 2 was not statistically significant, with a p-value of 0.4991 and an F-statistic of 1.446.

5.4. **Side-By-Side Comparison of Text**

While similarity, stylometry, and readability can provide a lot of information, it is also important to do a side-by-side comparison of the texts to see what ChatGPT emulates well and what it does not. It is also key to see what textual components are left out or added in by ChatGPT.



NOTES TO CONSOLIDATED FINANCIAL STATEMENTS

Note 1: Basis of Presentation

Bank of Montreal ("the bank") is a chartered bank under the *Bank Act (Canada)* and is a public company incorporated in Canada. We are a highly diversified financial services company, providing a broad range of personal and commercial banking, wealth management and investment banking products and services. The bank's head office is at 129 rue Saint-Jacques, Montreal, Quebec. Our executive offices are at 100 King Street West, 1 First Canadian Place, Toronto, Ontario. Our common shares are listed on the Toronto Stock Exchange ("TSX") and the New York Stock Exchange.

We have prepared these consolidated financial statements in accordance with International Financial Reporting Standards ("IFRS") as issued by the International Accounting Standards Board ("IASB"). We also comply with interpretations of IFRS by our regulator, the Office of the Superintendent of Financial Institutions Canada ("OSFI").

Our consolidated financial statements have been prepared on a historic cost basis, except for the revaluation of the following items: assets and liabilities held for trading; financial assets and liabilities measured or designated at fair value through profit or loss ("FVTPL"); financial assets measured or designated at fair value through other comprehensive income ("FVOCI"); financial assets and financial liabilities designated as hedged items in qualifying fair value hedge relationships; cash-settled share-based payment liabilities; defined benefit pension and other employee future benefit liabilities; and insurance-related liabilities.

These consolidated financial statements were authorized for issue by the Board of Directors on December 3, 2019.

(a) BMO's Basis of Presentation Note Disclosure.

Write the Basis of Preparation note disclosure for the Bank of Montreal for the 2019 Annual Report.

Basis of Preparation Note Disclosure for Bank of Montreal
2019 Annual Report

The Bank of Montreal (BMO) is a leading North American financial institution that offers a range of financial products and services to personal, commercial, corporate, and institutional customers in Canada, the United States, and around the world.

The financial statements included in this 2019 Annual Report have been prepared in accordance with International Financial Reporting Standards (IFRS) as issued by the International Accounting Standards Board (IASB). The financial statements have been prepared on a going concern basis, which assumes that BMO will continue in operational existence for the foreseeable future and will be able to realize its assets and discharge its liabilities in the normal course of business.

BMO's functional and presentation currency is the Canadian dollar. The financial statements have been prepared on the historical cost basis, except for certain financial instruments, such as investments in available-for-sale financial assets and derivatives, which are measured at fair value.

This basis of preparation disclosure is provided to give an overview of BMO's activities and is not a substitute for reading the bank's 2019 Annual Report in its entirety. The 2019 Annual Report includes additional information about BMO's financial results, business activities, risk factors, and other important disclosures.

For more information about BMO, please visit the bank's website at www.bmo.com.

(b) ChatGPT's Basis of Presentation Note Disclosure for BMO.

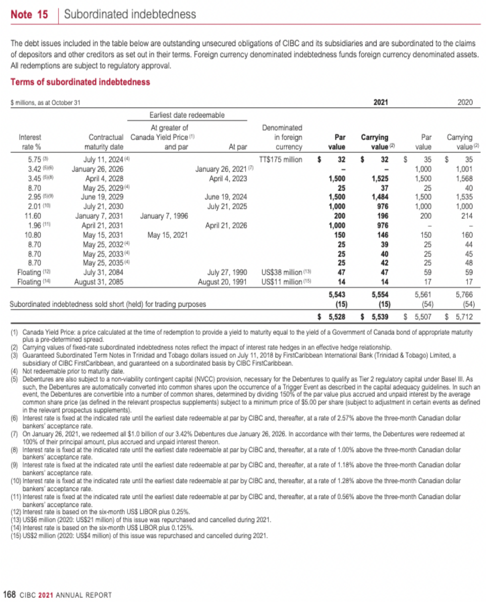
Figure 1. Comparison of Note Disclosure 1

For disclosure 1 — Basis of Presentation — the Bank of Montreal's 2019 note disclosure is presented alongside ChatGPT's generated text in Figure 1, where (a) presents the note disclosure as written by BMO and (b) provides ChatGPT's generated note disclosure for BMO. Boxes have been added to draw attention to missing information as well as additional information that is not normally provided as part of the disclosure.

It is clear that ChatGPT understood the query, as it has begun the note disclosure with "Basis of Presentation". It did not provide all of the required information as this note disclosure is usually several pages long and discusses the accounting policies used in much more detail. As such, we truncated the bank's original text to include only the general information when analyzing the similarity, stylometry, and readability in an effort to compare "like" text based on ChatGPT's limitations.

ChatGPT does capture a number of relevant information points such as bank name, what products and services the bank offers, and that the statements have been prepared under IFRS. It does miss some important information such as the bank's charter, the fact that it is a public company, the bank's head office address, that is traded on the Toronto Stock Exchange (TSX) and New York Stock Exchange (NYSE) , and that the consolidated financial statements were authorized for issue by the Board (along with the date of authorization). (See blue boxes in figure (a)).

We also noticed that ChatGPT included text that does not appear in any of the five banks' "Basis" note disclosures — a boilerplate advisory that the disclosure is not a substitute for reading the bank's Annual Report in its entirety (even though the note disclosure is *in* the annual report), and a "For more information" notice that includes the bank's website. We found this very interesting as ChatGPT has added this in. Given that the number of Canadian Annual Reports is far less ubiquitous than American Annual Reports, we consulted the United States' largest bank JPMorgan Chase & Co and reviewed its "Basis of Presentation" note disclosure [29]. This extra information was not present in JPMorgan's note disclosure for 2017–2021 either. This raises an interesting question as to why ChatGPT is providing the boilerplate advisory (when the note is part of the Annual Report), and why it is including the website, when this information is not normally provided as part of this disclosure.



Note 15 | Subordinated indebtedness

The debt issues included in the table below are outstanding unsecured obligations of CIBC and its subsidiaries and are subordinated to the claims of depositors and other creditors as set out in their terms. Foreign currency denominated indebtedness funds foreign currency denominated assets. All redemptions are subject to regulatory approval.

Terms of subordinated indebtedness

$ millions, as at October 31					2021		2020	
		Earliest date redeemable						
Interest rate %	Contractual maturity date	At greater of Canada Yield Price [(1)] and par	At par	Denominated in foreign currency	Par value	Carrying value [(2)]	Par value	Carrying value [(2)]
5.75 [(3)]	July 11, 2024 [(4)]			TT$175 million	$ 32	$ 32	$ 35	$ 35
3.42 [(5)(6)]	January 26, 2026		January 26, 2021 [(7)]		–	–	1,000	1,001
3.45 [(5)(8)]	April 4, 2028		April 4, 2023		1,500	1,525	1,500	1,568
8.70	May 25, 2029 [(4)]				25	37	25	40
2.95 [(5)(9)]	June 19, 2029		June 19, 2024		1,500	1,484	1,500	1,535
2.01 [(10)]	July 21, 2030		July 21, 2025		1,000	976	1,000	1,000
11.60	January 7, 2031	January 7, 1996			200	196	200	214
1.96 [(11)]	April 21, 2031		April 21, 2026		1,000	976	–	–
10.80	May 15, 2031	May 15, 2021			150	146	150	160
8.70	May 25, 2032 [(4)]				25	39	25	44
8.70	May 25, 2033 [(4)]				25	40	25	45
8.70	May 25, 2035 [(4)]				25	42	25	48
Floating [(12)]	July 31, 2084		July 27, 1990	US$38 million [(13)]	47	47	59	59
Floating [(14)]	August 31, 2085		August 20, 1991	US$11 million [(15)]	14	14	17	17
					5,543	5,554	5,561	5,766
Subordinated indebtedness sold short (held) for trading purposes					(15)	(15)	(54)	(54)
					$ 5,528	$ 5,539	$ 5,507	$ 5,712

[(1)] Canada Yield Price: a price calculated at the time of redemption to provide a yield to maturity equal to the yield of a Government of Canada bond of appropriate maturity plus a pre-determined spread.

[(2)] Carrying values of fixed-rate subordinated indebtedness notes reflect the impact of interest rate hedges in an effective hedge relationship.

[(3)] Guaranteed Subordinated Term Notes in Trinidad and Tobago dollars issued on July 11, 2018 by FirstCaribbean International Bank (Trinidad & Tobago) Limited, a subsidiary of CIBC FirstCaribbean, and guaranteed on a subordinated basis by CIBC FirstCaribbean.

[(4)] Not redeemable prior to maturity date.

[(5)] Debentures are also subject to a non-viability contingent capital (NVCC) provision, necessary for the Debentures to qualify as Tier 2 regulatory capital under Basel III. As such, the Debentures are automatically converted into common shares upon the occurrence of a Trigger Event as described in the capital adequacy guidelines. In such an event, the Debentures are convertible into a number of common shares, determined by dividing 150% of the par value plus accrued and unpaid interest by the average common share price (as defined in the relevant prospectus supplements) subject to a minimum price of $5.00 per share (subject to adjustment in certain events as defined in the relevant prospectus supplements).

[(6)] Interest rate is fixed at the indicated rate until the earliest date redeemable at par by CIBC and, thereafter, at a rate of 2.57% above the three-month Canadian dollar bankers' acceptance rate.

[(7)] On January 26, 2021, we redeemed all $1.0 billion of our 3.42% Debentures due January 26, 2026. In accordance with their terms, the Debentures were redeemed at 100% of their principal amount, plus accrued and unpaid interest thereon.

[(8)] Interest rate is fixed at the indicated rate until the earliest date redeemable at par by CIBC and, thereafter, at a rate of 1.00% above the three-month Canadian dollar bankers' acceptance rate.

[(9)] Interest rate is fixed at the indicated rate until the earliest date redeemable at par by CIBC and, thereafter, at a rate of 1.18% above the three-month Canadian dollar bankers' acceptance rate.

[(10)] Interest rate is fixed at the indicated rate until the earliest date redeemable at par by CIBC and, thereafter, at a rate of 1.28% above the three-month Canadian dollar bankers' acceptance rate.

[(11)] Interest rate is fixed at the indicated rate until the earliest date redeemable at par by CIBC and, thereafter, at a rate of 0.56% above the three-month Canadian dollar bankers' acceptance rate.

[(12)] Interest rate is based on the six-month US$ LIBOR plus 0.25%.

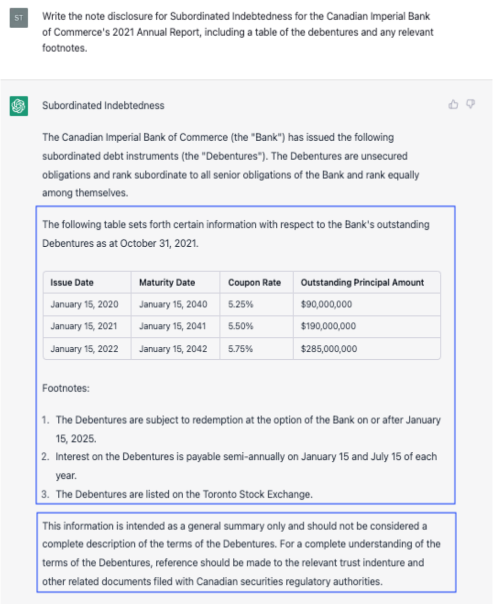
[(13)] US$6 million (2020: US$21 million) of this issue was repurchased and cancelled during 2021.

[(14)] Interest rate is based on the six-month US$ LIBOR plus 0.125%.

[(15)] US$2 million (2020: US$4 million) of this issue was repurchased and cancelled during 2021.

168 CIBC **2021** ANNUAL REPORT

(a) CIBC's Subordinated Indebtedness Note Disclosure.



Write the note disclosure for Subordinated Indebtedness for the Canadian Imperial Bank of Commerce's 2021 Annual Report, including a table of the debentures and any relevant footnotes.

Subordinated Indebtedness

The Canadian Imperial Bank of Commerce (the "Bank") has issued the following subordinated debt instruments (the "Debentures"). The Debentures are unsecured obligations and rank subordinate to all senior obligations of the Bank and rank equally among themselves.

The following table sets forth certain information with respect to the Bank's outstanding Debentures as at October 31, 2021.

Issue Date	Maturity Date	Coupon Rate	Outstanding Principal Amount
January 15, 2020	January 15, 2040	5.25%	$90,000,000
January 15, 2021	January 15, 2041	5.50%	$190,000,000
January 15, 2022	January 15, 2042	5.75%	$285,000,000

Footnotes:

1. The Debentures are subject to redemption at the option of the Bank on or after January 15, 2025.
2. Interest on the Debentures is payable semi-annually on January 15 and July 15 of each year.
3. The Debentures are listed on the Toronto Stock Exchange.

This information is intended as a general summary only and should not be considered a complete description of the terms of the Debentures. For a complete understanding of the terms of the Debentures, reference should be made to the relevant trust indenture and other related documents filed with Canadian securities regulatory authorities.

(b) ChatGPT's Subordinated Indebtedness Note Disclosure for CIBC.

Figure 2. Comparison of Note Disclosure 2

ChatGPT's performance for disclosure 2 was a good attempt, but it only provided "dummy" information, and generated the same dummy information for every bank, when queried. Although the table was stripped out for the analysis, it was necessary to get ChatGPT to generate the footnotes. ChatGPT also included an advisory that the debenture information was only a summary and that for a complete understanding, the trust indentures and other related documents should be referred to . While Canadian banks do provide supplementary documents on their trust indentures that users *can* consult, the note disclosure is required to provide all of the relevant information on the debentures. In CIBC's footnotes (Figure 2, (a)), the relevant high level information is provided in the footnotes and there is no reference to debenture documents filed with regulators included in its footnotes (or for any of the other four banks' disclosures). Again, we also checked JPMorgan Chase & Co's subordinated debt disclosure, and there is no mention of referring to supplementary debenture regulatory filings [29]. Therefore, we conclude that ChatGPT has added this advisory on its own.

6. Conclusion, Future Work, and Limitations

This research has provided a benchmark for ChatGPT's current abilities to write financial statement note disclosures. This benchmark is important as it identifies where the gap is between the generated text and the desired/needed text. It also identifies areas where, for use in the financial world, ChatGPT needs to be further trained. Our research highlights that note disclosures are *currently* challenging for ChatGPT. We also draw attention to the fact that while ChatGPT is not able to fully provide the necessary output just yet, its text is much more readable (and therefore accessible) than that of its human counterparts, particularly for disclosure 1.

These results create exciting opportunities for further research. A big question that we would like to address in future research is how LLMs (Large Language Models, like

ChatGPT) can be used to help improve the readability of financial statement note disclosures. The original intent and purpose of the stock market was to make it open to everyone — professional and lay-person alike. As financial reporting and market regulations have evolved, the original intent is becoming more difficult to achieve, as the communications from company to shareholder (or potential shareholder) have become exceedingly long with very complicated text. The potential ability of LLMs to distill the message to one that is more universally accessible is an area of research that is worth exploring.

The high similarity results for both human and ChatGPT's texts on a YoY basis demonstrate that both use a "boilerplate" approach for the note disclosures. Contemporary research is interested in the detection and removal of boilerplate in web corpora. An unintended consequence of premature boilerplate removal, however, could render financial reports unintelligible, as key parts of the reports could be removed. Also, given the scarcity of financial statement text as well as the opportunities for fraudulent or incorrect text to be generated, more work is needed on how LLMs can augment financial reporting text, and how fraudulent/incorrect text can be detected quickly.

There are several important limitations of this research: we only evaluated the text from Canada's five major banks (RBC, CIBC, BNS, TD, and BMO). It may be that ChatGPT's performance is better in generating the disclosures of other Canadian banks. Also, we only selected two note disclosures. Although these are required (and fairly standardized) note disclosures, it may be that ChatGPT's performance is better when generating other note disclosures. Finally, ChatGPT was trained on the common crawl web corpora which consists of 12 years of common crawl data [30]. That means that for each of the 5 banks, there are only 12 annual reports that ChatGPT has seen. This could have a material effect on the outcome of its generation.

References

[1] S. Harvey. *Advice for managing the length of annual reports.* https://www.fm-magazine.com/news/2017/dec/managing-the-length-of-annual-reports-201717989.html. Last Accessed: 2023-16-1. 2017.

[2] Government of Canada. *International Financial Reporting Standards (IFRS).* https://www.canada.ca/en/revenue-agency/services/tax/businesses/topics/international-financial-reporting-standards-ifrs.html. Last Accessed: 2023-16-1. n.d.

[3] Chartered Professional Accountants of Canada. *Understanding Reports on Financial Statements.* https://www.cpacanada.ca/-/media/site/operational/rg-research-guidance-and-support/docs/01878-rg-understanding-reports-on-financial-statements-jan-2020.pdf?la=en&hash=BE02C8C0BDC4B7FB908004C113D9538AE266FBB9. Last Accessed: 2023-16-1. 2020.

[4] D. E. Kieso, J. J. Weygandt, T. D. Warfield, I. M. Wiecek, and B. J. McConomy. *Intermediate Accounting, Volume 1.* John Wiley & Sons, 2019.

[5] J Hughes and A Fisher. *Reading Financial Statements - What Do I Need To Know?* https://www.cpacanada.ca/-/media/site/business-and-accounting-resources/docs/reading-financial-statements----what-do-i-need-to-know.pdf?la=en&hash=3BDE48F69C73D2C4C022935CBC6404261DE764F9. Last Accessed: 2023-16-1. 2014.

[6] Deloitte. *IAS 1 — Presentation of Financial Statements.* https://www.iasplus.com/en/standards/ias/ias1. Last Accessed: 2023-16-1. n.d.

[7] Deloitte. *IFRS 7 — Financial Instruments: Disclosures.* https://www.iasplus.com/en/standards/ifrs/ifrs7. Last Accessed: 2023-16-1. n.d.

[8] OpenAI. *ChatGPT.* https://openai.com/blog/chatgpt/. Last Accessed: 2023-16-1. 2022.

[9] A. Huang et al. "Similarity measures for text document clustering". In: *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand.* Vol. 4. 2008, pp. 9–56.

[10] A. W. Qurashi, V. Holmes, and A. P. Johnson. “Document processing: Methods for semantic text similarity analysis”. In: *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE. 2020, pp. 1–6.

[11] R. R. Singh, D. Koundal, and R. Tiwari. “Linguistic Approach for Authentic Authorship”. In: (2021).

[12] N. Smeuninx, B. De Clerck, and W. Aerts. “Measuring the readability of sustainability reports: A corpus-based analysis through standard formulae and NLP”. In: *International Journal of Business Communication* 57.1 (2020), pp. 52–85.

[13] R. J. M. Ventayen. “OpenAI ChatGPT Generated Results: Similarity Index of Artificial Intelligence-Based Contents”. In: *Available at SSRN 4332664* (2023).

[14] S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner. “Mathematical capabilities of ChatGPT”. In: *arXiv preprint arXiv:2301.13867* (2023).

[15] *SEDAR Homepage*. n.d. URL: https://www.sedar.com/homepage_en.htm.

[16] Royal Bank of Canada. *Investor Relations*. https://www.rbc.com/investor-relations/. Last Accessed: 2023-11-2. n.d.

[17] Canadian Imperial Bank of Commerce. *Investor Relations*. https://www.cibc.com/en/about-cibc/investor-relations.html. Last Accessed: 2023-11-2. n.d.

[18] J. Wang and Y. Dong. “Measurement of text similarity: a survey”. In: *Information* 11.9 (2020), p. 421.

[19] V. Arnaboldi, A. Passarella, M. Conti, and R. I. Dunbar. “Chapter 5 - Evolutionary Dynamics in Twitter Ego Networks”. In: *Online Social Networks*. Ed. by V. Arnaboldi, A. Passarella, M. Conti, and R. I. Dunbar. Computer Science Reviews and Trends. Boston: Elsevier, 2015, pp. 75–92. ISBN: 978-0-12-803023-3. DOI: https://doi.org/10.1016/B978-0-12-803023-3.00005-9. URL: https://www.sciencedirect.com/science/article/pii/B9780128030233000059.

[20] F. D. Laramée. “Introduction to stylometry with Python”. In: *The Programming Historian* 7 (2018).

[21] J. Burrows. “‘Delta’: a measure of stylistic difference and a guide to likely authorship”. In: *Literary and linguistic computing* 17.3 (2002), pp. 267–287.

[22] No author listed. *spacy readability*. https://spacy.io/universe/project/spacy_readability. Last Accessed: 2023-16-1. n.d.

[23] J. R. Layton. “A chart for computing the Dale-Chall Readability Formula above fourth grade level”. In: *Journal of Reading* 24.3 (1980), pp. 239–244.

[24] Moraine Park Technical College. *What Flesch Reading Ease Score Should My Content Have?* https://www.morainepark.edu/help/. Last Accessed: 2023-16-1. n.d.

[25] B. Walker. *16 Millionaires Who Made Their Fortunes Without a College Degree*. https://financebuzz.com/millionaires-without-a-college-degree. Last Accessed: 2023-16-1. 2022.

[26] R. Schäfer. “Accurate and efficient general-purpose boilerplate detection for crawled web corpora”. In: *Language Resources and Evaluation* 51 (2017), pp. 873–889.

[27] J. Leonhardt, A. Anand, and M. Khosla. “Boilerplate removal using a neural sequence labeling model”. In: *Companion Proceedings of the Web Conference 2020*. 2020, pp. 226–229.

[28] J. Bose. “Extraction of Relevant Images for Boilerplate Removal in Web Browsers”. In: *arXiv preprint arXiv:2001.04338* (2019).

[29] JPMorgan Chase & Co. *INVESTOR RELATIONS - Annual Report & Proxy*. https://www.jpmorganchase.com/ir/annual-report. Last Accessed: 2023-16-1. n.d.

[30] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.