Evaluating the Output of Large Language Models: An Explorative Corpus-based Approach

DANIEL BAUER

ACM Reference Format:
Daniel Bauer. 2024. Evaluating the Output of Large Language Models: An Explorative Corpus-based Approach. 1, 1 (March 2024), 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In their attempt to describe the term *corpus linguistics*, McEnery and Hardie [13] call it an area that focuses upon a set of procedures for studying language, which, in turn, allows us to use existing and even find new theories of language. While this definition may traditionally have been restricted to language as it is produced by humans, recent advances in the field of Natural Language Processing (NLP) offer a plethora of texts that, while not originally created by any human, still deserve to be analysed from a linguistic perspective. Considering the fact that the output of a language model essentially forms a corpus, investigating these texts most definitely warrants the usage of corpus linguistics tools. Thus, this paper aims to provide an explorative study on how such techniques can be used on data that is produced by language models. On a general level, the goal of this paper is to test the application of corpus linguistic techniques to gain new insights into the output produced by Large Language Models. More precisely, this goal can be divided into several general research questions:

(1) Can corpus linguistic methods be used to gain new insights into the output of LLMs?
(2) In which cases can corpus linguistic methods be useful when considering generated output? In which cases not so much?
(3) Can these methods be used to provide some degree of further explainability?

In order to find potential answers to these questions, I test different methods taken from corpus linguistic analysis on a practical example. Concretely, I compare linguistic data taken from the British National Corpus to data produced by a LLM fine-tuned on said corpus data. Most importantly, I consider the linguistic data both quantitatively and qualitatively. Furthermore, I consider differences between different text categories present in the chosen corpus, providing a potential application for further research. Concretely, I formulate the following research questions for my explorative case study:

(1) How does the output of the fine-tuned models compare to the corresponding dataset used to fine-tune the model? Are there notable differences between the text types?
(2) How does the output of the fine-tuned models differ when it is trained for more epochs? Is it more or less similar to the original dataset?

Author's address: Daniel Bauer, Daniel.Bauer@stud.uni-regensburg.de.

Manuscript submitted to ACM

By answering these questions, I aim to provide a case study that helps evaluate the usage of different corpus linguistic tools to study the linguistic properties of artificially generated text data. The entire code that was used for this paper can be accessed via the corresponding repository[1].

2 RELATED WORK

While different NLP techniques have been used to enhance the field of corpus linguistics overall [24] and for specific use cases, such as enhancing software tools [1] or automatic annotation [23], corpus linguistic methods have hardly been employed to evaluate the output produced by LLMs. Nonetheless, some efforts have been made to investigate LLM-generated text from a linguistic perspective. Muñoz-Ortiz et al. [16], for instance, compare New York Times articles to Llama-generated news, concluding that human texts exhibit more scattered sentence length distributions, and more aggressive emotions than LLM-generated texts. However, they purely prompt the base version of the respective model, whereas this paper aims to evaluate models that were fine-tuned on the dataset that the output will be compared to. Furthermore, this paper will use a more state-of-the-art option in the form of Llama-2. Providing a more practical use case of displaying linguistic diversity of text datasets, Reif et al. [19] offer a tool to analyze synthetic text data in different ways. Concretely, the text is clustered along syntactic, lexical, and semantic axes, allowing for both an overview and the inspection of specific examples. This aims to address not only problems of repetition that may occur in the output produced by LLMs, but also to allow the inspection of the generated output in a general sense.

In an extended sense, this paper also relates to synthetic data in general, as I investigate and evaluate the output that was (synthetically) produced by LLMs, which could then potentially be used to extend and enhance different datasets. Guo et al. [8] highlight the problems that the usage of synthetic data could have on the linguistic diversity of models trained on such data. Thus, studying the differences in linguistic diversity between the original training data and that produced by a fine-tuned model can be a first step for mitigating the issues described by [8].

Nonetheless, specific corpus linguistic studies of LLM output have not yet been conducted. This paper aims to provide a starting point in the form of a general structure and assessment of how such studies could be conducted.

3 METHODS

3.1 General approach

On a general level, the approach is based on evaluating and comparing the output produced by fine-tuned models to the dataset that the respective models were fine-tuned on. During this process, the model itself will be treated as a black box. In other terms, this means that the concrete workings and architecture of the chosen model will not be considered when evaluating the output produced by said model. Instead, only the original training dataset and the generated output will be investigated. Two factors will be varied during this process: First, the three different text categories provided by the selected corpus and, second, the amount of epochs the respective model is trained for. Concretely, the amount of epochs was varied from 2 to 4 and 8. This was done in the hope of obtaining models that provide different quality of output data without changing the underlying model architecture, considering the fact that training up to 4 epochs of data has almost negligible changes of loss compared to having unique data [15].

[1] https://github.com/danielbauer1860/LDS_Project

3.2 Model

Llama 2[2] [21] has the major advantage of being an open-source model which can be used free of charge and comes close to GPT-3.5 on academic benchmarks [21]. Hence, it is ideally suited for this project. It must be stated, however, that this only applies to the variant with 70 billion parameters. Given that the variant with 70 billion parameters would require more computing power and training time than would be feasible within the methodological approach I chose, I have to settle for the 7 billion parameter variant. Nonetheless, this version still remains quite competitive and still clearly outperforms similarly sized variants of other open-source alternatives such as Falcon [18] or MPT [14] on major benchmarks [21].

3.3 Dataset

The British National Corpus (BNC) can be defined as a synchronic, general, monolingual, and mixed sample corpus [3]. It was published in the year 1994 and consists of a total of roughly 100 million words [2]. It was composed with the goal of characterizing the state of contemporary British English in its various social and generic uses [2]. Generally, it is split into a written and spoken section, with the former comprising ninety percent and the latter the remaining ten percent of the corpus. For the purposes of this paper, a smaller scale version of the BNC, named BNC Baby [4], was chosen. This was done in order to allow for a faster training process while still maintaining the high quality of representative data featured in the BNC. In order to scale down the BNC to the BNC Baby, the sections contained within the original corpus had to be rearranged. This resulted in four sections, each containing roughly one million words: Fiction, Newspapers, Academic, and Spoken. Given that the focus on this paper is on written texts, the spoken section is disregarded. Compared to the original BNC, the other three sections were changed in slightly different ways:

Fiction. The fictional section is based on the "imaginative" category of the original BNC. Pres, having adults as target audience and the genre label "W fict prose" [6]. Additionally, only one title was allowed per author.

Newspapers. Despite the fact that newspapers were no distinct category in the original BNC, they still constituted a very large portion in it. Thus, the creators of BNC Baby were able to extract articles by the information contained in the text headers. The data in this section is obtained from thirteen different newspapers, with 60% of data coming from five national papers and 40% from eight local papers. While the creators tried to sample the texts in an even distribution for each category and newspaper respectively, due to the variation in size of newspaper texts, some differences in the amount of data obtained from each newspaper remain [6].

Academic. Similarly to the newspaper section, academic texts form no distinct category in the original BNC. Thus, texts also had to be identified via their headers. In total, 30 texts were sampled from 501 texts that fall under the term "academic writing", with no specific target being set with regards to a proportion between texts from periodicals to texts from books.

3.4 Implementation

3.4.1 Preprocessing. To ensure proper fine-tuning of the generator model, preparing the original dataset to be in the best possible form is of the utmost importance. Given that the BNC Baby can only be obtained in XML, several challenges had to be overcome to prepare the dataset in a satisfactory manner.

[2] https://huggingface.co/meta-llama

Table 1. BNC Baby

Section	Texts
Fiction	25
Newspapers	97
Spoken	30
Academic	30

Removing the headers containing metadata. Naturally, each text contained some amount of meta information, describing the text source, length, domain, and other information provided by the creators of the corpus. For the purposes of this paper, however, this information is irrelevant, since it would only inject misleading tokens into the fine-tuning process. Hence, headers were removed. Furthermore, many files contained headings and author information in lines three and four. Since it can be argued that neither headings nor author names reflect the given text properly, these were also removed to provide a further degree of consistency throughout the text.

Removing tags. Considering the XML format that the BNC Baby is provided in consists of the raw text being augmented with different XML tags to enhance it with additional information, these tags also had to be removed in order to obtain the raw text.

Remove unnecessary newlines. Since the original files were nicely structured over numerous lines, reading those files in python added several new line tokens. To ensure further clarity of the dataset, those were also removed.

Things that were not removed. While the afore-mentioned removals ensure a higher quality dataset, several things were explicitly kept to maintain the integrity of the original text type. These include: Roman Numerals, scientific symbols, numbers in general, and other special characters. Given that especially academic texts contain a considerable amount of these characters, removing those would substantially change the unique properties of that text type, in turn rendering comparison between different text types less interesting. Nonetheless, this must be kept in mind when considering the results.

The preprocessed dataset was then compiled and exported to be used in the respective fine-tuning pipelines.

3.4.2 Fine-tuning. Considering that I use Google Colab for the fine-tuning process, major system constraints had to be overcome in order to properly train Llama-2. Concretely, even the smallest Llama-2 version, which comes at 7 billion parameters, does not fit into the 15GB of VRAM that Google Colab's T4 GPU offers. Nonetheless, as described in [17], this can be solved by employing low-rank adaptation (LoRA) [10], and, more concretly, QLoRA [5]. As for concrete implementation, this means that Llama-2 was loaded in 4-bit using the 4-bit NormalFloat (NF4) datatype, as proposed by [5], with a LoRA configuration being set up before the actual fine-tuning process. This approach allowed for the 7 billion parameter version to be successfully loaded on the T4 GPU. Llama-2 was then fine-tuned on each of the three relevant text types using the hyperparameters displayed in Table 2. These are, to some degree, also based on the work of [17]. However, some adjustments had to be made to deal with the afore-mentioned RAM limitiations. Concretely, the batch size and gradient accumulation steps were reduced to one each. The only parameter that was varied throughout the process, was the number of epochs, which was varied from 2 to 4 and 8. In total, this means that nine final models were obtained, the training time of each can be seen in Table 3. Furthermore, Figure 1 shows the development of the

Table 2. Parameters used for fine-tuning

Hyperparameter	Value
Batch Size	1
Gradient Accumulation Steps	1
Number of Train Epochs	2, 4, 8
Optimizer	Paged AdamW (32-bit)
Learning Rate	4×10^{-5}
Max Gradient Norm	0.3
Warmup Ratio	0.03
Learning Rate Scheduler Type	Constant

Table 3. Fine-tuning time

Epochs	Model	Time
2	fictional	08:19min
	academic	12:24min
	newspaper	29:58min
4	fictional	18:42min
	academic	20:32min
	newspaper	56:53min
8	fictional	38:29min
	academic	44:09min
	newspaper	118:14min

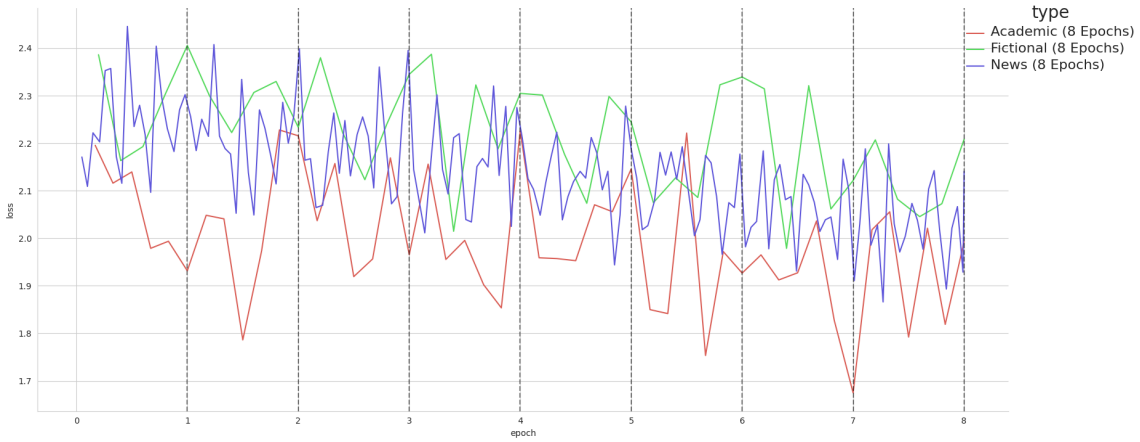

Fig. 1. Training Loss over 8 Epochs

training loss for the version trained over 8 epochs. As evident from this figure, the loss converges for each the text types.

Table 4. Tokens in each output corpus

Category	Epochs	Tokens
Academic	2	26580
	4	22480
	8	24698
Fictional	2	30961
	4	22210
	8	26359
News	2	25503
	4	20896
	8	27456

3.4.3 Generation of Data. The fine-tuned models were then used to generate new texts. The first step to do so involved the extraction of prompts from the original dataset. This was done in order to ensure that the generated dataset would roughly be based around similar topics as the original dataset. Nonetheless, the prompts were taken from all three text types. The prompt extraction was conducted in a two-step random sampling process: First, I randomly extracted 20 texts from each text type from which I sampled a random sentence in the second step. To ensure a qualitative prompt, the length of the sampled sentence had to be at least 5 tokens. These 60 sentences were then given as prompts to each of the nine models fine-tuned in the previous step.

The hyperparameters for generation were selected in accordance with the work provided by [20], who chose to employ nucleus sampling [9] for their data generation. In concrete terms, this means that a top-p value of 0.92 and a temperature of 0.7 were the relevant parameters. Additionally, the maximum length of the generated texts was set to 1000 tokens.

Furthermore, the generation process had to be redone multiple times until the final results were satisfactory for proper statistical analysis. Most notably, the original output seemed to consistently devolve into repeated sequences, a common problem that is alleviated by nucleus sampling [9], but not entirely eliminated by it [7]. This could also be seen in the first results generated by the fine-tuned models, with some texts being free of excessive repetition and others repeating more quickly. While this is certainly interesting in itself, the underlying repetition problem makes statistical analysis of the raw output almost impossible and would not offer many insights. Therefore, the texts were regenerated with an additional hyperparameter in the form of a repetition penalty, which was set to 1.1. This immediately produced more realistic output free of excessive repetitions.

In total, the generation of each dataset took around an hour.

3.4.4 Postprocessing. Given that the selected hyperparameters for generation provided output without any major problems, no post-processing had to be conducted on the actual text itself. Nonetheless, the output was changed from the .csv to the .txt format to allow for easier and more consistent qualitative analysis. Table 4 displays the final amount of tokens of each output corpus. As evident, these can range from 20 000 to 30 000.

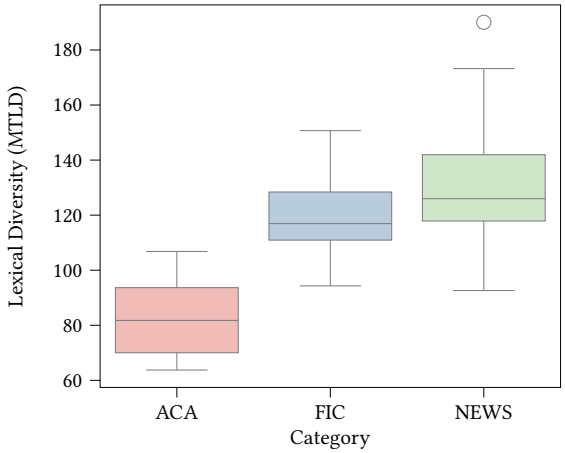

Fig. 2. Lexical Diversity in the BNC Baby

4 RESULTS

4.1 Quantitative

4.1.1 Lexical Diversity. As listed by [12], there are a number of different measures for lexical diversity. However, [12] further suggests that the measure of textual lexical diversity (MTLD) stands out as not correlating with text length. Given that there are large differences between the text lengths of the texts from the BNC and the output produced by the fine-tuned models, MTLD seems like the natural choice.

BNC Baby. The lexical diversity of the original BNC Baby is displayed in Figure 2. Shapiro-Wilk tests conducted for each of the original BNC Baby categories revealed a normal distribution for academic ($W = 0.94$, $p = 0.12$) and fictional texts ($W = 0.97$, $p = 0.73$) but not for newspaper texts ($W = 0.97$, $p = 0.01$). Further, a significant Levene's test revealed heteroskedasticity between the text types: $F(2, 179) = 3.19$, $p = 0.04$. With these requirements, a non-parametric Kruskal-Wallis test was conducted, revealing a significant difference in lexical diversity between the groups: $H(2) = 72.12$, $p = 2.18 \times 10^{-16}$.

Academic. For the academic output displayed in Figure 3, the lexical diversity of all outputs is distributed normally, except the one produced by the model trained for only 2 Epochs ($W = 0.92$, $p = 0.001$). Additionally, a significant Levene's test showed unequal variance between the texts: $F(3, 206) = 6.92$, $p = 1.85 \times 10^{-4}$. Thus, a Kruskal-Wallis test was also chosen to compare these texts, showing a significant difference: $H(3) = 40.22$, $p = 9.59 \times 10^{-9}$. In combination with the boxplot displayed in Figure 3, this implies that the output produced by the models fine-tuned on the academic dataset is significantly more lexically diverse than the academic portion of the BNC itself.

For the generated output of the three fine-tuned models, a Levene's test did not yield significant results, demonstrating equal variance: $F(2, 177) = 0.23$, $p = 0.80$. Additionally, a Kruskal-Wallis test did not show significant differences between the groups: $H(2) = 1.21$, $p = 0.54$.

The results of these tests show that the output's lexical diversity of the models fine-tuned on the academic dataset did not change significantly the more epochs the model was trained for. However, in combination with the boxplot

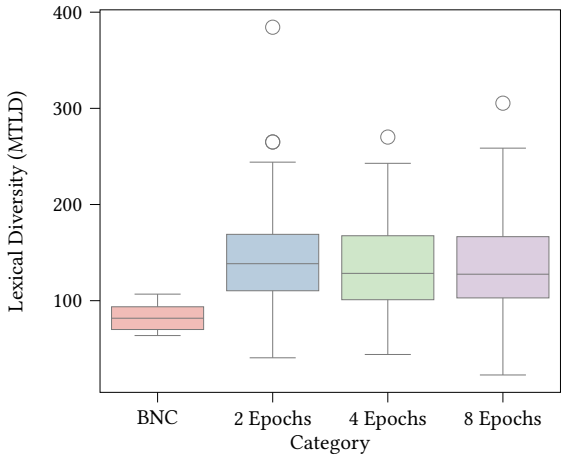

Fig. 3. Lexical Diversity in the Academic Output

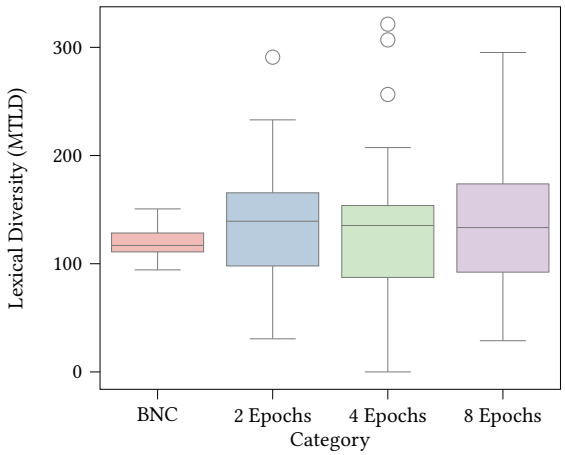

Fig. 4. Lexical Diversity in the Fictional Output

displayed in Figure 3, these results imply that the output produced by the models fine-tuned on the academic dataset is significantly more lexically diverse than the academic portion of the BNC itself.

Fictional. For the lexical diversity of the fictional texts, Shapiro-Wilk tests demonstrated normality for all distributions. Nonetheless, a significant Levene's test ($F(3, 201) = 7.38$, $p = 1.03 \times 10^{-4}$) showed a difference in variance. These requirements allow for the usage of a Welch ANOVA, which was also significant: $F(3, 111.46) = 3.02$, $p = 0.03$. This demonstrates a significant difference between the groups displayed in Figure 4.

For the generated output, a Levene's test was not significant, demonstrating equal variance: $F(2, 177) = 0.62$, $p = 0.54$. Given that all data was distributed normally, a one-way ANOVA could be performed, showing no significant differences between the three groups: $F(2, 177) = 0.23$, $p = 0.79$.

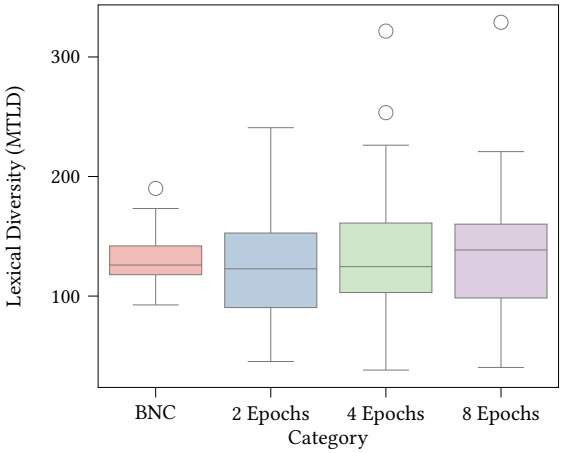

Fig. 5. Lexical Diversity in the News Output

Similarly to the academic portion, there seems to be a significant difference in lexical diversity between the fictional texts in the original BNC and those produced by the fine-tuned Llama-2 models.

News. As suggested by Figure 5 the news datasets have unequal variance, which is further underlined by a Levene's test yielding significant results: $F(3, 273) = 13.59$, $p = 2.74 \times 10^{-8}$. Furthermore, Shapiro-Wilk tests revealed that only the 2 epoch version ($W = 0.96$, $p = 0.07$) is distributed normally. Thus, a Kruskal-Wallis test was conducted which revealed no significant differences between the four news datasets: $H(3) = 2.02$, $p = 0.57$. This entails that the fine-tuning process retained a similar amount of lexical diversity as present in the original dataset, which did not change over the amount of training epochs.

Comparing only the output of the trained models to each other, a Levene's test did not show significant results, showing homoscedasticity between the output generated by the fine-tuned models: $F(2, 177) = 0.15$, $p = 0.86$. Furthermore, a Kruskal-Wallis test did not show significant differences between the groups: $H(2) = 1.35$, $p = 0.51$.

Conversely to the academic and fictional texts, no significant differences between the original newspaper texts and the generated ones could be detected.

Comparison between text types. Finally, considering that the lexical diversity of the outputs generated by the Llama-2 models does not differ significantly, the output of the variants trained for 8 Epochs were compared. A Levene's test comparing the lexical diversity of the three models yielded no significant results, demonstrating equal variance: $F(2, 177) = 0.86$, $p = 0.43$. Furthermore, a Kruskal-Wallis test showed no significant differences between the lexical diversity of the output generated by the three models: $H(2) = 0.48$, $p = 0.79$. This entails that the difference in lexical diversity that is present in the original BNC texts was not retained during the fine-tuning process.

Summary of important results. The most notable result from the lexical diversity data is the fact that the significant difference between the text types present in the BNC Baby data is not replicated by the models trained for 8 Epochs. This suggests that simply fine-tuning the model on datasets stemming from different text types did not transfer the lexical diversity of the respective text type - at least not to a significant extent. Furthermore, the three models of each text types do not display any significant differences when compared to each other, with the academic and fictional

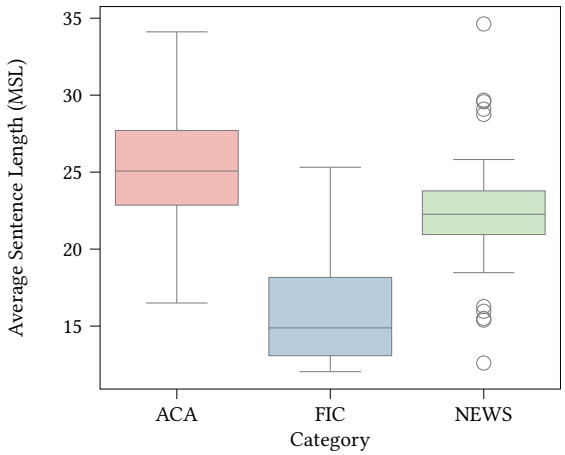

Fig. 6. Average Sentence Length in the BNC Baby

output being noticeably different from the respective BNC Baby portion. This entails two things: First, fine-tuning Llama-2 for more epochs did not impact the lexical diversity to a significant extent. Second, in combination with the comparative results discussed before, the pre-training of the model appears to influence the lexical diversity more than the fine-tuning process.

4.1.2 Sentence Length. In addition to focusing on the lexical aspect of a text in the form of investigating its lexical diversity, examining the average sentence length of a text is a valid option to also consider the syntactic complexity of texts [11].

BNC Baby. Looking at the distribution of average sentence length between the three different text categories in the BNC Baby, only the academic portion is distributed normally ($W = 0.97$, $p = 0.50$). A significant Levene's test showed homoscedasticity: $F(2, 179) = 1.87$, $p = 0.16$. The subsequent Kruskal-Wallis test demonstrated a significant difference between the average sentence length of the text types: $H(2) = 56.12$, $p = 6.50 \times 10^{-13}$.

Academic. The average sentence length of the models fine-tuned on the academic portion of the BNC Baby were only distributed normally in the case of the original BNC data ($W = 0.97$, $p = 0.50$) and the variant trained for 4 Epochs ($W = 0.97$, $p = 0.23$). Furthermore, comparing the BNC Baby data to the generated one via a Levene's test did provide significant results, showing heteroscedasticity: $F(3, 206) = 3.81$, $p = 0.01$. A significant Kruskal-Wallis test also demonstrated differences between the average sentence lengths: $H(3) = 29.01$, $p = 2.00 \times 10^{-6}$.

Looking at the only the generated data, a Levene's test showed equal variance: $F(2, 177) = 1.52$, $p = 0.22$. The Kruskal-Wallis test did not yield significant results, revealing no relevant differences between the average sentence length of the generated texts: $H(2) = 1.03$, $p = 0.60$.

Fictional. Of the models fine-tuned on the fictional texts, both the 2 ($W = 0.98$, $p = 0.26$) and 8 Epoch ($W = 0.97$, $p = 0.20$) variant displayed normality. Notably, a Levene's test showed no significant results, showing equal variance: $F(3, 201) = 2.05$, $p = 0.11$. Furthermore, a Kruskal-Wallis test showed no significant differences between the average sentence lengths for the fictional portion: $H(3) = 4.61$, $p = 0.20$.

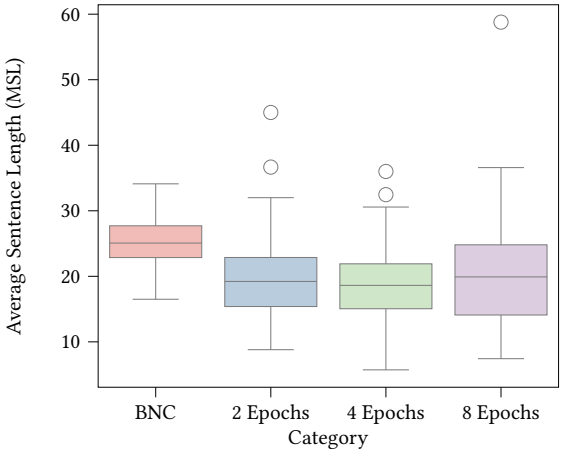

Fig. 7. Average Sentence Length in the Academic Output

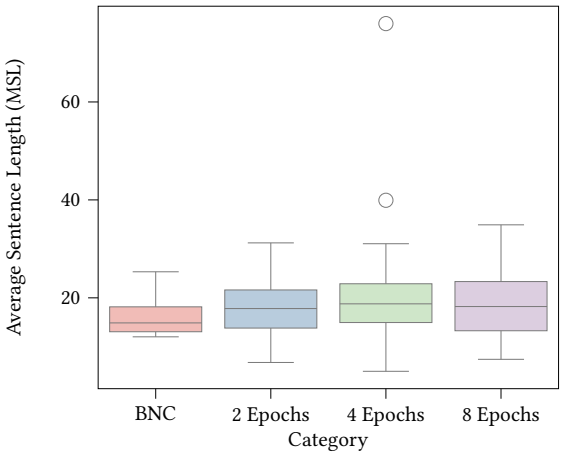

Fig. 8. Average Sentence Length in the Fictional Output

Regarding only the generated data, a Levene's test showed equal variance: $F(2, 177) = 0.44$, $p = 0.64$. As expected from the previous results, the Kruskal-Wallis test did not display significant results: $H(2) = 0.51$, $p = 0.77$.

News. For the texts produced by the models fine-tuned on the newspaper portion, both the 4 ($W = 0.98$, $p = 0.51$) and 8 Epoch ($W = 0.98$, $p = 0.28$) versions show normal distribution with regards to average sentence length. A subsequent significant Levene's test further demonstrated unequal variance: $F(3, 273) = 17.11$, $p = 3.30 \times 10^{-10}$. The Kruskal-Wallis test revealed a significant difference between the sentence length of the texts produced by the newspaper models: $H(3) = 22.43$, $p = 5.30 \times 10^{-5}$.

Comparing only the model output against each other, the text produced by the three models demonstrates equal variance in their average sentence length, as underlined by a significant Levene's test: $F(2, 177) = 0.88$, $p = 0.42$.

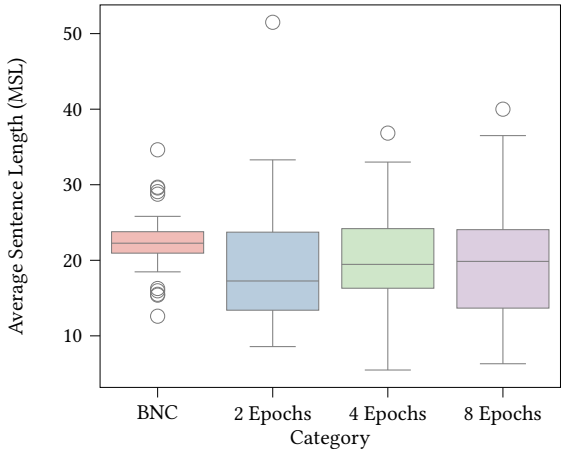

Fig. 9. Average Sentence Length in the News Output

Additionally, a Kruskal-Wallis test did not show any significant differences between the output of the three models: $H(2) = 1.86$, $p = 0.39$.

Comparison between text types. Similarly to lexical diversity, the statistical tests regarding average sentence length were also performed on the respective 8 Epoch variant of each text type. In the first step, this showed a Levene's test with no significant results: $F(2, 177) = 0.45$, $p = 0.64$. Further, the Kruskal-Wallis test was also not significant, pointing at no notable differences between the 8 epoch variants trained on different text types: $H(2) = 0.60$, $p = 0.74$.

Summary of important results. The results regarding the average sentence lengths of the investigated texts point at similar insights that were already observed in the section on lexical diversity. First, the significant difference in sentence length in the BNC Baby, as displayed in Figure 6, is not mirrored in the output produced by variants fine-tuned for 8 epochs. Second, comparing only the generated data to itself did not reveal any significant differences. Third, the academic and newspaper output differed significantly from their respective BNC Baby counterparts with regards to average sentence length. These results have similar implications as those stated in the section on lexical diversity, namely that fine-tuning the models did not impact the models syntactic complexity to a significant manner. In combination with the results obtained from the lexical diversity section, one can thus claim that neither the lexical nor the syntactic intricacies of each text type were accurately replicated consistently - at least from a quantitative perspective.

4.2 N-Grams

Moving beyond techniques for quantitative analysis, the output produced by Llama-2 should also be considered from a qualitative perspective. While there are a number of different angles that this could be done from, investigating n-grams serves as good starting point to not only consider single words by themselves, but rather see them occurring in phrases, which can potentially even make for basic collocations [22]. Considering that the most interesting aspect of a language model's output is not the necessarily the single words that it produces but rather the context and conditions under which certain words are produced, n-grams can offer numerous insights regarding the generated texts.

Table 5. The most common tri-grams in the academic texts

Rank	**BNC Baby**	Freq	**2 Epochs**	Freq	**4 Epochs**	Freq	**8 Epochs**	Freq
1	in terms of	338	one of the	12	be able to	9	city limits area	44
2	there is a	301	as well as	9	the coypu man	8	areas outside city	43
3	one of the	275	whether or not	9	a number of	8	outside city limits	43
4	a number of	272	types of batteries	7	a lot of	7	rural areas outside	40
5	as well as	264	between genesis and	6	in the house	6	area surrounding rural	39
6	it is not	258	eigenvalues and eigenvector	6	one of the	6	surrounding rural areas	39
7	some of the	251	of the study	6	you need to	6	limits area surrounding	23
8	part of the	246	that he was	6	be a writer	5	country side area	22
9	there is no	245	the fact that	6	some of the	5	countryside country side	22
10	per cent of	244	to do so	6	two big guys	5	area surrounded countryside	21
11	the number of	240	type of battery	6	what kind of	5	side area surrounding	21
12	the use of	230	an eigenvector of	5	you want to	5	surrounded countryside country	20
13	the fact that	211	it was a	5	a couple of	4	there was a	7
14	in order to	191	man on the	5	all the way	4	one of the	6
15	in which the	184	on the right	5	and it was	4	the specimen is	6

Table 6. The most common tri-grams in the fictional texts

Rank	**BNC Baby**	Freq	**2 Epochs**	Freq	**4 Epochs**	Freq	**8 Epochs**	Freq
1	one of the	392	one of the	8	conceptualization conceptualization conceptualization	93	boy in the	28
2	out of the	383	as well as	7	we need to	15	in the striped	28
3	it was a	353	due to the	7	the elementary divisors	11	the boy in	25
4	there was no	215	is given by	7	elementary divisors of	9	the striped pajamas	25
5	a lot of	203	the number of	7	upper hessenberg matrix	9	be able to	14
6	for a moment	195	you want to	7	be able to	8	one of the	12
7	he had been	188	in front of	6	secretary of state	8	cos θb	11
8	she had been	176	it was a	6	the matrix is	8	sin θb	11
9	in front of	173	the social sciences	6	the secretary of	8	mg cos θ	9
10	it had been	165	the use of	6	an upper hessenberg	7	mg sin θ	9
11	was going to	164	a room with	5	it was a	7	the fact that	9
12	it was the	163	an article for	5	of the elementary	7	as well as	8
13	the end of	158	for a moment	5	one of the	7	θbmg	8
14	that he was	155	he said I	5	part of the	7	we need to	7
15	that he had	153	I want to	5	roots of the	7	a number of	6

4.2.1 Academic. Considering the most frequent tri-grams appearing in the academic texts, as displayed in Table 5, one can immediately notice *in terms of* as standing out as the most frequently occurring one in the BNC Baby data. Compared to the lists for the other text types (Table 6, Table 7), this phrase appears characteristic of the academic texts contained within the BNC Baby. Similar claims could also be made for *a/the number of*, *the fact that*, or *in order to* - phrases that one would also typically associate with more academic texts.

Nonetheless, these phrases are not quite as present in the output generated by the fine-tuned models. While *the fact that* and *a number of* appear in the versions trained for 2 and 4 epochs respectively, *in terms of* is not present in none of the fifteen most common tri-grams that were considered. Still, especially the 2 epoch version appears to relatively frequently produce output that could be considered academic jargon, such as *eigenvalues and eigenvectors* and *types of batteries*. Unfortunately, however, the tri-grams in the 8 epoch version appear to have been skewed by a repetition problem, which was also confirmed by a manual look at the data. This entails that despite the fact that a repetition penalty parameter was used during generation, the issue could not be prevented absolutely. Nevertheless, investigating n-grams appears like a valid way to catch such cases.

Table 7. The most common tri-grams in the newspaper texts

Rank	**BNC Baby**	Freq	**2 Epochs**	Freq	**4 Epochs**	Freq	**8 Epochs**	Freq
1	one of the	411	first citizens bank	15	be able to	15	the number of	8
2	the end of	219	ordinary and necessary	10	the combined transformation	9	one of the	8
3	out of the	190	and necessary expenses	9	a lot of	6	do you think	6
4	per cent of	164	there is no	9	as a result	6	in order to	6
5	the year old	159	it was a	8	the working class	6	in the first	6
6	a lot of	158	that it is	8	we need to	6	looked at her	6
7	as well as	157	with the verb	8	would have been	6	what do you	6
8	in the first	143	said mr stibbard	7	I want to	5	whether or not	6
9	part of the	135	the design argument	7	in other words	5	williams was the	6
10	there is a	134	a lot of	6	one of the	5	as one of	6
11	electronic edition of	132	one of the	6	that there are	5	by williams was	5
12	it is a	131	a room with	5	that there is	5	for a moment	5
13	to be a	124	the subject of	6	the end of	5	found by williams	5
14	may not be	123	a result of	5	the identity matrix	5	have to be	5
15	it will be	122	agrees with the	5	the total work	5	he did not	5

4.2.2 Fictional. Many of the most frequent tri-grams present in the fictional texts taken from the BNC Baby are ones that one would typically expect in a narrative text, such as third person phrases along the verb *be*, such as *he had been*, *she had been*, *it was the*, *that he was*, or *that he had*. Given that these texts are typically telling a story, these constructions are unsurprising. In a similar vein, descriptors of time and place, such as *for a moment*, *the end of* and *in front of* also seem to be typical for these types of texts. Interestingly, the most frequently occurring tri-gram of the original BNC Baby *one of the* also appears as the most frequent one for the 2 epoch version, while also appearing within the top fifteen of the other model's output.

Standing out as another case of a repetition problem, *conceptualization* appears in sequence a total number of 93 times in a text produced by the 4 epoch variant. Investigating this further, this appears to have been caused by the generated text appending a "keyword" section to its text, where it eventually ended up simply repeating *conceptualization* over and over again.

Furthermore, while the most frequent tri-grams of the 8 epoch variant may look suspiciously like another instance of a repetition problem, a closer look revealed a text that was framed as a movie review of "The Boy in the Striped Pajamas". Nonetheless, these tri-grams also display numerous instances of mathematical jargon being used, such as referring to matrices, as in *upper hessenberg matrix* or *the matrix is*, or making explicit mathematical expressions as in the tri-grams using a greek *theta* or *sin* and *cos*. This could, for instance, be caused by a prompt nudging the output towards mathematical topics, which would, in turn, imply that the fine-tuning alone does not necessarily "overwrite" the information obtained in the prompt.

4.2.3 Newspapers. Regarding the newspaper section in the BNC Baby, one entry clearly stand out in the form of *electronic edition of*, probably appearing in the context of newspapers being accessed via their respective electronic versions. Compared to the other text types, however, the stylistic properties of newspaper texts do not become explicitly apparent. It could be argued that phrases such as *there is a*, *it is a*, *to be a*, *may not be*, or *it will be* might be representative of a matter-of-fact way of describing events and carefully considering implications of events while still maintaining a professional distance.

In a similar vein, the tri-grams present in the output produced by the models trained on the newspaper texts appear equally unassuming, with no glaring instances of the afore-mentioned repetition problem being apparent. Additionally,

the most frequent tri-gram in the BNC Baby *one of the*, is also present within the fifteen most common ones of the output produced by the three models.

Comparison between text types. Comparing the results between the text types, one must note that the newspaper texts appear to be the only ones free from any repetition problems. However, this must be taken with caution, as the other text types only appear to fall victim to this problem in the 8 epoch version of the academic model and the 4 epoch version of the fictional model, with only one out of the sixty texts of each text types being the one at fault. Overall, this clearly demonstrates the effectiveness of implementing a repetition penalty.

Furthermore, the produced output seems to mimic some of the features of the original training corpus, as can be seen in the usage of academic jargon in the academic output, as displayed in Table 5. However, this must also be put into perspective, as the fictional texts also appear to talk about academic subjects.

Additionally, it appears that the most frequent tri-gram of each BNC Baby text category can also be found in the fifteen most frequent tri-grams of the generated output quite consistently, with the only exception being academic texts.

Overall, however, it must be noted that the overall frequency of tri-grams present in the generated output do not allow for any substantial conclusions to be made. Thus, further research may increase the amount of generated data that is investigated.

5 DISCUSSION

These results have demonstrated several things. First, it appears that both the difference in lexical diversity and sentence length between the text types that was evident in the BNC Baby did not get reproduced by the fine-tuned models. In combination with the fact that these metrics did not change significantly for the output of models that were trained for more metrics, it appears that fine-tuning the model did not substantially transfer the lexical diversity or the syntactic complexity of the base dataset to the output. Second, the repetition problem, while alleviated by the implementation of a repetition penalty during the generation process, did not get eliminated entirely. Further research may investigate, how this problem affects other models and how this could be optimized for Llama-2 specifically. Third, while the tri-grams demonstrated that each text type possesses characteristic phrases that are used quite frequently in that respective category of the BNC, this is not directly replicated in the generated data. However, this must be seen considering two caveats. On the one hand, the sample of the produced output may simply be too low and, on the other hand, we do not know the distribution of tri-grams in the dataset that was used to pre-train Llama-2. Nonetheless, also considering the fact that the fictional texts displayed phrases that one would typically associate with academic jargon, it appears that the texts were also influenced by the provided prompts quite noticeably. Thus, while the output texts appear to be influenced by the dataset it was fine-tuned on, factors such as the pre-trained nature of Llama-2 and the provided prompts appear to have influenced the output in a noticeable way.

Based on the previously discussed results, several answers to the overarching research questions regarding the usage of corpus linguistics for the evaluation of LLM-generated output can be formed. First, it has become clear that the methods used throughout this paper can indeed help gain new insights regarding the text produced by LLMs. Specifically, the measures of linguistic diversity and average sentence length have offered clear insights as to how the output of the fine-tuned models differs from the original BNC Baby counterparts. In this case, it has become clear that fine-tuning alone did not lead the output to mimic the lexical diversity and sentence length of the original dataset. Nonetheless, these metrics can be used to evaluate the effectiveness of a training process while treating the model itself as a black box. Further research could, for instance, consider not only fine-tuning a pre-existing model, but conducting

such studies for the training of a neural model from scratch. Other use cases for these metrics may consider different types of input datasets or compare different models against each other. Overall, these quantitative techniques can pose a valid tool in understanding the generated data on a more abstract linguistic level. Second, however, investigating tri-grams, while unveiling certain issues that arose during the generation of the data, did not provide the same level of insight as the quantitative measures discussed beforehand. Still, they helped uncover problematic cases that could potentially lead to issues if the dataset would be used in further downstream tasks, such as using it to enhance a dataset, which could then, in turn, be used to fine-tune another model. Nonetheless, using only n-grams to investigate the exact output of the language models, while providing a slight overview of the texts' contents, did not unveil the exact intricacies of each text type. Thus, further research could resort to additional qualitative assessment besides relying on n-grams. Third, it can be claimed that using the selected tools can, indeed, help better explain the output of the model when it is fine-tuned. In this case, for instance, it has become clear that the linguistic diversity and sentence length are, for the most part, influenced by the underlying pre-training. This could, for instance, explain discrepancies between the original dataset and the output produced by the fine-tuned models. Furthermore, investigating the n-grams has unveiled certain problematic cases, explaining certain problematic cases. Nonetheless, it must be kept in mind that the selected approach is based on treating the model as black box, which, naturally, severely limits the potential for explainability to hypotheses we can make based on observations and statistical analysis. Nevertheless, similar approaches may thus allow for more approachable cross-disciplinary usage of LLMs for different research questions, especially from the linguistic perspective.

6 CONCLUSION

Throughout this paper, it has become clear that a corpus linguistic investigation of LLM-generated texts can yield interesting new insights, such as how fine-tuning influenced the lexical and syntactic properties of the output text compared to the original dataset. Nonetheless, this study was limited by several factors. First, both fine-tuning nine models and generating a dataset for each of the models was rather time-consuming, especially considering the fact that the generated output had to be reiterated upon by implementing an additional hyperparameter. Further research could hence limit itself to fewer models, allowing them to go further in-depth in their analysis. Second, focusing purely on n-grams for the qualitative perspective did not reveal the entire depths of the produced output. Thus, the dataset could be looked at from further angles, which could potentially unveil additional properties that this study could not uncover. Third, given that the model itself was only the 7 billion parameter version of Llama-2, the findings drawn from the results presented in this paper can not necessarily be generalized for other state-of-the-art language models.

Considering these limitations, future work can improve on the foundation provided by this study in several directions. First, different architectures apart from Llama-2 could be compared to each other. Second, additional corpus linguistic methods may be applied to investigate the output produced by language models. Third, different corpora can be used to train or fine-tune a model. Fourth, the size of the generated output corpora could be increased to allow for the better investigation of word frequencies and the potential use of other corpus linguistic methods.

Considering corpus linguistics as a tool to investigate deeper linguistic research questions, future research may also concern itself with studying such questions in the context of language models. These could, for instance, concern themselves with the analysis of different specific phenomena and how they behave in the context of a language model's output. Ultimately, looking at the output produced by language models through the lens of corpus linguistics can well be an option to help us better understand and evaluate synthetically created texts.

REFERENCES

[1] Laurence Anthony. 2023. Corpus AI: Integrating Large Language Models (LLMs) into a Corpus Analysis Toolkit. Presentation given at the 49th Annual Conference of the Japan Association for English Corpus Studies (JAECS), Kansai University, Osaka, Japan. https://osf.io/srtyd/

[2] Guy Aston and Lou Burnard. 1998. *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh.

[3] Lou Burnard. 1995. Users Reference Guide for the British National Corpus. https://homepages.abdn.ac.uk/k.vdeemter/pages/teaching/NLP/practicals/bnc-doc.pdf

[4] BNC Consortium. 2007. British National Corpus, Baby edition. http://hdl.handle.net/20.500.12024/2553 Oxford Text Archive.

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314 [cs.LG]

[6] edited by Lou Burnard. 2008. Reference Guide to BNC Baby (second edition). (2008). http://www.natcorp.ox.ac.uk/corpus/baby/manual.pdf

[7] Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A Theoretical Analysis of the Repetition Problem in Text Generation. arXiv:2012.14660 [cs.CL]

[8] Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text. arXiv:2311.09807 [cs.CL]

[9] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. arXiv:1904.09751 [cs.CL]

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL]

[11] Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15 (2010), 474–496. https://api.semanticscholar.org/CorpusID:17189214

[12] Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42, 2 (01 May 2010), 381–392. https://doi.org/10.3758/BRM.42.2.381

[13] Tony McEnery and Andrew Hardie. 2011. *What is corpus linguistics?* Cambridge University Press, 1–24.

[14] MosaicML NLP Team. 2023. *Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs.* www.mosaicml.com/blog/mpt-7b Accessed: 2023-05-05.

[15] Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling Data-Constrained Language Models. arXiv:2305.16264 [cs.CL]

[16] Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting Linguistic Patterns in Human and LLM-Generated Text. arXiv:2308.09067 [cs.CL]

[17] Armin Norouzi Norouzi. 2023. Mastering Llama 2: A Comprehensive Guide to Fine-Tuning in Google Colab. (2023). https://medium.com/artificial-corner/mastering-llama-2-a-comprehensive-guide-to-fine-tuning-in-google-colab-bedfcc692b7f Accessed: 2024-03-14.

[18] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. arXiv:2306.01116 [cs.CL]

[19] Emily Reif, Minsuk Kahng, and Savvas Petridis. 2023. Visualizing Linguistic Diversity of Text Datasets Synthesized by Large Language Models. arXiv:2305.11364 [cs.CL]

[20] Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J. Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services* 64 (2022), 102771. https://doi.org/10.1016/j.jretconser.2021.102771

[21] 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. https://arxiv.org/pdf/2307.09288

[22] M. Weisser. 2016. *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis.* Wiley. https://books.google.de/books?id=5vAwBgAAQBAJ

[23] Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of AI-assisted pragmatic annotation: The case of apologies. arXiv:2305.08339 [cs.CL]

[24] Qiuying Zhao. 2022. Review of Natural Language Processing for Corpus Linguistics. *Corpus Pragmatics* 6, 4 (01 Dec 2022), 311–314. https://doi.org/10.1007/s41701-022-00127-6